Spring 1-1-2015

Essays on Military and Civilian Manpower Planning

Jose Alberto Ramirez
University of Colorado at Boulder, josergond@gmail.com

Follow this and additional works at: https://scholar.colorado.edu/opim_gradetds

Part of the Business Administration, Management, and Operations Commons

Recommended Citation
https://scholar.colorado.edu/opim_gradetds/3

This Dissertation is brought to you for free and open access by Operations & Information Management at CU Scholar. It has been accepted for inclusion in Operations & Information Management Graduate Theses & Dissertations by an authorized administrator of CU Scholar. For more information, please contact cuscholaradmin@colorado.edu.
Essays on Military and Civilian Manpower Planning

by

Jose A. Ramirez

B.S., University of Notre Dame, 1994
M.S., University of Texas at Austin, 2008

A thesis submitted to the
Faculty of the Graduate School of the
University of Colorado in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Operations Management, Leeds School of Business
2015
This thesis entitled:
Essays on Military and Civilian Manpower Planning
written by Jose A. Ramirez
has been approved for the Operations Management, Leeds School of Business

________________________
Thomas W. Vossen

________________________
Manuel Laguna

Date ____________________

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.
Disclaimer

The views expressed in this dissertation are those of the author and do not reflect the official policy or position of the United States Army, Department of Defense, or the U.S. Government.
Civilian and military manpower planning is commonly conducted by human resource (HR) planners having little or no experience using and developing quantitative methods. If an organization forecasts an increase or decrease in production, HR planners must respond with recommended changes to the employee footprint. It is vital for planners to properly identify personnel needs by job skill and tenure in order to properly recommend the required changes in total personnel. The use of mathematical models can provide valuable support when making manpower planning decisions, such as personnel recommendations based on characteristics of a set of employees or issuance of incentive bonuses based on a company’s skill shortages.

In the manpower planning process two important levers are available to a company to shape its workforce—recruitment and retention of personnel, and these can be employed individually or collectively. For example, a company may augment its workforce or replace departing personnel by recruiting new employees, or it can seek to retain incumbent employees by incentivizing or de-incentivizing them with respect to their retention decisions. In this dissertation, we first present mathematical models that can assist civilian and military personnel planners in balancing recruitment and retention. Both hiring and retention incur costs: monetary compensation may be required to entice current employees to remain, and recruiting employees will almost certainly involve occupational training and reduce productivity until this training is complete. Thus, in both military and civilian sectors, this choice as to whether to hire or retain can be of the utmost importance.

Effective manpower planning requires a thorough understanding of employee retention behavior, that is, the likelihood that employees will stay with the organization. In this dissertation, we therefore also propose models that can assist HR planners in estimating retention probabilities
of employee-groups fitting particular profiles. Accurately forecasting the retention propensity of a
group of employees is key because it helps a company prepare for the possible departure of these
employees, while also identifying possible reasons (via a set of attributes) motivating their decision
to leave.

The issues outlined above may require the use of large data sets, which can complicate real-
time analysis efforts. As such, the final chapter of this dissertation presents methods for extracting
a representative subset of employee data. The resulting data sets can simplify subsequent analysis
efforts and enable real-time analysis of retention and recruitment decisions. The resulting models
have applications beyond manpower planning and may provide a general framework for alleviating
the burden of conducting analysis using large data sets in a variety of settings.
Dedication

To my beloved children, Katia and Isaiah. Always reach for the stars.
Acknowledgements

I would like to thank everyone who made this dissertation possible. First, I would like to thank the United States Military Academy at West Point, for giving me the opportunity to pursue advanced civil schooling and for funding this research. My advisor, Thomas W. Vossen, who spent countless hours mentoring me and guiding this research. My committee, Dan Zhang, Manuel Laguna, Stephen Lawrence, and Marco Better for all of their insights which helped direct and focus my research efforts. Robinson Lemos from the Universidade do Estado de Mato Grosso (Brasil) for his research assistance and for sharing his Java programming knowledge.

Thanks to all my friends at OptTek System, especially James Kelly, Jay April, and Fred Glover, for their help and support during the completion of this work. Additionally, a special thanks to Mark Zais for being my studying partner in our doctoral courses and for providing valuable feedback on my ideas regarding proposed dissertation topics.

I am grateful to my parents, Fidel and Felicitas Ramirez, for their encouraging words and repeated motivational words of “échale ganas.” I wish to extend a sincere and heartfelt thank you to my loving wife, Maria Esther, who closed her own window-of-opportunity in life so that mine could be opened; I shall never forget your sacrifice. Finally, to my daughter and son, Katia and Isaiah, thank you for understanding when you repeatedly heard the words, “Dad has to study.”

Thank you all for your untiring support.
Contents

Chapter

1 Introduction 1
1.1 The Issue of Balancing Recruitment and Retention in Military Personnel Planning . 2
1.2 Two Methods of Predicting Employee-Group Retention Behavior . . . . . . . . . . . . 3
1.3 Surrogate Data Representations for Feature Selection and Prediction . . . . . . . . 5

2 Balancing Recruitment and Retention in Military Personnel Planning 7
2.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 7
2.2 Literature Review . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 10
2.3 Background and Data Source . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
2.3.1 Data Extraction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 13
2.4 An Empirical Study of Retention Rates . . . . . . . . . . . . . . . . . . . . . . . . . 17
2.4.1 Variable definitions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 17
2.4.2 Logistic Regression Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 19
2.4.3 Observations and Insights . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
2.5 A Retention and Recruitment Model . . . . . . . . . . . . . . . . . . . . . . . . . . . 26
2.5.1 An Extension to Multiple Retention-Eligible Personnel Groups . . . . . . . . . . 29
2.6 Numerical Experiments of Recruitment and Retention Model . . . . . . . . . . . . . . 30
2.6.1 Army Enlisted Empirical Study Parameters and Assumptions . . . . . . . . . . . 31
2.6.2 Binomial Modeling Observations and Insights using Mean Values . . . . . . . . 33
2.6.3 Binomial Modeling Observations and Insights using Group Approach 37

2.7 Summary and Future Directions 41

3 Two Methods of Predicting Employee-Group Retention Behavior 44

3.1 Introduction 44

3.2 Literature Review 46

3.3 Group-based Retention Trees 49

3.3.1 Decision Trees 50

3.3.2 Toward Group-based Decision Trees 52

3.3.3 Constructing Group-based Decision Trees 54

3.4 K-nearest Neighbor (KNN) Methods 57

3.5 Employee Workforce Empirical Study 60

3.5.1 Personnel Data Extraction 60

3.5.2 Experimental Setup 66

3.5.3 Empirical Results 67

3.6 Conclusions and Further Research 74

4 Surrogate Data Representations for Feature Selection and Prediction 76

4.1 Introduction 76

4.2 Literature Review 79

4.3 Preliminaries 82

4.4 Generating Tranches 83

4.5 Signatures of Variables and Vectors 86

4.6 Selecting a Surrogate Set $X^*$ to Replace $X$ 87

4.7 Exploiting the Tranches for Feature Selection 89

4.7.1 Simple Feature Selection Strategies 90

4.7.2 Accounting for Interdependencies 90

4.7.3 Exploiting Co-Diversity for Feature Selection 91
Tables

Table

2.1 Summary Statistics of Predictors for Infantry Career Field (Four Year Reenlistment, Station Stabilization Option) .......................... 19
2.2 Model Fit Statistics .................................................. 21
2.3 Testing Global Null Hypothesis: Beta=0 ............................. 22
2.4 Analysis of Maximum Likelihood Estimates ......................... 22
2.5 Odds Ratio Estimates .................................................. 23
2.6 Retention Rate Scenarios for First-Term Infantry Soldiers .......... 25
2.7 Enlisted Army Strengths and Losses (2006-2009) .................... 31
2.8 Binomial Modeling Parameters: Army Recruiting and Retention .... 32
2.9 Average Variable Values for Mean Value Experiment ............... 34
2.10 Empirical Results for Infantrymen using Binomial Recruiting and Retention Model (Attribute Means) ........................................... 35
2.11 Converted Values for Continuous Variables ......................... 37
2.12 Parameter Specifications for Each Group and their Respective Optimal Retention Rates ......................................................... 39
2.13 Empirical Results for Infantrymen Using Group Binomial Recruiting and Retention Model ......................................................... 39

3.1 Simple Retention Tree Example ...................................... 52
Figures

2.1 Data Structure of the Total Army Personnel Database ........................................ 14
3.1 Numerical Example of Simple Greedy Retention Tree ........................................ 52
3.2 Visual Representation of Delete and Rebuild Move ............................................. 56
3.3 CHAID Results: Partial Decision Tree Illustration (Test Set) ............................. 69
3.4 Best GRASP Retention Tree for RCL = 4 (first generated tree, Test Set) .............. 69
3.5 Pareto Analysis of Employee-Group Personnel Losses (Test Set) ......................... 71

B.1 Derived E(q,b) for Additive Case: $c_r/c_u = 0.5$ ............................................. 127
B.2 Simulated E(q,b) for Additive Case: $c_r/c_u = 0.5$ ............................................ 127
B.3 Derived E(q,b) for Additive Case: $c_r/c_u = 0.75$ ............................................. 127
B.4 Simulated E(q,b) for Additive Case: $c_r/c_u = 0.75$ ............................................ 127
B.5 E(w) for Additive Retention Case: $c_r/c_u = 0.25, \alpha = 30$ ............................ 128
B.6 E(w) for Additive Retention Case: $c_r/c_u = 0.75, \alpha = 30$ ............................ 128

C.1 Derived E(q,b) for Multiplicative Case: $c_r/c_u = 0.5$ ......................................... 134
C.2 Simulated E(q,b) for Multiplicative Case: $c_r/c_u = 0.5$ ..................................... 134
C.3 Derived E(q,b) for Multiplicative Case: $c_r/c_u = 0.75$ ......................................... 134
C.4 Simulated E(q,b) for Multiplicative Case: $c_r/c_u = 0.75$ ..................................... 134
C.5 E(w) for Multiplicative Retention Case: $c_r/c_u = 0.25, \alpha = 30$ ..................... 135
C.6 E(w) for Multiplicative Retention Case: $c_r/c_u = 0.75, \alpha = 30$ ..................... 135
Chapter 1

Introduction

Civilian and military manpower planning is usually conducted by human resource (HR) planners having little or no experience using and developing quantitative methods. This observation is based on the author’s extensive experience interacting with civilian and military HR planners. If an organization forecasts an increase or decrease in production, HR planners must respond with recommended changes to the organization’s employee footprint. Not only is it vital for planners to properly recommend the required changes in total personnel, but they must also be able to properly identify personnel needs by job skill and tenure. In these types of scenarios, the use of mathematical models can provide valuable support in manpower planning decisions, such as personnel recommendations based on the characteristics of a set of employees or issuance of incentive bonuses based on a company’s skill shortages.

In the manpower planning process, there are two important levers with which an organization can shape its workforce. These are recruitment of new employees and retention of incumbent personnel, and the organization can choose to use these individually or collectively. For example, a company can resort to recruiting new employees to augment or replace departing personnel, or it can choose to incentivize or de-incentivize incumbent personnel to stay. In this dissertation, we first study mathematical models that can assist the civilian and military personnel planner in balancing recruitment and retention in meeting personnel needs. Both hiring and retention incur costs; monetary compensation may be required to entice current employees to remain, and recruiting employees will almost certainly involve occupational training and reduced productivity
until this training is complete. Thus, in both military and civilian sectors, the choice as to whether to hire or retain can be of the utmost importance.

Effective manpower planning requires a thorough understanding of employee retention behavior, that is, the likelihood that employees will stay with the organization. In this dissertation, we therefore propose models that can assist HR planners in estimating the retention probabilities of employee-groups fitting particular profiles. Accurately estimating the retention propensity of a group of employees sharing a common profile is vital, because it helps a company prepare for the possible departure of employees fitting the profile, while also identifying possible reasons (via a set of attributes) behind their decisions.

The issues we outline above may require the use of large data sets, which can complicate real-time analysis efforts. Therefore, the final chapter of this dissertation presents methods for extracting a representative subset of employee data. The resulting data sets can simplify subsequent analysis efforts and enable real-time analysis of retention and recruitment decisions. The resulting models have applications beyond manpower planning and may provide a general framework for alleviating the burden of conducting analysis using large data sets in a variety of settings.

In the remainder of this chapter, we provide an introduction to each of the topics pertinent to this decision-making context.

1.1 The Issue of Balancing Recruitment and Retention in Military Personnel Planning

Current Army manpower planning includes little integration between recruitment and retention efforts of military personnel. One organization may determine recruiting needs using statistical methods or large-scale optimization models to forecast personnel losses, another may attempt to determine retention requirements based on historical averages, and a third may determine monetary bonuses offered to eligible personnel to entice them to stay in the Army. Bonus levels are determined in a rather ad-hoc manner and, overall, coordination between various divisions within the Army is
It is important to note that the hiring and firing of military personnel is not as fluid as it is in the corporate world since military personnel are subject to annualized contractual obligations, and vacancies are filled either by new recruits or reenlisting soldiers. Therefore, we propose a modeling framework that integrates recruitment and retention decisions for military personnel planning.

Specifically, in Chapter 2 we propose an analytical model that coordinates recruitment and retention decisions by jointly setting recruitment level and retention bonuses. This model can be viewed as a novel application of the newsvendor model with pricing, which has been studied extensively in operations management.

To evaluate our model, we conduct an empirical study on the determinants of reenlistment behavior of first-term, retention-eligible military personnel. We conduct analysis using a proprietary data set belonging to the U.S. Department of Defense (DOD). In particular, since we view monetary compensation as an important factor in our application of the newsvendor model, we study the response of military personnel to retention bonuses. Currently, the only personnel information the Army uses to determine future personnel strengths is pay grade, time in service, gender, and, to a small degree, a soldier’s military occupational specialty (MOS). In our analysis, we employ more detailed personnel characteristics to make inferences regarding employee retention behavior.

Counterfactual experiments based on our empirical results suggest that coordinating recruitment and retention bonus decisions might enable the military to better maintain its enlisted force and could generate substantial savings in personnel costs.

1.2 Two Methods of Predicting Employee-Group Retention Behavior

The most popular approaches to making inferences regarding employee retention are time series forecasting techniques and basic regression-type models. The use of decision trees to analyze employee retention behavior, however, remains largely unexplored, although they have been used to evaluate employee retention at the individual level. However, it has become increasingly important
for a company to be able to determine the retention propensity of a group of employees possessing a certain set of attributes. For example, a company may show interest in determining the retention probability of employees possessing the following attributes: 20-25 years old, female, mechanical engineer, working in the Midwest, located in the United States of America. Traditional decision tree methods are ill-equipped to estimate these types of employee-group retention probabilities given that their focus is estimating an individual’s retention probability.

In Chapter 3, we construct non-binary decision trees which we refer to as retention trees, using the principles of metaheuristics and optimize \( k \)-nearest neighbor weights (KNN) by using commercially available software. We present a greedy, randomized adaptive search (GRASP) retention tree approach as an improvement to a private optimization software company’s current decision tree method and propose a KNN approach for future implementation. We define non-mutually exclusive homogeneous groups based on employee attribute commonalities and seek the retention probabilities of these groups, whereas ordinary decision tree approaches seek to classify or determine probabilities of individual entities. Although one can extract individual employee probabilities from our defined groups, our models do not seek to optimally estimate individual retention decisions. Given the objective of our retention trees, they in fact perform poorly at estimating individual retention probabilities. Likewise, commercial software is designed to estimate and predict individual retention decisions, and performs poorly at estimating employee-group probabilities. The goal of our research is to estimate the retention probabilities of employees that possess a particular set of personal attributes. We provide a visual solution representation of the decision tree that is easily interpretable by either the HR planner or a company’s upper management. As previously mentioned, the objective of the methods we propose is not to optimally classify individual employees, but to generate decision trees and optimal nearest neighbor weights that produce the best employee-group retention probabilities that are employee-attribute based.

Whereas military personnel planning is constrained by numerous Department of Defense policies and an employee’s contractual obligation, civilian manpower can be more difficult to manage given its more liberal policies in fulfilling recruitment and retention targets. An empirical
study using a proprietary employee data set from a Fortune 500 company suggests that estimating employee-group probabilities with our retention tree and KNN methods would produce more accurate group probability estimates than commercially available software. Thus, our models empowers the human resource planner with the ability to understand and alter employee-group retention behavior.

1.3 Surrogate Data Representations for Feature Selection and Prediction

Often, the task of utilizing an extremely large employee database to make forecasts becomes extremely computationally expensive. To alleviate this computational burden and maintain a good level of classification accuracy, Chapter 4 introduces models to reduce the amount of data required while still providing accurate forecasting capabilities. Therefore, we introduce the concept of surrogate data representations to generate a representative subset $X^*$ of an original (larger) data set $X$, permitting $X^*$ to be used in place of $X$ in such data mining procedures as feature selection and classification. We first explain how we partition observations into bins termed “tranches” via our tranche-generation process.

We explain the concept of co-diversity and discuss our unique diversity-screening methods to enhance our modeling approach. Co-diversity is inversely related to correlation in that in our feature-selection model seeks to maximize co-diversity between features and minimize co-diversity between each feature and the response variable in a procedure analogous to those statistical methods that minimize correlation among independent variables. We exploit these co-diversity measures via a screening method, and given these measures, use them as inputs into a pair of novel binary quadratic programming models. The main purpose of our binary quadratic programming models is to support feature selection and forecasting. Employed together or individually, our surrogate data representations and feature selection methods can assist quantitative analysts in their forecasting efforts.

For HR planners with the task of forecasting employee retention behavior, utilizing employee
data and quantitative models to conduct analyses is critical. However, modeling employee behavior poses difficulties when the size of the personnel database is prohibitively large. Random sampling to extract a subset of the data constitutes one possible approach, but this type of sampling can lead to excessive loss of pertinent employee information. In an empirical study using U.S. Army retention data, our results show that the performance of our surrogate data sampling and feature-selection approaches perform better than random sampling and standard statistical feature-selection strategies. Also, use of classification and regression tree (CART) forecasting analysis indicates that our methods for predicting soldier retention propensity lose only a very small percentage of predictive accuracy. An important managerial implication of our results is the ability to provide, given a short time-line in which to complete real-time analysis, statistical and other modeling results surpassing those based on standard random sampling and traditional statistical approaches.
Chapter 2

Balancing Recruitment and Retention in Military Personnel Planning

2.1 Introduction

In the manpower planning process, there are two important levers a company can use to shape its workforce—recruitment and retention of personnel, and these can be used individually or collectively. For example, a company can recruit new employees to augment its existing workforce or replace departing personnel, or incentivize or de-incentivize incumbent personnel to stay.

In current Army manpower planning, there is little integration between recruiting and retention efforts of military personnel. One organization may determine recruiting needs using statistical methods or large-scale optimization models to forecast personnel losses, while another may attempt to determine retention requirements based on historical averages, and a third may determine monetary bonuses to offer so as to entice eligible personnel to stay in the Army. Bonus levels are determined in a rather ad-hoc manner, and overall coordination between the various organizations is limited.

In this chapter we first present mathematical models that can assist the civilian and military personnel planner in balancing recruitment and retention. In both the military and civilian sectors, how an organization handles a shortage of personnel in key occupational areas determines the type and level of costs incurred. Monetary compensation may be a required to retain current employees, and hiring new employees entails training costs and reduced productivity until completion of training. Also, the hiring and firing of military personnel is not as fluid as it is in the corporate world, since military personnel are subject to annualized contractual obligations, and vacancies can
be filled either by new recruits or reenlisting soldiers.

The main thrust of our research is to provide a modeling framework that integrates recruitment and retention decisions in military personnel planning. In the Army, lower ranking enlisted soldiers constitute a large proportion of the entire personnel inventory. For this reason, the Army is highly dependent on the lower enlisted ranks to accomplish unit missions. What differentiates a young, first-term soldier from a fresh military recruit is the valuable military experience of the former. In our opinion, at least three years of military experience is needed to train a new recruit to the same capability level as the first-term soldier. Thus, if the military offers bonus incentives to incumbent soldiers, not only will it retain soldiers capable of immediately contributing to their military units, but it will also save the Army personnel costs associated with training and lower productivity of new recruits. In this research, we find that monetary bonuses have a profound effect on retention behavior. Utilizing our approach to the Army will entail offering bonuses to incumbent soldiers at lower levels than historically offered during fiscal years 2007 through 2009, increasing retention rates by military occupation, and reducing recruitment levels to fill personnel gaps in the lower enlisted ranks. Our point is not to replace all recruits with young, first-term soldiers, but rather to evaluate the potential benefits of “balancing” recruitment and retention within a constrained military budget and across competing military programs.

A soldier’s decision to remain in the Army involves numerous factors. Some are personal, such as gender and family status, and others are non-personal in nature such as macro-economic state, current and eminent military conflicts, promotion opportunities and retention bonuses. While most non-personal factors cannot be easily altered, the military can readily change reenlistment bonuses to influence reenlistment decisions. Indeed, historically, retention bonuses have been used extensively by the military to maintain a target enlisted force.

To understand how reenlistment bonuses affect Army personnel’s reenlistment decisions, we first conduct an empirical analysis on a proprietary data set. Our data source includes historical bonus records, records of retention-eligible personnel, and recruiting and retention missions (personnel objectives) within a three year timeline. We employ a logistic regression model with covariates
including personal attributes, such as age, gender, family size, race, and military occupation, as well as non-personal attributes, such as U.S. employment rates and bonuses acting as covariates.

Our empirical analysis reveals a number of interesting insights. First, a soldier’s experience level may have a profound effect on his/her decision to stay or leave the Army. The less experience a soldier has, the less likely he will be to reenlist and to remain in the Army. Also, a non-intuitive result of our empirical analysis indicates that a soldier’s initial contract length (i.e., service term) is highly correlated with staying in the Army, that is, the longer the initial contract, the more likely the soldier reenlists. Therefore, to retain personnel that join the Army under long contract terms (i.e., greater than three years), they should be immediately targeted upon becoming reenlistment eligible. Of greater importance with respect to our research study, we find that soldiers’ reenlistment decisions are highly sensitive to bonus offers.

In the analytical model we propose, recruitment and retention decisions are coordinated by jointly setting retention bonus amount and recruitment levels. Our model is similar to the widely studied newsvendor model with pricing (Petruzzi and Dada, 1999), with the bonus amount corresponding to the pricing decision. Our model does have several unique characteristics, however. First, the target strength of the military force is fixed. Instead, uncertainty in our model stems from the uncertain responses of Army personnel to retention bonuses. Second, there is no overage cost as standard military practice dictates that, once all demand is fulfilled, no further reenlistment requests are accepted.

We model military personnel’s reenlistment decisions as a Bernoulli process for given bonus amounts. Therefore, at a given bonus level, the total number of reenlisted personnel follows a binomial distribution, since the number of military personnel eligible for reenlistment is fixed. We then use a normal distribution to approximate this binomial distribution.

A counterfactual experiment reveals that applying the newsvendor problem approach to manpower planning may contribute to significant manpower cost savings when compared to current manpower planning tools utilized by the U.S. Army. Experimental results show that clustering military personnel in groups based on service term contract, marital status, experience level, num-
The remainder of this chapter is organized as follows. Section 2.2 presents a brief literature review. Section 2.3 presents some background on the U.S. Army’s enlisted personnel system and explains how we cleaned and merged several data sources. Section 3.5.3 describes our retention rate empirical analysis using logistic regression. Section 2.5 proposes an analytical model for joint retention bonus and recruitment decisions. Section 2.6 reports our numerical results, and Section 2.7 presents our conclusions.

2.2 Literature Review

Our research follows two primary streams within the literature, one relates to the newsvendor problem with pricing and the other relates to military personnel planning. For the former, Chan et al. (2004) provide a comprehensive review of coordination of pricing and inventory decisions. The authors classify inventory models by such characteristics as horizon length, prices, demand type, demand function form, and capacity limits. Khouja (1999) discusses single-period stochastic inventory (i.e., newsvendor) problems and presents a very detailed look at their history and various extensions. Goh et al. (1993) present a two-stage perishable inventory model, while Raafat (1991) provides a survey on deteriorating inventory models. Beyond single-period and multi-period inventory models, Cheung and Yuan (2003) show an infinite horizon inventory model with commitment in which the buyer is required to make a periodic order commitment according to general discrete demand distributions. Lastly, the recent paper by Li and Zhang (2013) provides important insights into the concept of “pre-order” strategies for a perishable product. In the pre-order scenario, the authors explain that offering items for pre-order benefit both seller and buyer. One assures the buyer of receiving the items on time, while the seller can use pre-order quantities to determine demand when actual demand is known.

The model we introduce in this chapter closely relates to the work Petruzzi and Dada (1999)
present. In their seminal paper, the authors adopt as the overall goal of their model to maximize the expected profit of selling a single product. To set the stage for this optimization problem, a decision maker determines how much perishable product to stock for some pre-determined selling period. Unlike previous efforts, the authors consider demand and selling price as endogenous parameters. By setting stocking item quantity and selling price simultaneously, Petruzzi and Dada (1999) formulate the newsvendor problem in a novel way. The authors first consider a price-setting firm that stocks a single product, faces a random price-dependent demand function, and has as its objective determining jointly a stocking quantity, \( q \), and selling price, \( p \), to maximize expected profit. The authors state that randomness in demand is price independent and can be modeled either in an additive or a multiplicative fashion. Petruzzi and Dada (1999) also provide valuable theoretical insights into the structure of the problem which we employ in constructing our own inventory modeling approach. Our approach, however, excludes overage costs since we assume that, once we achieve a target number of personnel (inventory) for a particular time period, we no longer permit more recruits and reenlistments.

As for military personnel planning, Gass (1991) provides readers with a comprehensive survey of military manpower models. Eiger et al. (1988) provide an excellent description of what is essentially the Army’s entire current Enlisted Specialty Model (ES). This annualized linear program incorporates weights on deviations in the objective function. Corbett (1995) formulates a linear programming model to manage the Army’s Accession and Branch detail program. Rodgers (1991) proposes a multi-objective linear program for Navy enlisted personnel that forecasts monthly strengths, promotions, and accession goals over a short time horizon. Liang and Lee (1985) describe the Navy manpower system and explain the numerous complexities involved in assignment of personnel. Albrecht (1979) investigates the effects of labor substitution in the military. Cai et al. (2010) present a mathematical model to determine the optimal numbers of employees in different categories to meet uncertain demands for manpower having certain skills. Lastly, Mehlmann (1980) applies dynamic programming (DP) to determine optimal recruitment and transition decisions in a civilian manpower system, while Morton and Popova (2004) formulate a Bayesian Stochastic Pro-
programming (SP) model to apply to an employee scheduling problem. The papers described above illustrate the use of the primary mathematical models that apply to solving the military manpower problem (MMP). To the best of our knowledge, our research is the first to present a newsvendor problem approach to the MMP that integrates recruiting and retention.

Empirical studies in the area of retention modeling have taken on many forms, and we highlight a few contributions below. Sheridan (1992) studies the retention rates of college graduates and discovers that variation in cultural values significantly affect the rates at which newly hired employees voluntarily terminate employment. Koys (2001) hypothesizes that employee satisfaction, organizational citizenship behavior, and employee turnover influence profitability and customer satisfaction. Similarly, Tansky and Cohen (2001) conduct an empirical study at a major Midwestern hospital and found that organizational commitment and perceived organizational support are correlated with satisfaction with career development. The study by Guthrie (2001) found a positive association between the use of high-involvement work practices and employee retention and firm productivity. The author comments that employees' positive perceptions and strong correlations between working conditions and mission attachment suggest that mission can still play a significant role in employee retention by reducing dissatisfaction with pay and career advancement. Recently, Murphy et al. (2013) examines the relationships between perceptions of job insecurity, job embeddedness, and important individual work outcomes. The interesting results of their longitudinal study suggest that perceptions of job embeddedness fully mediate the relationship between perceptions of job insecurity and intention to remain.

Motivated by the research from the authors we mention in this literature review, this research proposes a framework that uses a two-stage newsvendor type approach to solving the joint recruitment and retention problem. First, we provide some background on our data set and then we present the results of our empirical analysis on retention rates.
2.3 Background and Data Source

2.3.1 Data Extraction

We consider three years of historical records from fiscal years 2007 through 2009 obtained from Headquarters, U.S. Army, Deputy Chief of Staff, Army G-1 (Personnel). The data is part of the Total Army Personnel Database (TAPDB). A schematic view of the TAPDB data structure is shown in Figure 2.1. The three data files that contain the most relevant data for our analysis are the inventory file, the transaction data file, and the reenlistment bonus data file.

The inventory file contains snapshots of the active Army enlisted force taken at monthly intervals. Each monthly snapshot contains a record for every currently enlisted soldier by social security number (SSN) at month’s end. The number of enlisted soldiers in the Army at any given point in time varies; during the historical period of our analysis, the size of the enlisted force fluctuated between 420,000 and 470,000 soldiers. Each record within this monthly data has 255 data fields (columns) containing a variety of data ranging from soldier attributes (e.g., gender, dependents, MOS, enlistment date, and civilian education level) to descriptive states (organization, organization type, location, etc.). As one can imagine, a database with a single monthly extract of approximately 450,000 records and 255 fields that is then multiplied by 36 months is enormous in size. The initially inventory file we construct includes approximately 16 million records.

In another file similar to the personnel inventory file, the Army compiles and stores transaction data in monthly increments. Each monthly segment compiled at the end of the month includes a record for any major personnel transaction that occurred during that month. There were 2,883,595 transaction records created in the three-year period we analyzed. Ultimately, we retain the transaction fields SSN, date, and loss type for use in our study. The loss type field reflects the type of personnel loss to the Army which includes voluntary and involuntary losses. The number of transaction records is later reduced significantly by filtering the transaction data for loss type equal to loss due to immediate reenlistment (LIMR) and loss due to expiration term of service (LETS). LIMRs are not true losses, since a soldier renewing his/her contract automatically
Figure 2.1: Data Structure of the Total Army Personnel Database
generates an immediate loss transaction and then an immediate gain transaction. Actual losses
due to expiration term of service (ETS) are the only type of losses relevant to our analysis because
the other loss types are either involuntary or related to some condition that precludes reenlistment,
such as retirement.

As part of our data processing, we conducted further data cleanup, which Appendix A ex-
plains in detail. We now briefly highlight the main data cleanup tasks. First, we limited our analysis
to soldiers that reenlist for four years. In order to better analyze retention behavior, we narrowed
our analysis to four-year reenlistees because soldiers behave differently based on their chosen reen-
listment term. Given this limitation, we randomly selected a sample of personnel that voluntarily
departed the military. The size of this sample is based on the proportion of four-year reenlistees
compared to the entire reenlistment population. Second, since there are a significant number of
soldier occupations (over 150), we limited our analysis to the most popular military occupation,
the infantrymen. Finally, the Army allows soldiers to reenlist under five different retention options.
So, to further refine our analysis, we only considered infantrymen that reenlisted under the most
popular option, station stabilization. Given this refined data set, we proceed to assigning them
reenlistment bonus values.

Determining whether a bonus affects a soldier’s decision to reenlist or leave, as opposed to
merely analyzing its effect on reenlistees, is crucial to our analysis. Reenlistment bonus data is the
most difficult to obtain and reconstruct. Throughout this chapter, we use the terms “ETS’ed” and
“ETS’ing” to denote a soldier that had completed or would complete his/her initial contractual
obligation (no reenlistment). For the period of our analysis, we obtained historical bonus data for
reenlistees from Army Human Resource Command (HRC). This data set includes bonuses that
were awarded to soldiers that reenlisted. To this end, we assigned a bonus amount to eligible Army
personnel based on observed historical bonus data for Army personnel that chose not to reenlist.
In our study, we assigned an ETS’ing soldier a reduced bonus value based on the calculated bonus
rates for his/her particular MOS whereas reenlisting personnel are assigned the bonus values they
actually received.
Analysis of the bonus data reveals some interesting observations regarding the behavior of a soldier’s response to bonus offers. First, a soldier does not always reenlist when he/she becomes reenlistment eligible. Second, the longer a soldier waits to make his decision the more likely he/she will encounter several different bonus offers that may increase or decrease. The reason a soldier may not make an immediate reenlistment decision is that he/she may be indecisive about signing a new service contract. Or, a soldier is “gaming” the Army retention system and is hoping for a higher bonus offer during his/her retention eligibility. Last, not all reenlisting soldiers receive a maximum bonus offer upon their decision to stay in the Army.

Given the observations mentioned above, determining the maximum bonus offer for use in our analysis is based on historical bonuses awarded by Military Occupation Specialty (MOS) for each month between October 2006 and September 2009. To assign these maximum bonus amounts to a soldier within a particular MOS that departed the military within a specific month, we first determined the maximum bonus offers by month during the soldier’s previous 3- to 24-month retention window, which represents the number of months prior his/her ETS (end of contract). By Army regulation, a soldier can only reenlist during this retention window. For example, if several cannon crewmembers (13B) accepted bonus offers between $2,000 to $20,000 for a particular month, the servicemembers are assigned a value of $20,000 for the month in question in our research. We then determined the maximum of the “maximum bonus” offers to an ETS’ing soldier during the span of his/her reenlistment window of opportunity. Given this calculated maximum bonus amount, we proceed to the next step.

Subsequently, we determine a bonus rate for each MOS to ascertain, on average, what percentage of the maximum bonus the reenlisted personnel took during the time they became retention eligible, up to the month they reenlisted. For example, if a soldier reenlisted 20 months prior to his/her ETS and accepted a bonus of $15,000, he/she in effect “declined” bonus offers of $20,000, $15,000, $7,500 and $4,500 for months 21 through 24, and the bonus rate we assign to this soldier is .75. By accumulating all the soldier bonus rates for all three years of historical data, we determine average bonus factors for each MOS. It is important to note that these bonus rates are
applied to job-specific maximum bonus amounts assigned to departing personnel. Applying this rate ensures that overly high bonus amounts are not assigned to personnel that departed and thus severely skew our regression analysis to determine the appropriate retention rate function for our recruiting-retention model.

One key input into our recruitment and retention model is determining the retention rate of incumbent first-term soldiers. Before introducing our recruitment and retention model, we first explain how we derive the retention rate function, \( r(b) \), the most important input into our model. Therefore, the next section explains in detail, the analysis we conducted to derive the retention rate equation most appropriate for input into our model. An introduction to our recruitment and retention model follows this retention rate empirical analysis.

2.4 An Empirical Study of Retention Rates

2.4.1 Variable definitions

As previously noted, we consider numerous variables for inclusion in our empirical analysis. The inventory and transaction tables shown in Appendix A provide a snapshot of the personal attributes obtained for each enlisted soldier in the Army by combining variables from all data sources. These variables are deemed potentially relevant based on the author’s first-hand experience in working with military manpower models. These tables provide the reader with a concise description of key variables, which we later redefine to make them more easily accessible to the reader during the description of our analytical approach which follows. We further elaborate on these variables in the following paragraphs.

Additional endogenous and exogenous variables, such as age, civilian education level, and employment rate of change, are added. For modeling purposes, we convert several explanatory variables into binary form. Table 2.1 shows these newly converted and defined variables, which we now discuss.

Following completion of our preliminary analysis, we discovered that several variables are
not statistically significant for our analysis. Since a large proportion of first-term soldiers are predominantly high school graduates, we found CIV ED QY (which refers to a soldier’s civilian education level) uninformative and excluded it from our analysis. We convert the REDCAT CD (which refers to a soldier’s racial background) to a binary variable that identifies a soldier as belonging to the WHITE race or non-WHITE race. Our initial regression analysis deems the WHITE variable as statistically insignificant to our work, and so we exclude it before conducting the final regression analysis. Also, METS QY (which refers to a soldier’s months until expiration term of service) is excluded since its value is almost identical in definition to AOS (which refers to a soldier’s additional obligation of service). For example, if a soldier reenlists for four years but has six months remaining on his/her contract, the soldier’s METS QY is six months, while his/her AOS value is 3.5 years (four years less six months).

Additionally, we define variables related to soldier age as binary variables as well. We define AGEGROUP22 to include personnel with ages of 18 through 22 ($18 \leq age \leq 22$), AGEGROUP25 to include personnel with ages of 23 through 25 ($22 < age \leq 25$), and AGEGROUP30 to include personnel with ages of 26 through 30 ($25 < age \leq 30$). Again, we exclude the extremely small population of personnel beyond the age of 30 since we are more concerned with young reenlisting soldiers and their ability to replace a young recruit. Similar to the WHITE variable, our preliminary analysis shows that a person’s age is not significant since most reenlistment eligible soldiers are below the age of 22.

For the minor dependents variable, we again employ binary variables to account for the different numbers of children. MINORDEP1 denotes a soldier having one child, MINORDEP2 a soldier having two children, and MINORDEP3 a soldier having three or more children. Soldiers having no children serves as the reference variable in our empirical analysis. We construct the EMPLOYMENT exogenous variable from data obtained from the Bureau of Labor Statistics (2013). Using its monthly data, we calculate a twelve-month moving average of the monthly change in employment (number of new jobs created). These employment percentages are multiplied by one hundred to make better inferences on the impact of small changes to the rate as it relates to a
soldier’s propensity to remain in the Army. Thus, a one-unit change in this variable equates to a 1/100 percent change in employment rate.

Moreover, we further filter and test the variables listed in Table A.1 by further excluding variables we deem irrelevant for the purpose of this analysis based on the author’s experience working with enlisted personnel data. Thus, such variables as TOE_UNIT (mission or training focus unit) and AFQT_PCT_QY (Armed Forces Qualification Test Score) are excluded since they are highly correlated with the primary MOS variable. Table 2.1 presents the average values and ranges of our final modeling variables.

Table 2.1: Summary Statistics of Predictors for Infantry Career Field (Four Year Reenlistment, Station Stabilization Option)

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAX_BONUS</td>
<td>7.5872</td>
<td>3.5124</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>SERVICE_TERM</td>
<td>3.5458</td>
<td>0.7457</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>AOS</td>
<td>34.118</td>
<td>6.0641</td>
<td>24</td>
<td>47</td>
</tr>
<tr>
<td>MARRIED</td>
<td>0.3186</td>
<td>0.4661</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>EXPER</td>
<td>3.4216</td>
<td>0.9692</td>
<td>.4</td>
<td>7.3</td>
</tr>
<tr>
<td>MINORDEP0</td>
<td>0.7851</td>
<td>0.4109</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MINORDEP1</td>
<td>0.1478</td>
<td>0.355</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MINORDEP2</td>
<td>0.0508</td>
<td>0.2198</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>MINORDEP3</td>
<td>0.0163</td>
<td>0.1266</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>WHITE</td>
<td>0.7722</td>
<td>0.4196</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>EMPLOYMENT</td>
<td>-2.1763</td>
<td>16.2515</td>
<td>-36</td>
<td>18</td>
</tr>
</tbody>
</table>

2.4.2 Logistic Regression Results

We employ a logistic regression to analyze Army personnel reenlistment decisions (stay or leave). The literature adopts a similar approach; see, e.g., Stevens (2009). To set up the logistic regression, we scale down our bonus variable by dividing the raw values by 1,000. For the purpose of our analysis, we use a total of 10 covariates (interval and categorical) and, as stated before, limit our regression model to the infantryman career field. Using all the key variables in the model, we fit a binary logistic regression model to our data as follows. First, we designate the variables MARRIED,
WHITE, MINORDEP1, MINORDEP2, and MINORDEP3 as categorical variables using 0 as the reference parameter value. Second, we designate a stepwise variable selection procedure, set REUP as the binary response variable, cap iterations at 1,000, and set the model convergence parameter to 1E-8.

Our variable-selection procedure considers a large number of variables for inclusion in the model. However, using a \( p \)-value of 0.10 as the cutoff for a variable to remain in the final regression analysis eliminates many of these. Also, employing only one occupation in our regression analysis eliminates the necessity of including an enormous number of categorical variables in the regression model. Prior to the running of the model, possible variable selection techniques include forward selection, backward elimination, stepwise (combination of backward and forward selection), and best subset selection. Again, we use a stepwise variable selection to construct our regression equation since, as Der and Everitt (2001) state, “In the best of all worlds the final model selected by each of these procedures would be the same.” We conduct numerous experiments using a combination of one to eleven covariates in the model to determine its most logical form. Again, given the enormous number of modeling choices, we follow the recommended steps set forth by Shtatland et al. (2007) to arrive at the best form of the regression equation given that our model consists of categorical and continuous variables. The steps are as follows: construct a full stepwise sequence, minimize information criteria (IC) on the full stepwise sequence, and specify critical \( p \)-values based on IC for stepwise procedures.

Using the stepwise variable selection process, we evaluate variables entering the regression equation based on their \( p \)-values (difference from zero test) in the Chi-squared test. Even if a variable “leaves” the model during the selection process, it is eligible for consideration to re-enter the regression; predictors could routinely enter and leave the regression equation several times during the iteration process. As for the model’s convergence status, the maximum-likelihood algorithm successfully achieves convergence at the 1E-8 convergence criterion. As previously stated, we omit the WHITE (racial profile) and MINORDEP1 variables and so do not appear in the final form of the regression equation. Also, the most significant value for the number of minor dependents
variable is three or more children. The variables we include in Table 2.1 but not appearing in
the regression equation were found to represent attributes not statistically significant at the 0.10
\( p \)-value level for the infantryman occupation.

Before presenting the final form of the logistic regression for our retention rate calculations,
we first interpret the standard statistical test results. Table 2.2 displays the Akaike Information
Criterion (AIC), Schwarz Criterion (SC), and Log-likelihood statistics \((-2 \text{LogL})\), all of which
indicate that the model including both intercept and covariates performs better (i.e., has lower
model fit statistics) than the one fitted to have only an intercept. This table shows that our model
with its chosen attributes is an appropriate and useful tool to help determine retention propensity.

Table 2.2: Model Fit Statistics

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Intercept Only</th>
<th>Intercept &amp; Covariates</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>1954.503</td>
<td>859.615</td>
</tr>
<tr>
<td>SC</td>
<td>1959.800</td>
<td>907.282</td>
</tr>
<tr>
<td>-2 Log L</td>
<td>1952.503</td>
<td>841.615</td>
</tr>
</tbody>
</table>

Table 2.3 shows the statistical test results with respect to the null hypothesis that at least
one of the regression coefficients is not equal to zero. All test results suggest that, at a \( p \)-value
of 0.10 we reject the null hypothesis that at least one regression coefficient equaled zero. The
Chi-squared test statistics shown in Table 2.3 are associated with extremely small \( p \)-values,
which reflect the probability of obtaining those test statistic values given that the null hypothesis is true.
Thus, these results indicate that at least one model coefficient is not equal to zero. Furthermore,
Table 2.4 presents the maximum likelihood estimates (MLE) of the parameters associated with
variables remaining in the Logit model. As is evident from the table, having one child (i.e., the
variable MINORDEP1) is not statistically significant at the 10% level. Given this result, one can
conclude that having one child is equivalent to having no children at all, since we use the "no
children" option as the reference value in our logistic regression analysis.

Table 2.5 shows the odds ratio for the different covariates. These are obtained by exponenti-
ating the MLE estimates. As a reminder, the difference in the logs of two odds is equal to the log of the ratio of these two odds. Thus, one can interpret the odds ratio as follows: For a one-unit change in the independent variable, the odds ratio for a positive outcome is expected to change by the respective coefficient, given that the other variables in the model are held constant (Lewis, 2007). In the next section, we provide our insights into the meaning of the odds ratio estimates.

Table 2.3: Testing Global Null Hypothesis: Beta=0

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio</td>
<td>1110.887</td>
<td>8</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Score</td>
<td>808.118</td>
<td>8</td>
<td>&lt; .0001</td>
</tr>
<tr>
<td>Wald</td>
<td>390.8247</td>
<td>8</td>
<td>&lt; .0001</td>
</tr>
</tbody>
</table>

Table 2.4: Analysis of Maximum Likelihood Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>DF</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>Wald Chi-Square</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>-4.6369</td>
<td>0.7760</td>
<td>35.7046</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>MAX_BONUS</td>
<td>1</td>
<td>0.4353</td>
<td>0.0401</td>
<td>117.7771</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>SVC_TERM</td>
<td>1</td>
<td>2.899</td>
<td>0.1915</td>
<td>229.2321</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>AOS</td>
<td>1</td>
<td>0.1222</td>
<td>0.0161</td>
<td>57.3868</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>MINORDEP1</td>
<td>1</td>
<td>0.0742</td>
<td>0.2725</td>
<td>0.0741</td>
<td>0.7854</td>
</tr>
<tr>
<td>MINORDEP2</td>
<td>1</td>
<td>0.7787</td>
<td>0.4344</td>
<td>3.213</td>
<td>0.0731</td>
</tr>
<tr>
<td>MINORDEP3</td>
<td>1</td>
<td>1.5813</td>
<td>0.8785</td>
<td>3.2398</td>
<td>0.0719</td>
</tr>
<tr>
<td>EXPER</td>
<td>1</td>
<td>-3.6621</td>
<td>0.1879</td>
<td>379.7407</td>
<td>&lt; 0.0001</td>
</tr>
<tr>
<td>MARRIED</td>
<td>1</td>
<td>0.4691</td>
<td>0.2189</td>
<td>4.5928</td>
<td>0.0321</td>
</tr>
</tbody>
</table>

Following Table 2.4, the final logit model to determine retention rates is given by

\[
\text{logit}(Y) = -(4.6369) + (0.4353) \text{MAX\_BONUS} + (2.899) \text{SVC\_TERM} \\
+ (0.1222) \text{AOS} + (0.7787) \text{MINORDEP2} + (1.5813) \text{MINORDEP3} \\
- (3.6621) \text{EXPER} + (0.4691) \text{MARRIED}. \tag{2.1}
\]
Table 2.5: Odds Ratio Estimates

<table>
<thead>
<tr>
<th>Effect</th>
<th>Point Estimate</th>
<th>95% Wald Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAX_BONUS</td>
<td>1.545</td>
<td>1.429 1.672</td>
</tr>
<tr>
<td>SVC_TERM</td>
<td>18.156</td>
<td>12.475 26.424</td>
</tr>
<tr>
<td>AOS</td>
<td>1.130</td>
<td>1.095 1.166</td>
</tr>
<tr>
<td>MINORDEP2 (1 vs 0)</td>
<td>2.179</td>
<td>0.930 5.105</td>
</tr>
<tr>
<td>MINORDEP3 (1 vs 0)</td>
<td>4.861</td>
<td>0.869 27.201</td>
</tr>
<tr>
<td>EXPER</td>
<td>0.026</td>
<td>0.018 0.037</td>
</tr>
<tr>
<td>MARRIED (1 vs 0)</td>
<td>1.598</td>
<td>1.041 2.455</td>
</tr>
</tbody>
</table>

2.4.3 Observations and Insights

The results of the logistic regression in Table 2.5 contain both intuitive and counter-intuitive results with respect to the infantry soldiers we consider in our study. First, these results indicate that a soldier offered a $1,000 bonus is 1.545 times more likely to remain in the Army than one offered no bonus. Interestingly enough, a married soldier is 1.598 times more likely to remain in the Army than an unmarried soldier. Not surprising is that a young, first-term infantry soldier with three children is 4.861 times more likely to stay in the Army when compared to a soldier having no children. As previously mentioned, we exclude the MINORDEP1 variable from future analysis given its extremely high $p$-value. The most statistically significant variable is SVC_TERM; the longer a soldier’s initial contract, the better are the odds that the soldier will remain in the Army.

Quite surprisingly, the federal employment variable suggests it has little bearing on a soldier’s decision to stay in the Army and is not significant at the 10% alpha level, suggesting that if one uses the number of new jobs created every month to gauge the state of the economy, the creation or reduction in national employment opportunities does not influence an infantryman’s retention decision. Based on personal work experience in military personnel planning prior to conducting this empirical analysis, we ascertain that a soldier is more likely to stay in the Army if the new contract term offer is the minimum (24 months is the minimum value). Thus, soldiers are more willing to remain in the Army if they are offered shorter contract offers. However, it is surprising that our analysis infers that a soldier is more likely to reenlist when he signs a contract having a
term much greater than 24 months as suggested by the AOS variable’s odds ratio of 1.130. Thus, as his/her “net” contract increases, the probability he or she stays in the Army also increases. This result contradicts our perception that a soldier is willing to stay in the Army if offered a shorter contract length. If our initial perceptions were correct, the AOS odds ratio should be less than 1. To provide a broader view of the impact the regression coefficients have on retention propensity, we will now present some hypothetical soldier attribute “scenarios.”

One can see the varying retention probabilities we produce by the different model scenarios shown in Table 2.6. As is evident from the table, retention rates change considerably as the soldier’s attributes in the regression model change. As bonus offers change, retention rates increase or decrease depending on such attributes as marital status and number of children. These different scenarios highlight the importance of partitioning personnel into groups to better predict retention propensity. Again, the scenarios show how the probabilities differ with increases in experience level (retention rate decreases), increases in initial service term contract (retention rates increase), and, more importantly, the varying bonus levels, which retention rates are extremely sensitive. The bonus attribute is clearly relevant in the time our data originated, and we surmise that it could become of greater value during times of more stringent DOD fiscal policies. In the near future, this empirical analysis will continue to provide significant insights into the influence of bonuses on future soldiers’ retention behavior. A budget-constrained conservative scenario occurred in fiscal year 2014 and is currently occurring in fiscal year 2015, given the steep reduction in Army personnel levels and DOD program cuts. In the next section, after deriving the retention rate equation (Equation 2.1) for use in our recruiting-retention model, we define our binomial model parameters and assumptions and begin our empirical study.
<table>
<thead>
<tr>
<th>Scenario Description</th>
<th>Bonus ($1000s)</th>
<th>Service Term (years)</th>
<th>AOS (months)</th>
<th>Minor Dep2</th>
<th>Minor Dep3</th>
<th>Experience (years)</th>
<th>Married</th>
<th>Retention Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmarried, 3 years experience, 2 children, 4-year contract, and no bonus</td>
<td>0</td>
<td>4</td>
<td>24</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>42</td>
</tr>
<tr>
<td>Unmarried, 3 years experience, 2 children, 4-year contract, and low bonus</td>
<td>5</td>
<td>4</td>
<td>24</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>87</td>
</tr>
<tr>
<td>Married, 2.5 years experience, 3 children, 3-year contract, and no bonus</td>
<td>0</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>1</td>
<td>2.5</td>
<td>1</td>
<td>65</td>
</tr>
<tr>
<td>Married, 2.5 years experience, 3 children, 3-year contract, and medium bonus</td>
<td>8</td>
<td>3</td>
<td>30</td>
<td>0</td>
<td>1</td>
<td>2.5</td>
<td>1</td>
<td>98</td>
</tr>
<tr>
<td>Married, 4 years experience, no children, 42 months of AOS, and no bonus</td>
<td>0</td>
<td>4</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>3.5</td>
<td>1</td>
<td>44</td>
</tr>
<tr>
<td>Married, 4 years experience, no children, 42 months of AOS, and medium bonus</td>
<td>10</td>
<td>4</td>
<td>42</td>
<td>0</td>
<td>0</td>
<td>3.5</td>
<td>1</td>
<td>98</td>
</tr>
<tr>
<td>Unmarried, 4.5 years experience, 2 children, 5-year contract, and no bonus</td>
<td>0</td>
<td>5</td>
<td>30</td>
<td>1</td>
<td>0</td>
<td>4.5</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Unmarried, 4.5 years experience, 2 children, 5-year contract, and high bonus</td>
<td>25</td>
<td>5</td>
<td>30</td>
<td>1</td>
<td>0</td>
<td>4.5</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td>Unmarried, 5 years experience, 1 child, 6-year contract, and low bonus</td>
<td>5</td>
<td>6</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>92</td>
</tr>
<tr>
<td>Married, 5 years experience, 1 child, 6-year contract, and low bonus</td>
<td>5</td>
<td>6</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>95</td>
</tr>
</tbody>
</table>
2.5 A Retention and Recruitment Model

This section introduces an analytical model that balances recruitment and retention decisions. Our empirical study in Section 3.5.3 shows that military personnel’s reenlistment decisions can be strongly influenced by reenlistment bonuses. A vacancy in the armed forces can be filled either by a new recruit or a reenlisted soldier. We consider a model that minimizes total personnel cost by choosing the total number of new recruits and the bonus offer for reenlistment-eligible soldiers. Before we formally introduce the model, we list the model parameters below.

\[ Q \] target inventory (demand)
\[ Q_0 \] retention-eligible inventory during the period
\[ q \] number of recruits enlisting in the Army
\[ b \] amount of bonus offered
\[ \chi(b) \] stochastic retention yield given bonus amount \( b \)
\[ c_r \] cost of recruiting per soldier
\[ r(b) \] retention rate
\[ c_u \] penalty cost of having less personnel than mission requirements (demand)

We consider a single-period model from the Army’s perspective, where a period can be interpreted as a fiscal year. We assume that the number of retention-eligible personnel at the beginning of the period is \( Q_0 \), while end-of-period target strength we define as \( Q \). The objective of the model is to minimize total expected cost by jointly determining a recruiting quantity, \( q \), and a bonus level, \( b \). A central component of our model is the number of people that reenlist given a bonus amount \( b \), which we represent using a random variable \( \chi(b) \).

Given a pair \((q,b)\), the total expected cost is given by

\[
\Pi(q,b) = E[c_u(Q - q - \chi(b))^+ + b \min\{Q - q, \chi(b)\}] + c_r q.
\] (2.2)
In the above, the expectation is taken with respect to $\chi(b)$. Expression (2.2) is closely related to the widely studied newsvendor problem with pricing (Petruzzi and Dada, 1999). Unlike the standard newsvendor problem, we do not include overage costs since we assume that once we achieve the target number of personnel, we no longer permit additional recruits and reenlistees. We assume linear recruitment cost, so that $q$ new recruits can be obtained at a cost of $c_r q$. The cost function, $\Pi(q, b)$, includes recruiting cost, retention bonuses rendered, and underage costs, $c_u$, where $c_u$ is assumed to be greater than the cost of recruiting, $c_r$, so as to avoid a trivial solution.

One important assumption in our model is that we can essentially substitute a new recruit for any type of non-attrition related Army personnel loss. This is a reasonable assumption because of the Army promotion system. To give an example, suppose a Sergeant Major (the highest enlisted rank) retires in the month of July. Then, August will see a promotion chain effect. One Master Sergeant promoted to replace the SGM, one Sergeant First Class promoted to replace the Master Sergeant, and so on and so forth (AR 600-8-19, 2010). The Army promotion system is built in such a way that the higher enlisted ranks always maintain approximately 100% strength. Hence, at the end of the promotion system process, we are left with lower enlisted ranks with no replacement. This is where our model comes into play, attempting to maintain a high fill rate for the lower ranks by offering bonuses to incumbent first-term soldiers (FTMs) and recruiting new soldiers. Additionally, we only consider FTM retention-eligible soldiers in our model when we “balance” recruiting and retention.

We derive and analyze additive and multiplicative random yield approaches to our problem (see Appendices B and C), but we deem a binomial random yield methodology as the most appropriate approach given the nature of military recruiting and retention.

A Binomial Model for $\chi(b)$. For a given bonus amount $b$, suppose the reenlistment probability for each eligible military personnel is $r(b)$. Then, the total number of reenlistments $\chi(b)$ follows a binomial distribution with parameters $Q_0$ and $r(b)$. Since $Q_0$ is typically quite large, we approximate $\chi(b)$ using a normal random variable with mean $Q_0 r(b)$ and standard deviation
\[ \sqrt{Q_0r(b)(1-r(b))}; \text{i.e.,} \]

\[ \chi(b) = Q_0r(b) + \sqrt{Q_0r(b)(1-r(b))}\epsilon, \]  

(2.3)

where \( \epsilon \) is a standard normal random variable.

Plugging (2.3) into (2.2) leads to

\[ \Pi(q,b) = E[c_u(Q - q - Q_0r(b) - \sqrt{Q_0r(b)(1-r(b))}\epsilon)^+ \]

\[ + b \min\{Q - q, Q_0r(b) + \sqrt{Q_0r(b)(1-r(b))}\epsilon\} + c_r q. \]  

(2.2)

Transforming \( q \) using the variable \( z = (Q - q - Q_0r(b))/\sqrt{Q_0r(b)(1-r(b))} \), the expected cost function becomes the following:

\[ \Pi(z,b) = (c_u - b) \int_{-\infty}^{z} (\sqrt{Q_0r(b)(1-r(b))}(z-u))f(u)du \]

\[ + (c_r - b)\left(Q - Q_0r(b) - z\sqrt{Q_0r(b)(1-r(b))}\right) + bQ. \]  

(2.4)

Using Leibniz’s rule and the chain rule, we arrive at the following first order equation with respect to \( z \) and solve for \( z \):

\[ \frac{\partial\Pi(z,b)}{\partial z} = (c_u - b)\sqrt{Q_0r(b)(1-r(b))}F(z) - (c_r - b)\sqrt{Q_0r(b)(1-r(b))} \]

\[ 0 = (c_u - b)\sqrt{Q_0r(b)(1-r(b))}F(z) - (c_r - b)\sqrt{Q_0r(b)(1-r(b))} \]

\[ 0 = (c_u - b) F(z) - (c_r - b) \]

\[ z = F^{-1}\left(\frac{c_r - b}{c_u - b}\right). \]  

(2.5)

Next, by defining \( \Gamma(z) = \int_A^z [z-u] f(u)du \), we derive the first order equation with respect to \( b \):

\[ \frac{\partial\Pi(z,b)}{\partial b} = \Gamma(z)\left(-1 \cdot \sqrt{Q_0r(b)(1-r(b))} + \frac{(c_u - b)}{2\sqrt{Q_0r(b)(1-r(b))}} \right). \]

\[ Q_0 \frac{dr(b)}{db}\left(1 - 2r(b)\right) - (Q - Q_0r(b) - z\sqrt{Q_0r(b)(1-r(b))}) + \]

\[ (c_r - b)\left(-Q_0 \frac{dr(b)}{db} - \frac{z}{2\sqrt{Q_0r(b)(1-r(b))}}\right) \cdot (Q_0 \frac{dr(b)}{db}(1 - 2r(b)) + Q. \]  

(2.6)
As we can see from the first order expected cost function with respect to $b$, determining the optimal $b$ using an analytical approach is intractable. The analysis we require to derive the second order expected cost function with respect to $b$ becomes even more complex. For this reason, later in this chapter, we describe the use of graphical methods to empirically validate the convex nature of the expected cost function.

### 2.5.1 An Extension to Multiple Retention-Eligible Personnel Groups

Our model (Equation 2.2) so far assumes that the retention probability $r(b)$ is the same for all eligible military personnel. In this section, we consider an extension that classifies the retention-eligible population, $Q_0$, into a finite number of groups, $g = 1, \ldots, G$. Group $g$ consists of $Q_{0g}$ individuals and has retention rate function $r_g(b)$. We define each group as a collection of attributes that are unique in our empirical binary logistic regression analysis. For example, one group can represent a soldier that is married, has no children, and serves under a four-year contract, while another group can represent an unmarried soldier having two children and three years of military experience.

The number of reenlisted personnel from group $g$ is represented by $\chi_g(b)$, which again follows a binomial distribution with parameters $(Q_{0g}, r_g(b))$. The random variable $\chi_g(b)$ can be approximated by a normal random variable with mean $Q_{0g} r_g(b)$ and standard deviation $\sqrt{Q_{0g} r_g(b)(1 - r_g(b))}$; i.e.,

$$\chi_g(b) = Q_{0g} r_g(b) + \sqrt{Q_{0g} r_g(b)(1 - r_g(b))} \epsilon. \quad (2.7)$$

Assuming independence of groups, the total number of reenlisting personnel is given by

$$\chi(b) = \sum_{g=1}^{G} \chi_g(b) = \sum_{g=1}^{G} Q_{0g} r_g(b) + \sum_{g=1}^{G} \sqrt{Q_{0g} r_g(b)(1 - r_g(b))} \epsilon. \quad (2.8)$$

Using (2.8) in (2.2), we obtain

$$\Pi(q, b) = E \left[ c_u \left( Q - q - \sum_{g=1}^{G} Q_{0g} r_g(b) - \sum_{g=1}^{G} \sqrt{Q_{0g} r_g(b)(1 - r_g(b))} \epsilon \right)^+ \right]$$
\[ + b \min \left\{ Q - q, \sum_{g=1}^{G} Q_{0g} r_{g}(b) + \sum_{g=1}^{G} \sqrt{Q_{0g} r_{g}(b)(1 - r_{g}(b))} \epsilon \right\} \] + c_r q. \quad (2.9)\]

As before, in order to simplify our equation, we transform \( q \) to \( z \), by defining \( z \) as
\[
z = \left( Q - q - \sum_{g=1}^{G} Q_{0g} r_{g}(b) \right) / \sqrt{\sum_{g=1}^{G} Q_{0g} r_{g}(b)(1 - r_{g}(b))}. \quad (2.10)
\]

The final form of our expected cost function then becomes
\[
E[\Pi(z, b)] = (c_u - b) \int_{-\infty}^{z} \sqrt{\sum_{g=1}^{G} Q_{0g} r_{g}(b)(1 - r_{g}(b))} (z - u) f(u) du +
\]
\[
(c_r - b) \left[ Q - \sum_{g=1}^{G} Q_{0g} r_{g}(b) - z \sqrt{\sum_{g=1}^{G} Q_{0g} r_{g}(b)(1 - r_{g}(b))} \right] + bQ. \quad (2.11)
\]

Given that we only include explanatory variables having low correlations among themselves and to further simplify our analysis, we assume that all defined groups are jointly normally distributed and are uncorrelated. This assumption implies independence of groups and a covariance of zero (Verbeek, 2008), and allows us to add the standard deviations of each group as shown in Equation 2.8. Additionally, the sum of all the group retention totals must equal \( Q_0 \). This constraint ensures that every group is mutually exclusive and collectively exhaustive. In the following section, we present the results of our numerical experiments using both the generic and group-based binomial models.

### 2.6 Numerical Experiments of Recruitment and Retention Model

We will now consider a numerical study to investigate the value of balancing recruitment and retention by constructing realistic problem instances using historical data. Our model fills the personnel gap between the total end strength authorization established by the U.S. Congress (2010), and the “actual” personnel strength at the end of the fiscal year (September 30th of each fiscal year). The total enlisted Army strength at fiscal year \( t + 1 \) is defined as

\[
\text{Total Strength}_{t+1} = \text{Total Strength}_t + \text{Recruits}_t - \text{Losses}_t + \text{Dropped from Rolls Gains}_t.
\]
Dropped from Rolls (DFR) Gains are commonly referred to as Absent Without Leave (AWOL) Gains, such that once AWOL personnel are in military custody they are considered part of the Army strength; otherwise they are classified as Army losses.

In order to use the appropriate target $Q$ in our model, we must take into account officer and United States Military Academy (USMA) cadet growth/decline to ensure we correctly set the target enlistment. In practice, one determines the personnel shortfall for a fiscal year using a suite of Army loss forecasting models that determine projected losses based on seven years of historical rates. The monthly rates are determined with time series forecasting techniques such as exponential smoothing and primarily use three soldier attributes to determine the following historical loss rates: months of service, enlisted grade, month, and gender. Table 2.7 briefly summarizes the historical strength and personnel losses that we use in our model.

Table 2.7: Enlisted Army Strengths and Losses (2006-2009)

<table>
<thead>
<tr>
<th>Description</th>
<th>Fiscal Year (FY)</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Enlisted End Strength</td>
<td>2006</td>
<td>419,353</td>
</tr>
<tr>
<td></td>
<td>2007</td>
<td>433,109</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>451,846</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>457,980</td>
</tr>
<tr>
<td>Enlisted Losses</td>
<td>2007</td>
<td>71,861</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>68,566</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>70,103</td>
</tr>
<tr>
<td>First Term Retention</td>
<td>2007</td>
<td>27,307</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>28,119</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>24,827</td>
</tr>
<tr>
<td>Recruiting</td>
<td>2007</td>
<td>81,784</td>
</tr>
<tr>
<td></td>
<td>2008</td>
<td>83,710</td>
</tr>
<tr>
<td></td>
<td>2009</td>
<td>72,647</td>
</tr>
</tbody>
</table>

2.6.1 Army Enlisted Empirical Study Parameters and Assumptions

We begin by specifying the main parameters for our model as shown in Equation 2.4 and show our results in Table 2.8. We determine the parameters and target values based on three
years of historical data acquired from Headquarters, Department of the Army (HQDA), G1, and recruitment data cost acquired from Recruiting Command (2013). Note that we use a lower bound in our recruiting and retention model for $q$ because there are 36,487 attrition losses (adverse, administrative, and program managed losses) of personnel that could only be backfilled by recruits and not replaced by reenlistees because these losses were unforeseen by the Army. ETS losses are the only Army losses that can alter retention behavior. We establish the value of 36,487 by averaging FY 2007 (34,534), FY 2008 (37,647), and FY 2009 (37,280) attrition losses according to G1 (2010a) historical reports. Hence, to determine the overall enlisted target value, $Q$, we first calculate the difference between the FY09 and FY08 total strengths which denotes the shortfall due to changes in personnel authorizations. Next, we extract the historical ETS losses and first-term reenlistments for FY09. Last, we add together the recruiting and retention lower bound value, the authorization shortfall, and historical ETS losses and first-term reenlistments, resulting in a $Q$ target value of 98,064.

Prior to conducting our numerical experiments, we execute a short statistical analysis to determine the proportion of infantry career field soldiers relative to the enlisted Army population represented in our data set. Since 14.64% of the enlisted data constitutes infantry soldiers, the personnel targets for our modeling analysis are reduced to more realistic values by multiplying $Q$ by 14.64% to arrive at the $Q$ value shown in Table 2.8. Additionally, the $Q_0$ in our table is calculated from historical records based on FY09’s reenlistment-eligible infantry population.

Table 2.8: Binomial Modeling Parameters: Army Recruiting and Retention

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q$</td>
<td>14,357</td>
</tr>
<tr>
<td>$Q_0$</td>
<td>7,319</td>
</tr>
<tr>
<td>$q_{LB}$</td>
<td>4,619</td>
</tr>
<tr>
<td>$c_r$</td>
<td>$54,000$</td>
</tr>
<tr>
<td>$c_a$</td>
<td>$60,000$</td>
</tr>
<tr>
<td>$b_{min}$</td>
<td>$0$</td>
</tr>
<tr>
<td>$b_{max}$</td>
<td>$c_r$</td>
</tr>
</tbody>
</table>
As stated earlier, we determine the target parameters listed in Table 2.8 based on previously discussed U.S. Army historical records reduced to reflect the infantry force. Recall that $Q_0$ represents the average number of reenlistment-eligible FTM personnel during the retention period. Additionally, we determine the underage cost ($c_u$) for the first-termer population based on the composite rates published by the DOD comptroller’s office in each respective fiscal year (Comptroller, 2011). This composite rate is derived from the following soldier pay and incentives areas: basic pay, retired pay accrual, basic allowance for housing, basic allowance for subsistence, incentive and special pays, and permanent change of station. The goal of our model is to balance recruiting and retention with the objective of minimizing annual costs. As for the recruitment cost, $c_r$, the cost estimate is obtained from the U.S. Army’s accession command and is based on the assumption that a recruit conducts all his/her basic and advanced individual training at the same military installation. Lastly, as previously stated, the error term, $\epsilon$, is a standard normal random variable.

We now provide our results and expound upon our model excursions and insights.

### 2.6.2 Binomial Modeling Observations and Insights using Mean Values

To set the stage for our recruiting and retention scenario experiments using the generic binomial model, we assume every entity in the model possesses a certain set of attributes. In order to avoid marginalizing our model experiments by setting attributes such as marital status and service term as constant values into our model, we make use of our summary statistics and set each variable’s mean value as an input into the model. It is common practice to use mean variable values to conduct experiments of regression equations, as explained in Chambliss and Schutt (2013). The only exception among model inputs is the bonus amount (decision variable). The average inputs are shown in Table 2.9.

Using the “log odds” formula $\ln[p/(1-p)] = \beta_0 + \sum_i \beta_i x_i$, which implies $p/(1-p) = \exp^{\beta_0 + \sum_i \beta_i x_i}$, we then derive the retention probability as

$$r(b) = \frac{\exp^{\beta_0 + \sum_i \beta_i x_i}}{1 + \exp^{\beta_0 + \sum_i \beta_i x_i}}. \quad (2.12)$$
Table 2.9: Average Variable Values for Mean Value Experiment

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Average Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Service Term (SVC_TERM_CD)</td>
<td>3.5458</td>
</tr>
<tr>
<td>Marital Status (MARRIED)</td>
<td>.3186</td>
</tr>
<tr>
<td>Additional Obligation of Service (AOS)</td>
<td>34.118</td>
</tr>
<tr>
<td>Minor Dependents 2 (MINORDEP2)</td>
<td>.0508</td>
</tr>
<tr>
<td>Minor Dependents 3 (MINORDEP3)</td>
<td>.0163</td>
</tr>
<tr>
<td>Experience (EXPER)</td>
<td>3.4216</td>
</tr>
</tbody>
</table>

The values for the parameters $\beta_0$ and $\beta_i$ in Equation 2.12 originate from the coefficients of the previously defined Logit model (Equation 2.1). Because we assume the infantryman career field and women are not permitted to join the infantry ranks, we do not include the gender attribute in our model. In essence, our retention model is transformed to a model of the form, $r(b, X)$, where $X$ represents the itemized averages in the list shown in Table 2.9. To determine the range of retention rates based on the above parameters, we vary the average bonus value from 0 to $c_r$ (the maximum offer in our historical records is $25,000). Recall that we use the formula, $\chi(b) = Q_0r(b, X) + \sqrt{Q_0r(b, X)} \cdot (1 - r(b, X)) \mu_\epsilon$, to determine the optimal retention given $r(b, X)$. Given the minimum recruitment level we set based on involuntary personnel losses, we execute numerical experiments whose results we discuss in the paragraph that follows.

Below we explain the results shown in Table 2.10. Under the actual recruiting and retention cost (FY09), we conduct a simple calculation to arrive at the cost figure shown. Using the historical average first-termer infantry career field bonus value of $7,587, we calculate the actual cost by plugging it into our derived expected cost equation. As previously mentioned, we assume a recruiting cost of $54,000 per recruit for both actual and model cost calculations. This amount includes MOS training, which occurs at the same site as basic training, which is termed One Station Unit Retraining (OSUT). In actuality, the cost of recruiting is much greater for some MOSs since Advanced Individual Training (AIT) is required for many MOSs following completion of basic training. Note that there are no salary differences between recruits; all career fields receive
Table 2.10: Empirical Results for Infantrymen using Binomial Recruiting and Retention Model (Attribute Means)

<table>
<thead>
<tr>
<th>Description</th>
<th>Quantity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Recruiting Mission (q^*)</td>
<td>7,417</td>
</tr>
<tr>
<td>Optimal Retention Mission (\chi^*)</td>
<td>6,917</td>
</tr>
<tr>
<td>Personnel Shortfall</td>
<td>22</td>
</tr>
<tr>
<td>Optimal Recruiting and Retention Cost</td>
<td>$400,614,381</td>
</tr>
<tr>
<td>Optimal Bonus (b^*)</td>
<td>$12,291</td>
</tr>
<tr>
<td>Optimal Retention Rate (r(b^*))</td>
<td>94.5%</td>
</tr>
<tr>
<td>Actual Recruiting</td>
<td>10,636</td>
</tr>
<tr>
<td>Actual Retention</td>
<td>3,721</td>
</tr>
<tr>
<td>Actual Recruiting and Retention Cost</td>
<td>$500,152,738</td>
</tr>
<tr>
<td><strong>Cost Savings</strong></td>
<td><strong>$99,538,356</strong></td>
</tr>
</tbody>
</table>

the same compensation. Basic and AIT training cost in this aforementioned scenario equates to $73,000. Next we discuss our numerical results.

First, our results indicate optimal retention is achieved at 6,917, an increase of nearly 3,200 personnel from the actual mission achieved. This increase is rather substantial because retaining additional personnel saves the Army a considerable amount of money given the high cost of recruiting ($54,000) compared to the range of Army bonus offers historically awarded to incumbent infantry soldiers ($0 to $ 25,000).

Second, the *actual* retention value is obtained by calculating the difference between \(Q\) and \(q\). Henceforth, our modeling results show that recruiting decreases from 10,636 to 7,417, a 3,219 personnel reduction. This decrease is very promising since our analysis assumed only one MOS. Moreover, we would argue that the Army can inherit substantial cost savings by retaining more first-term personnel rather than recruiting and training new people.

Third, the manpower model does not fully meet its target, \(Q\). The model reports a shortage of only 22 personnel, which equates to a unit readiness cost (lack of adequate personnel to conduct a mission). The model attempts to fill all requirements but has no direct control of the stochastic aspect of the problem. The risk associated with large shortages is relatively small. For a shortage
greater than 100, the probability of this occurrence is below 0.1%. In contrast, if we assume the error term has a significantly greater standard deviation of 1,000, the risk associated with this parameter change rises to 49.8%. However, specifying an extremely large variance is unrealistic given our assumption that the error term in the retention model behaves as a standard normal random variable. So, although having such large personnel shortages is very low given our chosen parameter settings, any level of programmatic risk is detrimental to the Army given the importance of combat units deploying at 100% personnel strength.

Fourth, a modeling excursion that tries to capture a soldier’s behavior under high and low bonus offers leads to some interesting results. A bonus of $53,000, nearly equivalent to the cost of recruiting ($c_r$), leads to an expected retention rate of 99.99%. At this high bonus level, retention mission is 7,319, recruitment level is a low 7,038, and cost savings are in excess of $124M. However, this scenario is not optimal because infantrymen will reenlist at a much lower bonus value. On the contrary, when offered a meager $1,000 bonus, the infantry retention rate is 11.22%, resulting in lower reenlistments (821 soldiers). This scenario significantly raises Army manpower costs by ~$229M in comparison with the “current” Army program and is not a wise precedent to set. These results are interesting and provide us lower and upper bounds on recruiting and retention target values.

Finally, as evidenced by the results presented in Table 2.10, incorporating our model into its current recruiting and retention planning process would have saved the Army an estimated $99.5 million dollars in fiscal year 2009 for the infantry career field. We acknowledge that we made numerous assumptions with respect to our input parameters and modeling approach, and so our cost saving results, which are dependent on our model setup, are hypothetical. Nevertheless, these results indicate that substantial cost savings are possible.

Furthermore, the lower bound we have set for recruiting ensures the Army would add a significant number of fresh faces during each fiscal year and so postures the Army with personnel capable of filling future senior enlisted personnel requirements. However, we view a group-binomial model as more appropriate to use than our generic binomial model since it captures the different soldier
attributes that affect retention rate. It is worth noting that our numerical experiments graphically confirm the convex structure of the generic and group binomial models (see Appendix B). In the next section, we present our modeling results using the group retention approach (Equation 2.11).

### 2.6.3 Binomial Modeling Observations and Insights using Group Approach

Prior to presenting our results, we first explain how we determine groups. Based on our final regression equation with six independent variables, and, given that we vary the maximum bonus offers, we have five variables remaining in our regression with which to define groups. These remaining five variables allow us to create a total of 288 unique groups since every observation must have an attribute for service term, additional obligation of service, and experience. We define interval bins for the continuous variables, and Table 2.11 shows the resulting values. We acknowledge that discretization of continuous variables can lose important information communicated by continuous values, but we have created a sufficient number of bins to minimize the reduction in our model’s generalization accuracy. One should note that the maximum experience a first-term soldier can possess is six years since this is the maximum initial contract length offered by the Army. Also, the AOS variable cannot exceed 48 months since a soldier can only reenlist when he is within two years of his/her ETS date. Hence, the value of 48 months takes into account the soldier that reenlists for six years and has two years remaining on his/her current contract.

<table>
<thead>
<tr>
<th>Continuous Variable</th>
<th>Interval Definition</th>
<th>New Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EXPER (years)</td>
<td>(0 \leq EXPER \leq 2)</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>(2 &lt; EXPER \leq 3)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>(3 &lt; EXPER \leq 4)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>(EXPER &gt; 4)</td>
<td>5</td>
</tr>
<tr>
<td>AOS (months)</td>
<td>(0 \leq AOS \leq 24)</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>(24 &lt; AOS \leq 36)</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>(AOS &gt; 36)</td>
<td>48</td>
</tr>
</tbody>
</table>

We remove groups that do not have any observations and thereby reduce the number of unique
groups to 96. For the remaining groups, the rule of thumb for approximating a binomial random variable with a normal distribution is that the product $np$ (where $n$ is the number of trials, $p$ is the probability of success, and $np$ is the mean of the binomial random variable) should be greater than ten according to Ross (2007). In our case, we establish the requirement that $Q_{0g} \geq 20$. Only 20 groups met this requirement, and the frequency counts of the remaining 76 groups (i.e., those with frequency counts less than 20) are aggregated and their within-group attribute values are averaged. The final number of groups is therefore 21 as shown in Table 2.12. Referring to the previously defined function (Equation 2.12), we define the constant $\alpha_i$ as $\alpha_i = \exp^{\beta_0 + \sum_i \beta_i x_i}$. Again, we define $x_i$ as the value of group attribute $i$, and $\beta_i$ represents the group MLE estimate for the $ith$ attribute’s parameter as shown in Table 2.4. Note that the MLE for the bonus variable, $\beta_b$, is excluded from the constant $\alpha_i$, and its value for all groups is 0.4353. Table 2.12 shows the $\alpha_i$ constant values. Finally, the last column in the table shows the optimal group retention rates from our numerical experiments. These range from 4.70% to 99.95%, demonstrating the importance of the logistic regression empirical work and its sensitivity to bonus offers and soldier attributes.

Given the above-mentioned groups, we input these parameters into our group-binomial model (expected cost) to provide some additional insights as with the generic binomial model. We assume the same recruiting and underage costs and the same personnel targets as with the generic model. The maximum bonus offer ranges from $0 to $c_r$, and the results from our analysis are shown in Table 2.13.

In analyzing the results, we achieve optimal retention at 6,609 personnel, which is slightly lower (∼308 personnel) than the results from the general binomial model. However, this number is also an increase of 2,888 infantrymen over fiscal year 2009’s actual retention occurrence. Obviously, similar to the general binomial model, increasing retention significantly reduces personnel costs given the high cost of training and recruiting a new soldier. The recruiting mission decreases by 2,910 personnel, an increase of approximately 309 personnel over the general binomial model. Although this modeling approach yields an increase in recruiting compared with our other modeling approach, this output is more realistic given the various groups we use in our model to diversify
Table 2.12: Parameter Specifications for Each Group and their Respective Optimal Retention Rates

<table>
<thead>
<tr>
<th>Consolidated Group Parameter ($\alpha_i$)</th>
<th>Frequency</th>
<th>Group Retention Rate ($r_i(b^*)$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.34625</td>
<td>490</td>
<td>99.69</td>
</tr>
<tr>
<td>0.00889</td>
<td>336</td>
<td>89.27</td>
</tr>
<tr>
<td>0.55350</td>
<td>188</td>
<td>99.81</td>
</tr>
<tr>
<td>0.01421</td>
<td>611</td>
<td>93.01</td>
</tr>
<tr>
<td>0.00415</td>
<td>265</td>
<td>79.50</td>
</tr>
<tr>
<td>0.16143</td>
<td>317</td>
<td>99.34</td>
</tr>
<tr>
<td>0.00023</td>
<td>49</td>
<td>17.60</td>
</tr>
<tr>
<td>0.25805</td>
<td>135</td>
<td>99.59</td>
</tr>
<tr>
<td>0.07990</td>
<td>902</td>
<td>98.68</td>
</tr>
<tr>
<td>0.00205</td>
<td>491</td>
<td>65.75</td>
</tr>
<tr>
<td>1.45060</td>
<td>845</td>
<td>99.93</td>
</tr>
<tr>
<td>0.00096</td>
<td>243</td>
<td>47.23</td>
</tr>
<tr>
<td>0.03725</td>
<td>510</td>
<td>97.21</td>
</tr>
<tr>
<td>0.12772</td>
<td>216</td>
<td>99.17</td>
</tr>
<tr>
<td>0.00328</td>
<td>74</td>
<td>75.42</td>
</tr>
<tr>
<td>0.00005</td>
<td>133</td>
<td>4.70</td>
</tr>
<tr>
<td>2.31890</td>
<td>191</td>
<td>99.95</td>
</tr>
<tr>
<td>0.01737</td>
<td>160</td>
<td>94.20</td>
</tr>
<tr>
<td>0.00153</td>
<td>102</td>
<td>58.86</td>
</tr>
<tr>
<td>0.33474</td>
<td>149</td>
<td>99.68</td>
</tr>
<tr>
<td>0.10000 (Hybrid group)</td>
<td>912</td>
<td>99.15</td>
</tr>
</tbody>
</table>

Table 2.13: Empirical Results for Infantrymen Using Group Binomial Recruiting and Retention Model

<table>
<thead>
<tr>
<th>Description</th>
<th>Quantity/Amount($$$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Optimal Recruiting Mission ($q^*$)</td>
<td>7,726</td>
</tr>
<tr>
<td>Optimal Retention Mission ($\chi^*$)</td>
<td>6,609</td>
</tr>
<tr>
<td>Personnel Shortfall</td>
<td>22</td>
</tr>
<tr>
<td>Optimal Recruiting and Retention Cost</td>
<td>$418,345,257</td>
</tr>
<tr>
<td>Optimal Bonus ($b^*$)</td>
<td>$15,716</td>
</tr>
<tr>
<td>Actual Recruiting</td>
<td>10,636</td>
</tr>
<tr>
<td>Actual Retention</td>
<td>3,721</td>
</tr>
<tr>
<td>Actual Recruiting and Retention Cost</td>
<td>$500,152,738</td>
</tr>
<tr>
<td>Estimated Recruiting and Retention Cost (Group)</td>
<td>$419,529,354</td>
</tr>
<tr>
<td>Estimated Recruiting and Retention Cost using $b^<em>$ and $q^</em>$ from Generic Binomial Model</td>
<td>$459,464,631</td>
</tr>
<tr>
<td><strong>Cost Savings (Actual Cost - Group Cost)</strong></td>
<td><strong>$81,807,481</strong></td>
</tr>
</tbody>
</table>
various soldier attributes in order to determine the corresponding group retention rates. Similar to the general binomial model, based on Table 2.13, our stochastic retention model does not fully meet the objective, $Q$. The model foresees a shortage of 22 personnel, which equates to a very low readiness cost. More importantly, we see that the model recommends a bonus of $15,716 to achieve the retention rates shown in the table for the various groups. As is evident, this bonus value is much greater than that of the generic model, exceeding it by $3,425, but the group binomial approach captures the retention behavior of each uniquely defined group and its associated sensitivity to bonus offers.

Similar to the generic binomial model, we use the average bonus value of $7,587 as an input into our group binomial model to determine the actual cost as defined in Equation 2.11. Thus, by incorporating our model into its recruiting and retention planning process, while also making good use of historical enlisted personnel records, the Army could have saved over $81 million in fiscal year 2009 given our model setting. This is a decrease in savings of approximately $17.7 million compared to the alternative, general binomial model. However, when we input the optimal bonus amount, $b^*$, from the generic model into the group model, we see a cost increase of $41M over the group cost, thus demonstrating the model’s sensitivity to soldier attribute changes and its ability to capture retention behavior given the attribute diversity of the infantrymen population.

Hence, we have shown that both of our binomial models are capable of handling different “what if” scenarios and soldier attributes. The much simpler and less realistic generic binomial recruiting and retention model is useful if one wishes to simplify the manpower planning process. Our single-period, group-binomial model shows research promise and has vast potential for expansion and improvement, not only in the military sector but also in the corporate world. However, we believe that our group-binomial model is more appropriate to use in the military since it takes into account the different attributes (i.e., variables in the retention rate equation) soldiers can possess. Our results show that we can glean personnel cost reductions from the group-binomial model when recruitment and retention levels are decided jointly. In the next section, we discuss ways in which the binomial inventory manpower models presented here can be further expanded and analyzed.
2.7 Summary and Future Directions

This paper provides valuable insights into the area of military personnel planning. The idea of balancing recruiting and retention missions in unison, rather than addressing them in isolation as current Army models operate, is a very intriguing concept and promises to generate substantial cost savings to the military. Our models show that providing more retention bonuses to first-term soldiers increases retention rates and indirectly reduces training and recruiting costs for new personnel.

Fundamentally, retention bonuses are much less than the cost of recruiting, so retaining personnel by increasing bonuses is not a surprising result. However, the empirical study we present in this chapter gives the reader an idea of the impact of monetary bonuses on retention, and the possible benefits of balancing recruitment and retention. Although our model is too general to use as a manpower planning tool, it provides valuable insights into the retention behavior of first-term enlisted soldiers. If one considers the long term implications of this model, it is easy to see that it could potentially generate a higher quality of soldiers. Since reenlisted soldiers are trained and have military experience, this could potentially enhance the Army’s ability to conduct its future wartime missions which equates to a more lethal military force. Contrary to the Army’s current manpower modeling objective of recruiting a high number of personnel which adds a high number of young personnel to the Army inventory, our approach will produce less of these recruits and more incumbent soldiers. In the long run, this will result to an older composition of the force. Unfortunately, if the personnel structure (authorizations by job specialty) of the military force changes significantly in the future, this could potentially result in too many or too few personnel for a particular MOS which directly affects unit readiness. Also, if there are too many senior soldiers, then the aging force can result in a reduced ability to engage in military conflicts given the senior soldiers diminished ability to perform younger soldier tasks. It is essential that the ever-evolving Army have a minimum number of new soldiers every year to populate the military force in order to provide a fresh outlook on Army tasks and missions.
The new perspective we propose on military manpower planning is based on the idea that offering first-term incumbent personnel more and increased monetary bonuses can lead to substantial cost savings for the Army, especially if the Army is able to decrease the annual number of recruits while continuing to fill projected personnel shortages. Although this model provides insights at the macro level of detail and is not indicative of all the various Army Military Occupational Specialties (MOSs), the model shows promise in providing DOD an alternative mindset for conducting cost-benefit analyses regarding recruitment and retention.

In the future, one can develop individual recruitment and retention models aimed at balancing personnel shortages by career fields, or even groups of combat MOSs. Obviously, such drastic changes as incorporating career field groupings or individual military occupations will definitely change the form of the \( r(b, X) \) function. For example, if one foresees shortages in the cannon crew member occupation and the military police, one can modify the current model to represent recruiting and retention personnel based on their respective historical strengths, projected future shortages, and mission requirements. One can then set target bonus values based on a group of soldiers’ career fields or on strategically defined “group” membership.

Moreover, an extension of this research would be to include a pre-order policy such that a reenlisting soldier would be guaranteed a certain bonus amount and other incentives if the soldier commits to the Army at the earliest possible time within his/her reenlistment window. Such a model would benefit the Army because its soldiers’ committing early would enable the Army to create predictability in its personnel inventory and properly plan soldier and unit movements across the continental United States (CONUS) and throughout the world. This type of model would also eliminate soldiers that desire to “play the system” and fail to commit to the Army in hopes of taking a higher bonus offer as they near their expiration term of service. Personnel and pecuniary predictability is an essential factor in the Army’s short- and long-term budget planning process. Additionally, the Army could target bonuses according to MOS, increasing or decreasing bonus offers based on an MOS’s attributes such as shortages within the occupation field, desirability, or any other relevant measure.
In our numerical experiments, we set a minimum recruiting target for the time period of our analysis. One reason for doing so is to offset the Army personnel losses whose behavior we cannot affect. Such losses include personnel involuntarily released by the Army for disciplinary reasons, losses due to medical reasons, or combat-related fatalities. Additionally, we acknowledge that, in order to prevent future personnel shortfalls in the higher enlisted ranks, we must also determine minimum recruiting targets for each time period based on long-term personnel targets. Based on analysis of historical promotion rates from grades E1 to E9 (by occupation), future research would necessitate a more detailed minimum recruiting analysis (“fresh faces”) to ensure that higher rank personnel requirements are met in the future. This type of analysis can potentially create balance in the enlisted workforce.

For future research, this model could provide a basis for developing a multi-period manpower model that spans fiscal years for either short-term or long-term planning. In addition to incorporating a time variable, constructing a utility-based model is a further expansion. This utility model can also incorporate dynamic bonus offers as well as a multinomial logit model that takes into account a reenlisting soldier’s reenlistment options. For example, when a soldier makes the decision to stay in the Army, he chooses one of six reenlistment options that may or may not include a bonus. The current reenlistment options include: station stabilization, Continental United States (CONUS) location of choice, overseas assignment, Army Military Occupational Specialty (MOS) retraining, needs of the Army, and separation from the Army (civilian workforce option). It would be interesting to study the value a soldier places on each option and assign a monetary “value” to these choices. This type of analysis could assist the Army in determining the retention and reenlistment option probabilities of a specific group of soldiers (i.e., those sharing specific attributes), and possibly shape their decisions based on monetary incentives. In conclusion, the number of potential research arenas for our model in the military human resource personnel planning sector is substantial.
Chapter 3

Two Methods of Predicting Employee-Group Retention Behavior

3.1 Introduction

It is important for any organization to fully utilize its human capital, and the ability to classify people into strategically valuable groups can help in doing so. For instance, a company may wish to develop women for upper management positions or tailor specific incentive packages to attract and retain younger workers having specific skill sets. Strategically viewing employees as grouped into clusters that are not necessarily mutually exclusive can provide upper management with a broader perspective of both the common and unique characteristics of their employees.

Organizations have many tools with which to shape a workforce from its incumbent employees. Since a company invests a substantial amount of time and money training and developing employees, it is in the company’s best interest to identify personnel-specific attributes that could indicate their likelihood of remaining with or leaving the company. In both the industrial and government sectors, a large amount of employee data is collected. With this data at its disposal, management may attempt to shape the firm’s workforce by either increasing or decreasing employee incentives.

Effective manpower planning requires a thorough understanding of employee retention behavior, that is, the likelihood that employees will stay with the organization. Accurately estimating the retention propensity of a group of employees is vital because it prepares a company for the possible departure of employees fitting the particular demographic profile that defines the group, while also identifying the possible reasons (via a set of attributes) underlying their decision.
When the goal of a manpower analyst is to estimate the probability of an individual employee staying with a company for the subsequent year, one would gather the employee’s demographic data, and determine which proportion of the employee workforce remained with the company given that the employee exhibits a certain set of attributes (e.g., race, gender, tenure, and age). Common approaches to estimate such retention probabilities include Decision Tree-based methods and K-Nearest Neighbor (KNN) algorithms. Methods based on Decision Trees estimate these probabilities by creating a tree structure, where each internal node in the tree specifies a test of some attribute, and each branch descending from that node corresponds to one (or a subset) of the possible values of this attribute. The attributes of an employee thus create a path from the tree’s root node to a leaf node, which will have a value representing the estimated retention probability. KNN approaches, on the other hand, would estimate an employee’s retention probability by considering what fraction of its K nearest neighbors (that is, other historical employee observations) stayed, based on some measure of distance in the attribute space.

In this chapter, we focus on extending these approaches to settings where, instead of estimating the retention probability of a single employee, we are interested in estimating these retention probabilities for certain predefined sets of Employee Groups. Employee groups are characterized by a specific set of attributes, and if an employee matches the group attribute value specifications he or she is considered a member of that group. As stated above, the motivation for considering employee groups is to support and understand the impact of strategic workforce decisions.

To estimate these group-based retention probabilities, we first develop methods to construct what we refer as “retention trees.” A retention tree is a variant of a classification and regression tree (CART), that, given an employee group’s attributes, returns the probability that a member in that group will stays with an organization for a pre-specified additional time period. Unlike ordinary CART and regression methods, our model conducts a heuristic search of employee attributes and determines the best way to cluster employees along the tree nodes (attributes) and branches (attribute values), based on an objective function that uses the employee groups to measure the tree’s classification performance. In addition, we also develop KNN algorithms to estimate the
group-based retention probabilities. Specifically, our KNN approach allows potentially different weights for each of the $K$ neighbors, which are optimized using commercially available software. Furthermore, we consider the impact of imposing certain constraints on these weights. Numerical experiments suggest that both methods provide a higher prediction accuracy for the retention propensity of employee groups when compared with traditional methods that aim to maximize the prediction accuracy of individual employees.

The remainder of this chapter is organized as follows. First, we present a literature review that highlights the related research and explains how our approaches differ from existing methods. Second, we introduce our models and describe our retention tree and KNN modeling approaches. Third, we demonstrate the effectiveness of our proposed methods when compared to decision trees constructed using commercially available software. Lastly, we provide avenues for further research in the area of employee retention modeling.

3.2 Literature Review

This section presents a brief overview of research within the machine learning and marketing fields that most closely relates to our work with an emphasis on decision trees, which are widely used in academia and industry to support classification involving multi-stage decision making. We assume that the reader possesses basic knowledge of decision trees and KNN data mining techniques and therefore only provide a brief introduction to these methods. Readers unfamiliar with decision trees and the classification methodology are referred to Safavian and Landgrebe (1991a), who provide a thorough tutorial on the terminologies, tree structure design, potential uses, and limitations of decision trees. For a comprehensive study of KNN machine learning, we refer to Aha (1991) and Aha et al. (1991). Murthy (1998) provides a comprehensive survey on decision trees and points to the popularity of decision trees in areas such as statistics, pattern recognition, decision theory, machine learning, and artificial neural networks.

The most relevant statistical method related to our approach is the Chi-squared automatic interaction detector (CHAID) classification method. CHAID is a popular tree classifier currently
utilized both in research studies and in industry. For a detailed tutorial on CHAID and other supervised tree methods, see Ritschard (2010). CHAID’s key contribution is the manner in which it determines $n$-ary splits for each predictor (i.e., attribute, feature, etc.). This splitting method is key because almost all other decision tree algorithms consider only binary splits of independent variables. Essentially, CHAID seeks to merge categories (for any of the available predictors) that appear similar, so as to reduce the complexity of the decision tree. Categories are merged in a greedy fashion such that category pairs with the highest $p$-value are merged. This process continues until significant Chi-squared values are obtained for all remaining pairs. As such, CHAID can be viewed as a top-down way of testing group homogeneity. Although CHAID provides $n$-ary attribute splits, one drawback is the greedy fashion in which it searches the predictor solution space; hence, this optimization algorithm does not allow escape of a local optimal decision tree. The CHAID tree that will be generated is deterministic, in that at each node the algorithm selects the best local decision, but the produced decision tree will never change given this search procedure. Later in the chapter, we compare our results to CHAID’s and demonstrate how our models outperform CHAID.

As explained by Murthy (1998), most tree induction systems use a greedy approach, top-down, one node at a time. The author states that the idea of determining globally optimal decisions is not novel but is very difficult to implement given the complexity of some tree structures. Constructing trees that can partially or, if possible, exhaustively look ahead is studied in the statistics literature by Elder (1995) and Chou (1991). Quite amazingly, a majority of these studies indicate very little improvement in their lookahead procedures compared to those that use a greedy approach. In fact, Murthy and Salzberg (1995) demonstrate that “one-level lookahead does not help build significantly better trees and can actually worsen the quality of trees.”

Other decision tree methods unrelated to our study are prevalent in the area of pattern recognition. We refer the reader to Dattatreya and Kanal (1985) and Chandrasekaran and Goel (1988) for detailed discussions on the use of decision trees this context. In the area of inferential statistics, a very popular and often cited book by Breiman et al. (1983) provides a detailed look at classification trees from a statistical perspective. Two very popular machine learning algorithms,
ID3 and C4.5, are described by Quinlan (1993) using the ideas of Breiman et al. (1983) as a framework.

A related approach is the Bayesian CART algorithm proposed by Denison et al. (1998) who present a stochastic search for CART analysis based on Bayesian modeling principles. The authors utilize a forest of trees to approximate a probability distribution over the search space by using “reversible jump Markov chain Monte Carlo methods.” Their algorithm’s procedure regards the number of splitting nodes, their positions, and the questions at the nodes as unknowns. The appeal of their methodology is that it does not require prior knowledge of the tree structure, unlike other decision tree constructions that seek to optimize a tree. Although this model produces multiple sub-optimal trees based on the input data, its results provide insights into the range of “good” trees rather than a point estimate.

Since the main focus of our research is on retention trees, we only briefly discuss contributions in the area of KNN. Essentially, there are two KNN approaches: KNN for classification and KNN regression. Typically, these approaches assign equal weights to a set of attributes in a data set to predict a dependent variable (KNN classification) or to assign a predicted value to the regressand (KNN regression). For an overview of KNN classification and KNN regression we refer the reader to Shakhnarovich et al. (2006), Altman (1992), respectively. However, our focus is on assigning optimal weights (such that they sum to 1) for each of the $K$ neighbors, resulting in a probability assignment for each employee. There is some limited research on varying the $K$ weights such as Atiya (2005) that uses maximum likelihood estimators to determine optimal $K$ weights and enforcing the weights by a neighbor’s proximity to an observation; the closer the neighbor the greater the weight. However, the use of metaheuristic optimization techniques to determine the best $K$ neighbor weights individually has not been explored. Applications of KNN include the application of genetic algorithms to medical fraud detection (He et al, 1999). Their focus is determining optimal equal weights for a set of attributes, followed by a KNN algorithm to classify data. Tommola et al. (1999) employed a KNN as a wood-procurement planning tool. Their inventory model is useful in planning wood procurement within the forest industry. In the area of manpower planning, Jantan
et al. (2010) propose a KNN technique for employee development and future job performance. The authors use a distance metric to measure the similarity of employees (observations) in the human resources (HR) data. Gupta et al. (2006), who present the idea of developing long-term relationships with customers, state that, in the marketing realm, one seeks to maximize customer lifetime value (CLV). Although KNN methods are not the focal point of his paper, they identify the need of such modeling approaches to maximize CLV. The authors review several CLV models in the marketing academic field, including ones dedicated to allocation of marketing resources for customer retention. The authors present several empirical insights from the marketing models they reviewed and state that customer retention is one area that can benefit from additional research.

To our knowledge, literature that studies employee retention propensity using decision trees and KNN approaches does not exist. We believe our research regarding retention trees and KNN to predict group-employee retention behavior (as opposed to classifying individual employees), presents a novel approach with opportunities for further research. We now present our model and explain the notation used throughout the chapter.

### 3.3 Group-based Retention Trees

As stated before, it may be more important for a company to determine the retention behavior of its workforce by capturing employee retention propensity of a group of employees rather than a single employee. We now provide a discussion of our decision tree approach to determine group-based retention behavior. Before presenting our retention tree model, we first define and introduce notation common to both our statistical analysis of employee data and our modeling approach. Then, we introduce our group-based retention tree problem and explain how it differs from ordinary retention trees. Finally, we discuss how we construct retention trees using on our novel group-based approach.
3.3.1 Decision Trees

We start with a brief overview of decision trees to introduce basic concepts, notation, and terminology. The inputs associated with the construction of a decision tree $T$ consists of a set of observations $N = \{1, \ldots, n\}$, which in our setting consist of employee records. In the remainder of this chapter, we use the term employee and observation interchangeably. Each observation is characterized by a set of attributes $A = \{1, \ldots, m\}$. Attribute values can be ordinal or categorical, and the possible values for attribute $j$ are given by its domain $v_j$. Moreover, each observation $k$ has a binary response variable (or dependent variable) $Stay_k \in \{0, 1\}$ that represents whether an employee remained with the company.

The nodes in a decision tree $T$ can be divided into a set of internal nodes $S$ and a set of terminal nodes $E$. Aside from the root node, every “child” node will be the descendant of a single parent. Each internal node $s \in S$ will have an associated attribute $a(s)$, and branch originating from this node will have an associated attribute value (or set of attribute values) from the domain of $a(s)$. Terminal nodes $e \in E$ have no descendants, but have a probability $p(e)$ associated with them. We discuss the calculation of terminal node probabilities later in the chapter.

A generic procedure for constructing decision trees is as follows (Breiman et al., 1983):

Step 1. Examine every allowable split (binary or n-ary) at every unexplored terminal node $e$ for each candidate attribute variable $j \notin A(i)$.

Step 2. Select and execute the best of these splits, adding a subtree composed of branches and new terminal nodes.

Step 3. Stop splitting on a node when some stopping rule is satisfied.

Here, $A(i)$ is the set of attributes associated with the parent nodes in the branches leading from the root node to node $i$. A split is the manner in which a particular attribute can be partitioned based on its range of values. For continuous attribute variables, a split in Step 1 can be executed with a question such as is $\{j > d\}$ for all $d$ in the range of attribute $j$. For categorical attribute variables that take on finite values, the questions take the form $\{j \in D\}$, where $D \subset v_j$. 
Commonly used procedures to construct ordinary regression trees, such as CHAID, employ a heuristic that is similar to the pseudo-code shown in Algorithm 1 below. For each unexplored terminal node, the procedure finds the best attribute to split. To determine the best attribute to split on, the decision is based on which of the nodes’ potential attributes contributes the most to a particular classification or probability estimation (e.g., stay or leave a company) according to some objective function \( f(T) \). In this algorithm, \( f(T) \) is usually represented by a so-called entropy function (purity test) or a Chi-squared test. For a more elaborate discussion of entropy functions see Safavian and Landgrebe (1991a). The algorithm stops when all terminal nodes have been explored and no split can further improve its classification accuracy, or if the terminal node size population is too small.

**Algorithm 1 Greedy Tree Construction Phase with Objective Function**

\[
\begin{align*}
T^* & \leftarrow \emptyset & \triangleright \text{Initialize best tree to empty} \\
f^* & \leftarrow \infty & \triangleright \text{Best objective value is a large positive number} \\
U & \leftarrow \text{root} & \triangleright \text{The only terminal node is the root} \\
\text{repeat} & & \triangleright \text{Stop when no improvement is possible (} f' \geq f^* \text{) or} \\
\quad f' & \leftarrow \infty & \triangleright \text{Is the addition an improvement?} \\
\quad \text{for all } i \in U & & \triangleright \text{Add best subtree} \\
\quad \quad \text{for } j = 1, \ldots, m : j \notin A(i) & & \triangleright \text{Delete } i' \text{ and add new terminal nodes to } U \\
\quad \quad \quad T & \leftarrow T^* \cup \text{Subtree}(i, j) & \triangleright \text{Add subtree of attribute } j \text{ to terminal node } i \\
\quad \quad \quad f(T) & \leftarrow \text{Evaluate}(T) & \triangleright \text{Calculate the classification accuracy of } T \text{ considering all values } (v_j) \text{ for attribute } j \\
\quad \quad \quad \text{if } f(T) < f' \text{ then} & & \triangleright \text{Record the best node-attribute pair} \\
\quad \quad \quad \quad (i, j') & \leftarrow (i, j) & \triangleright \text{Identify best pair} \\
\quad \quad \quad \quad f' & \leftarrow f(T) & \triangleright \text{Update best (local optimal) objective function value in current iteration} \\
\quad \quad \quad \text{end if} & & \triangleright \text{Stop when no improvement is possible (} f' \geq f^* \text{) or} \\
\quad \text{end for} & & \text{terminal node sizes are too small} \\
\text{end for} & & \triangleright \text{Stop when no improvement is possible (} f' \geq f^* \text{) or} \\
\text{if } f' < f^* \text{ then} & & \text{terminal node sizes are too small} \\
\quad T^* & \leftarrow T^* \cup \text{Subtree}(i', j') & \triangleright \text{Add best subtree} \\
\quad U & \leftarrow U \setminus \{i'\} \cup \{i \in \text{Subtree}(i', j')\} & \triangleright \text{Delete } i' \text{ and add new terminal nodes to } U \\
\text{end if} & & \triangleright \text{Stop when no improvement is possible (} f' \geq f^* \text{) or} \\
\text{until Stopping Criteria Satisfied} & & \text{terminal node sizes are too small}
\end{align*}
\]

Given an optimal tree \( T^* \), it is straightforward to determine a terminal node \( e_k \) for each observation \( k \), by following the splits defined in each internal node according to the attribute values of observation \( k \). Now, let \( N_e = \{ k \in N : e_k = e \} \) for each terminal node \( e \in E \). Then, the probability \( p(e) \) associated with a terminal node decision tree can be specified as

\[
p(e) = \frac{\sum_{k \in N_e} \text{Stay}_k}{|N_e|}.
\]
Figure 3.1 illustrates a sample decision tree, based on a data set with the attributes listed in Table 3.1. For simplicity, we define discrete values for each attribute. As an example, using this tree we would predict that the probability that a female employee with 1-3 years of experience will stay with the company equals 0.82.

![Figure 3.1: Numerical Example of Simple Greedy Retention Tree](image)

<table>
<thead>
<tr>
<th>Attribute $i$</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tenure (T)</td>
<td>1-3, 4-9, and over 10 years</td>
</tr>
<tr>
<td>Gender (G)</td>
<td>Male and Female</td>
</tr>
<tr>
<td>Classification (C)</td>
<td>Full-time and part-time employee</td>
</tr>
<tr>
<td>Location (L)</td>
<td>Latin America, U.S., and Europe</td>
</tr>
<tr>
<td>Age (A)</td>
<td>18-25 years, 26-35 years, 36-42 years, and over 42 years</td>
</tr>
</tbody>
</table>

### 3.3.2 Toward Group-based Decision Trees

While the decision trees outlined above focus on estimating the retention probability for individual employees, the focus of our research is on estimating the retention probabilities for certain strategically important employee groups. To illustrate the changes this will introduce, this section contrasts the calculation of true and estimated retention probability when individual
employees and employee groups are considered, and introduces the objective function that we will use to evaluate group-based retention trees.

When individual employees are considered, the true historical probability of staying is defined as

\[ P = \frac{1}{n} \sum_{k \in N} \text{Stay}_k, \]  \quad (3.2)

while the estimated retention probability is defined as

\[ \hat{P} = \frac{1}{n} \sum_{k \in N} p(e_k), \]  \quad (3.3)

given the retention probability estimates at each terminal node \( e \) of the decision tree. In words, the estimated retention probability is simply the average of the estimated retention probabilities for each employee.

When employee groups are considered, however, the true and estimated probabilities are calculated in a slightly different manner. In particular, let \( G \) represent a set of employee groups, and let \( N_g \subseteq N \) be the set of members of group \( g \) (\( g \in G \)). Then, for any group \( g \in G \), the true probability of staying is defined as

\[ P_g = \frac{1}{|N_g|} \sum_{k \in N_g} \text{Stay}_k, \]  \quad (3.4)

while the estimated retention probability is defined as

\[ \hat{P}_g = \frac{1}{|N_g|} \sum_{k \in N_g} p(e_k), \]  \quad (3.5)

again given the retention probability estimates at each terminal node \( e \) of the decision tree.

The objective function we will use to evaluate group-based decision trees is based on the average difference between the true (sample) group probabilities (\( P_g \)) and the estimated group probabilities (\( \hat{P}_g \)), and can be defined either using absolute deviations, that is,

\[ f(T) = \frac{1}{|G|} \sum_{g \in G} |P_g - \hat{P}_g|, \]  \quad (3.6)

or using mean-squared errors

\[ f(T) = \frac{1}{|G|} \sum_{g \in G} (P_g - \hat{P}_g)^2. \]  \quad (3.7)
Overall, the goal of our model will be to minimize the objective function value and provide a decision tree with the highest explanatory power regarding employee-group retention probabilities. Now that we have discussed how we use decision trees to make inferences about employee-group retention behavior, we next explain our procedure for constructing group-based decision trees.

### 3.3.3 Constructing Group-based Decision Trees

In this section, we describe our procedure for constructing group-based decision trees. The procedure builds on the greedy construction procedure outlined in Section 3.3.1, but differs in several key aspects. Aside from using the objective function outlined in the previous section (instead of the entropy functions that are typically used), we also allow $n$-ary splits on attribute-node pairs. This enables us to produce more compact and informative decision trees, because we do not have to explicitly create binary splits for every possible attribute-value pair. In addition, we also add a randomization step in the construction procedure, and use a local search procedure to further improve the resulting trees. We now describe these extensions in greater detail.

**Greedy Randomized Retention Tree Construction.** Instead of identifying a locally optimal attribute-node pair for which to execute a split, we modify the construction process to randomly select an attribute-node pair from a restricted candidate list of improving pairs. A candidate list (CL) is defined as the set of attribute-node pairs whose addition to the tree may improve the accuracy of the objective function, while a restricted candidate list (RCL) is a subset of this list. For example, if we designated a restricted candidate list (RCL) of $z$ candidates, this list would include the best $z$ improving pairs for a particular node in the partial tree. Using a probabilistic strategy, only one of the corresponding splits would be executed. The resulting construction procedure can produce a number of different retention trees by varying the size of the RCL, and is diverse in that it seeks to limit the impact of using locally optimal splits. Pseudo-code for the greedy randomized construction phase is shown in Algorithm 2.

**Local Search.** Once a full retention tree has been generated, we begin a local search procedure that
Algorithm 2 Greedy Randomized Retention Tree Construction Using a Restricted Candidate List (RCL)

\[
T^* \leftarrow \emptyset \\
f^* \leftarrow \infty \\
RCL \leftarrow z \\
U \leftarrow \text{root} \\
\text{repeat} \\
\quad f' \leftarrow \infty \\
\quad \text{for all } i \in U \text{ do} \\
\quad \quad \text{for } j = 1, \ldots, m : j \notin A(i) \text{ do} \\
\quad \quad \quad T \leftarrow T^* \cup \text{Subtree}(i, j) \\
\quad \quad \quad f(T) \leftarrow \text{Evaluate}(T) \\
\quad \quad \quad CL \leftarrow CL \cup (i, j, f) \\
\quad \quad \text{end for} \\
\quad \text{end for} \\
\quad (i', j', f') \leftarrow \text{Select}(CL, z) \\
\quad \text{if } f' < f^* \text{ then} \\
\quad \quad T^* \leftarrow T^* \cup \text{Subtree}(i', j') \\
\quad \quad U \leftarrow U \setminus \{i'\} \cup \{i \in \text{Subtree}(i', j')\} \\
\quad \text{end if} \\
\text{until} \text{ Stopping Criteria Satisfied}
\]

\(\triangleright\) Initialize best tree to empty
\(\triangleright\) Best objective value is a large positive number
\(\triangleright\) Define the candidate list size
\(\triangleright\) The only terminal node is the root
\(\triangleright\) Best objective function in current iteration
\(\triangleright\) Sweep through all unexplored nodes
\(\triangleright\) Sweep through all admissible attributes
\(\triangleright\) Add subtree to the terminal node
\(\triangleright\) Calculate the objective function value of \(T\)
\(\triangleright\) If adding the pair is an improvement, save it to the CL along with its objective function value
\(\triangleright\) Randomly select from the top \(z\) improving \((i, j)’\) pairs along with its objective function value
\(\triangleright\) Add the selected subtree
\(\triangleright\) Delete \(i’\) and add new terminal nodes to \(U\)
\(\triangleright\) Stop when no further improvement is possible \((f' \geq f^*)\) or terminal node sizes are too small

aims to improve this tree. The resulting procedure may thus be viewed as a Greedy Randomized Adaptive Search Procedures (GRASP) (Feo and Resende, 1995a). We begin the local improvement process by deleting an internal node (i.e., a node that is not a root or terminal node) in a greedy fashion from the fully constructed tree. The node that will replace this deleted node and its associated subtree cannot have the same associated attribute as the deleted node. In addition, the newly constructed subtree cannot use attributes that are situated along the branches from the deleted node to the root node.

A graphical illustration of this procedure is provided in Figure 3.2. In this example, we delete attribute A5 and its descendants. We then reconstruct in a greedy manner the subtree without using either attributes A1 or A2, which are in the path from the deleted node to the root. We do this with all remaining internal nodes (with associated attributes A2, A4, A7, A15, A17, and A9) to identify the best node to delete during each local search iteration. It is important to note that we conduct a local search procedure for every internal node and so identify a new tree (best tree) that yields the highest improvement in the objective function value. We continue conducting a local improvement steps on the best tree until no further improvement can be made to its objective
function value or the terminal node populations are too small to continue growing the tree.

Figure 3.2: Visual Representation of Delete and Rebuild Move

During the local search procedure, if the newly constructed tree improves the objective function value, the tree structure and its objective function value are saved for future comparison. Each greedy randomized construction and local search iteration can produce a different retention tree, and the best of those is returned after a prescribed number of GRASP iterations. In other words, we construct a greedy randomized tree and conduct a local improvement procedure for a pre-specified number of iterations. Finally, we return the best tree from all the retention trees that were generated. Algorithm 3 summarizes the pseudo-code for this procedure.
Algorithm 3  Local Search of Greedy Randomized Constructed Tree (GRASP Retention Tree)

\( T^* \)

\( f^* \)

\( \text{repeat} \)

\( \text{Local Search} \)

\( \text{for all } s \in S(T^*) \text{ do} \)

\( f' \leftarrow f^* \)

\( T' \leftarrow T^* \setminus \text{Subtree}(s, a(s)) \)

\( \text{for } j = 1, \ldots, m : j \notin A(i) \text{ and } j \notin a(s) \text{ do} \)

\( T' \leftarrow (T' \cup \text{Subtree}(s, j)) \)

\( f(T') \leftarrow \text{Evaluate}(T') \)

\( \text{if } f(T') < f' \text{ then} \)

\( (s, j') \leftarrow (s, j) \)

\( f' \leftarrow f(T') \)

\( \text{end if} \)

\( \text{end for} \)

\( \text{end for} \)

\( \text{if } f' < f^* \text{ then} \)

\( T^* \leftarrow T^* \setminus \text{Subtree}(s', a(s')) \cup \text{Subtree}(s', j') \)

\( f^* \leftarrow f(T^*) \)

\( \text{end if} \)

\( \text{until} \) Stopping Criteria Satisfied

3.4  \textbf{K-nearest Neighbor (KNN) Methods}

In addition to the use of decision trees, we also evaluate the potential of KNN methods to predict group-based retention probabilities. Given a set of employee observations as defined before, KNN methods aim to estimate the retention probability of an employee based on response variable values of the employee’s K nearest neighbors, predicated on some measure to determine the distance between observations. Our KNN method extends the model proposed by Atiya (2005), in that we modify Atiya’s KNN rule for classifying individual employees to estimate the posterior employee-group retention probabilities. Unlike traditional approaches that equally weigh each neighbor, this approach allows unequal weights to estimate an employee’s retention probability.

Clearly, a key component of any KNN algorithm is the metric that is used to measure the distance between observations. The survey conducted by Jiang et al. (2007) reviews appropriate distance metrics based on the type (scalar or categorical) of data that is used to describe the
observations. For our analysis, we utilize a distance metric similar to the Value Difference Metric (VDM) developed by Wilson and Martinez (1997). Specifically, given two observations that are characterized by the attribute vectors \( \mathbf{x} = (x_1, \ldots, x_m) \) and \( \mathbf{x}' = (x'_1, \ldots, x'_m) \), their distance is given by

\[
D(\mathbf{x}, \mathbf{x}') = \sqrt{\sum_{j=1}^{m} d^2_j(x_j, x'_j)}
\]  

(3.8)

where \( d_j(x_j, x'_j) \) is a metric that represents the distance between attribute values \( x_j \) and \( x'_j \) for attribute \( j \). In practice, the square root in Equation 3.8 is not used, because it does not change the relative distance measure among a pair of employees. The distance between attribute values is defined as

\[
d_j(x_j, x'_j) = \frac{|x_j - x'_j|}{4\sigma_j}
\]

(3.9)

for scalar attributes, while for nominal attributes we use

\[
d_j(x_j, x'_j) = \sqrt{(\gamma_{j,u} - \gamma_{j,u}')^2 + (\tau_{j,x_j} - \tau_{j,x'_j})^2} = \sqrt{2(\gamma_{j,x_j} - \gamma_{j,x'_j})^2},
\]

(3.10)

where \( \gamma_{j,u} \) is defined as the fraction of employees with an attribute value \( u \) for attribute \( j \) that stay, and \( \tau_{j,u} = 1 - \gamma_{j,u} \). For scalar attributes, this distance metric is a modification of the normalized Euclidean distance measurement. This distance measurement includes \( \sigma_j \), the standard deviation of attribute \( j \). In other words, the distance measure is normalized to use four standard deviations to scale each value for each scalar attribute, as opposed to using each attribute’s range. Wilson and Martinez (1997) explain that the reason for this is that 95% of the values in a normal distribution fall within two standard deviations of the mean. They infer that their scaling is a robust alternative in that the presence of outliers in a test set will not reduce the effect of normal instances in a data set. Their experimental results demonstrate that their distance measurement is superior to other distance metrics, including the popular Euclidean and Manhattan distance measurements.

Given an observation \( r \) characterized by the attribute vector \( \mathbf{x}_r \), we let \( l_k(\mathbf{x}_r) \) represent the index of its \( k \)th-nearest neighbor in the set \( N \) of observations according to the distance metric
outlined above. To estimate the retention probability for an employee with attributes $x_r$, we define

$$\hat{p}(x_r) = \sum_{k=1}^{K} w(k) \times \text{Stay}_{k}(x_r),$$

(3.11)

where we require that

$$\sum_{k=1}^{K} w(k) = 1.$$

(3.12)

Unlike traditional KNN approaches which assume equal weights for each neighbor, our approach allows unequal weights. We restrict these weights by imposing that

$$w(k) \geq w(k + 1) \quad \forall k : 1 \leq k \leq K,$$

(3.13)

where we assume $w(K + 1) = 0$. In other words, we require the weight to be non-increasing by distance. The use of unequal weights, restricted by constraint 3.13 was also used by Atiya (2005). His empirical study indicates that enforcing this constraint improves results. In contrast to Atiya, however, we optimize these weight using a metaheuristic search technique.

The objective function we use to optimize these weights is analogous to the objective function that was used to evaluate retention trees, which calculates the mean absolute deviation between the pre-defined true group retention probabilities and the estimated group retention probabilities. The only difference is that in this setting we define the estimated group retention probability as

$$\hat{p} = \frac{1}{n} \sum_{r \in N} \hat{p}(x_r),$$

(3.14)

where, as before, $x_r$ corresponds to the attribute vector of employee $r$. We use OptQuest, a commercially available optimization software, to optimize the KNN weights. OptQuest is instructed to stop when no improvements are made to the objective function (via a set of weights) after a fixed number of iterations.

We conclude this section with a small example that illustrates our overall KNN approach. Table 3.2 contains the observations for this example. The columns in the table, labeled $x_1$ through $x_5$, represent each observation’s attributes, while the response column denotes whether an employee stayed or departed the organization.
Table 3.2: KNN Basic Example Using Employee Data

<table>
<thead>
<tr>
<th>Individual Employee (Observation)</th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( x_5 )</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>5</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Assume all five employees belong to same employee group, so the true retention probability for this group is 0.6. Now, suppose \( K=2 \), and let us estimate Employee 1’s response value. Employee 1’s two nearest neighbors, using a standard Euclidean distance measure, are Employee 5 (distance = 1) and Employee 2 (distance = 2.236). Using a traditional KNN method with equal weights, the estimated retention probability for Employee 1 will equal \((1+0)/2 = 0.5\). Our approach, however, might assign unequal weights; for example, with weights \( w(1) = 0.8 \) and \( w(2) = 0.2 \), the estimated retention probability would become 0.2. Our optimization procedure will choose those weights such that the absolute error between the estimated and true employee group retention probabilities is minimized.

3.5 Employee Workforce Empirical Study

Numerical experiments are conducted for both the retention tree and KNN approaches. In this section, we first explain how we extract employee data, and then discuss the design of our empirical study. We end this section by presenting the results of our experiments.

3.5.1 Personnel Data Extraction

The data for our empirical analysis was acquired from a Fortune 500 company, whose name we omit for privacy reasons. The data set consists of monthly snapshots from January 2010 to June 2013, and contains 883,900 employee data records. Not all the employees in the data set are unique, and, in some cases, an employee can appear in the data set up to 36 times. The name of the company is intentionally omitted to maintain its privacy.
Because the objective of our analysis is to determine an employee group’s propensity to remain in the company (the binary response variable Stay), we remove personnel from our data set that are involuntarily terminated from the company, personnel with unknown reasons for termination, and personnel with unrealistic data entries. Following the removal of the 9,506 involuntary separations, the remaining data set consists of 869,394 monthly observations, of which only 8,296 voluntarily left the company (less than 1% of all employees) in any given month. This small percentage is problematic when trying to predict retention behavior, and therefore we convert the data into annualized counts by determining if a person remained in the company for an entire year. If a person left the company during a calendar year, the voluntary separation is recorded. Likewise, if a person “exists” on January 1st of year 1 and January 1st of year 2, he/she is counted as staying during year 1. Following this data conversion, 70,142 employees comprise our annualized data set. Of these employees, approximately 91% of the workforce remains with the company. It is important to note that since we annualize the employee data records, if an employee’s attribute value changes over the course of a year, the employee inherits the “end of year” attribute value.

The original employee data set consists of 17 independent variables as shown in Table 3.3. Period and Employee Number serve as row identifiers; each unique combination of Period and Employee number identifies a particular employee in a particular month. Job Level is removed, as it contains the exact same information as Summary Job Level for all records and is therefore redundant. At the company’s request, we also exclude the Salary feature from our analysis.

Two relationships between predictor variables are identified that require the exclusion of a predictor variable. First, the variables US/Outside US and Geographic Region contain identical information; Geographic Region simply identifies U.S. employees, as well as the region in which non-U.S. employees worked. Because these variables contain the same information, they have the potential to interfere with modeling results. We therefore exclude Geographic Region from the data set. Business Group Code and Division Code also contain redundant information; each division code is a combination of two or more business group codes. For reasons similar to those cited for the exclusion of Geographic Region, we exclude Division Code from the data set as well.
Table 3.3: Listing of All Attribute Values in Data Set

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Period</td>
<td>Month and Year of employee observation</td>
</tr>
<tr>
<td>Employee Number</td>
<td>Unique employee identification number</td>
</tr>
<tr>
<td>Business Group</td>
<td>Business group code</td>
</tr>
<tr>
<td>Revenue Type</td>
<td>Type of revenue generated by employee</td>
</tr>
<tr>
<td>US/Outside US</td>
<td>Within or outside the U.S.</td>
</tr>
<tr>
<td>Geographic Region</td>
<td>Region in U.S.</td>
</tr>
<tr>
<td>Job Level</td>
<td>A hierarchical structure of positions</td>
</tr>
<tr>
<td>Summary Job Level</td>
<td>Employee grade level</td>
</tr>
<tr>
<td>Gender</td>
<td>Employee gender</td>
</tr>
<tr>
<td>Race</td>
<td>Ethnicity of employee</td>
</tr>
<tr>
<td>Age</td>
<td>Age of employee in years</td>
</tr>
<tr>
<td>Years of Service</td>
<td>Years of service</td>
</tr>
<tr>
<td>Performance Rating</td>
<td>Employee evaluation rating</td>
</tr>
<tr>
<td>Employee Type</td>
<td>Full, temporary or part-time employee type</td>
</tr>
<tr>
<td>Exempt Flag</td>
<td>Salary versus hourly workers</td>
</tr>
<tr>
<td>Craft</td>
<td>Labor or non-laborer type of employee</td>
</tr>
<tr>
<td>Salary</td>
<td>Annual salary amount in U.S. dollars</td>
</tr>
<tr>
<td>Termination Reason</td>
<td>Reason for termination</td>
</tr>
</tbody>
</table>

Upon further examining the attributes we use in the employee-group definitions, we also exclude Business Group code, Race, Employee Type, and Exempt Flag attributes from our analysis since we do not use them in the subset of employee groups we select for our analysis. We explain how we selected these 25 groups at the end of this section. Of the remaining variables, Age and Years of Service are continuous values, Performance Rating (Pep Rating) is ordinal, while the other variables are nominal.

Initial statistical analysis of the annualized data shows that the workforce changes over the time horizon. Specifically, it varies in strength from approximately 17,000 employees to 21,000 employees during the time period of our analysis. Over the entire period, 39,994 unique employees are part of the company’s workforce.

The variable U.S./Outside U.S. shows that 69.17% of the company’s employees work in the U.S., while the Gender variable shows that 73.95% of the workforce is male. The variable Summary Job Level, which summarizes the company’s hierarchical structure of positions, contains
three categories: top, middle, and bottom, which correspond to a 71.14%, 17.84% and 11.02%, respectively.

As previously mentioned, we use two scalar attributes in our data set for our distance measurements, Age and YOS. We assume that the distribution of worker ages is normally distributed, with a mean and median of 43 years. The oldest worker is 90 and is eliminated from our analysis. The distribution of Years of Service is strongly skewed to the right; the greatest number of records has relatively few years of service, and yet there are a number of large outliers with the greatest value being 64. These outliers are removed from our data set. To facilitate our group analysis, prior to executing the employee-group membership evaluation in our algorithm, we convert scalar values into categories based on discrete intervals we define later in this chapter.

The only ordinal independent variable, Performance Rating (PEP Rating), classifies employees into five performance levels according to ratings assigned by their managers in each annual rating period. Just over half of the employees are rated in the middle of the scale, about one-quarter are rated as a 1 (the worst possible rating), and approximately 15% are given the highest possible rating of 5. Finally, the Craft variable records whether or not a worker is a laborer, and the data shows that 16.8% of the workforce are laborers.

Correlations between the metric predictor variables Age, Years of Service, and Salary are examined next. The highest correlation is 0.41, between Age and Years of Service. The other two correlations (between Age and Salary and between Years of Service and Salary) are both fairly low, between 0.2 and 0.3. This suggests that little redundant information exists between the three metric predictor variables, and thus each should contribute unique information to the model.

Relationships between each predictor variable and the response variable are examined next. As the response variable contains an overwhelmingly large number of stay decisions as opposed to leave decisions, identifying relationships between the predictor variables and the response is difficult. One category of the predictor variable Summary Job Level, however, is associated with a large number of leave decisions, suggesting that grade level may play an important role in the final model’s predictive power.
Prior to using the variables in the model, we convert string attribute values into numeric categorical variables. Since order (numerical value assigned to a range of attribute values) does not matter in our GRASP retention tree and KNN algorithms, we convert the Age and Year of Service variables to discrete numeric values (categorical) in order to properly identify employees’ membership in an employee group. The data conversion of the variables is shown in Table 3.4.

Table 3.4: Employee Attribute Value Conversions for Employee-Group Designations

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Original Values</th>
<th>Model Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revenue Type</td>
<td>Design/Build, Consulting, Operations Management, Project Management, Support/Admin</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>US or Outside US</td>
<td>Within the Continental U.S. or outside the U.S.</td>
<td>1, 2</td>
</tr>
<tr>
<td>Summary Job Level</td>
<td>Top, Middle, and Lower Tier Employee</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>Gender</td>
<td>Male, Female</td>
<td>1, 2</td>
</tr>
<tr>
<td>Age ID</td>
<td>&lt; 25 years, 26-35 years, 36-45 years, 46-54 years, 55-59 years, 60-61 years, 62-64 years, &gt; 65 years</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
</tr>
<tr>
<td>YOS ID</td>
<td>0-2 years, 3-5 years, 6-10 years, 11-20 years, 21+ years</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>PEP Rating</td>
<td>Low, Low-Medium, Medium, Medium-High, High</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>Craft ID</td>
<td>Laborer, Non-laborer</td>
<td>1, 2</td>
</tr>
</tbody>
</table>

The remaining attributes we use for all the models we experiment with are shown in Table 3.5 along with a summary of their definitions and converted discrete values.

Table 3.5: Final Employee Attribute List for Analysis

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Short Name</th>
<th>Data Type</th>
<th>Attribute Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revenue Type</td>
<td>A1</td>
<td>Categorical</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>US or Outside US</td>
<td>A2</td>
<td>Categorical</td>
<td>1, 2</td>
</tr>
<tr>
<td>Summary Job Level</td>
<td>A3</td>
<td>Categorical</td>
<td>1, 2, 3</td>
</tr>
<tr>
<td>Gender</td>
<td>A4</td>
<td>Categorical</td>
<td>1, 2</td>
</tr>
<tr>
<td>Age ID</td>
<td>A5</td>
<td>Categorical</td>
<td>1, 2, 3, 4, 5, 6, 7, 8</td>
</tr>
<tr>
<td>YOS ID</td>
<td>A6</td>
<td>Categorical</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>PEP Rating</td>
<td>A7</td>
<td>Categorical</td>
<td>1, 2, 3, 4, 5</td>
</tr>
<tr>
<td>Craft</td>
<td>A8</td>
<td>Categorical</td>
<td>1, 2</td>
</tr>
</tbody>
</table>

While the company provided us with 110 non-mutually exclusive employee groups we limited our analysis to only 25 groups, given the size of the employee database and to simplify our analysis.
We first selected all 12 groups that contain “$YOS \leq 3$” because employees with less than 3 years of service had the highest leave frequency. Following this filtering, we randomly selected 13 of the remaining company-defined groups to complete our 25-group list. The purpose of selecting these remaining 13 groups is to demonstrate our model’s accuracy in estimating employee-group retention probabilities in groups that had no association with the $YOS \leq 3$ attribute filter. The groups we use for our analysis are shown in Table 3.6. It is worth noting that a “-” value in the table denotes that the corresponding attribute is not a member of the particular group.

Table 3.6: Final List of Groups with Associated Attribute Values

<table>
<thead>
<tr>
<th>Group Number</th>
<th>A1</th>
<th>A2</th>
<th>A3</th>
<th>A4</th>
<th>A5</th>
<th>A6</th>
<th>A7</th>
<th>A8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>7</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>5</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>15</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>16</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>17</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>18</td>
<td>-</td>
<td>-</td>
<td>3</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>19</td>
<td>-</td>
<td>-</td>
<td>2</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>20</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>2</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>21</td>
<td>4</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>22</td>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>23</td>
<td>5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>25</td>
<td>2</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
3.5.2 Experimental Setup

To evaluate the performance of our models, our experiments were organized as follows:

1. Partition data into training and test sets in preparation for split-validation.
2. Construct retention tree and KNN models using the training set.
3. Determine retention probabilities using the test set, that is, run test cases with the trees and KNN weights determined in step 2.

A potential problem with this approach is the imbalanced nature of the data sets, which is expected in a data set that is biased towards people that remain with the company. The survey report by Chawla (2005) provides a thorough overview on this issue, while Chawla et al. (2011) introduce a novel idea called Synthetic Minority Over-sampling Technique (SMOTE). SMOTE is a data sampling method that oversamples the minority (abnormal) class and undersamples the majority (normal) class. It should be noted, however, that this was validated using continuous features almost exclusively, while its use with nominal attribute variables is largely left as an area of future research. In our setting the employees that stayed with the company (majority class) far outnumbers the employees that left the company (minority class). To analyze the potential benefits of oversampling, we created four "balanced" employee data sets using the SMOTE-N algorithm proposed in Chawla et al. (2011) using various parameter settings. Specifically, these data sets were created by oversampling the minority class by 200% and 400% and undersampling the majority class by 100% and 200%. However, our results indicated that forecasting accuracy suffered when using the newly created data sets, as the mean absolute deviation errors used in our objective functions range from 24% to 52%. In our experiments, we therefore resorted to using our original data set and did not use stratified sampling methods. Specifically, we performed split-validation of our original data set by randomly selecting 80% of the data to serve as a training set, and the remaining 20% to serve as a test set.

To provide a comparison with our results, we also used the commercially available software SPSS Statistics to construct a CHAID decision tree. For the CHAID tree, we specified a minimum
node size of 250 employees (minimum employee-group size), and a maximum tree depth of 5-levels to avoid trivial solutions (i.e., a leaf node for each employee) and growing extremely large decision trees. We imposed a similar restriction on the group-based retention trees we constructed. Furthermore, we allowed CHAID to group more than one attribute per node.

### 3.5.3 Empirical Results

We conducted a number of experiments to assess the potential benefits of our group-based retention tree and KNN approaches. First, we considered the performance of our retention tree models, focusing on the impact of the randomization built into our constructions and on comparing the resulting retention trees’ performance relative to the greedy CHAID tree construction procedure. Subsequently, we evaluated the performance of using non-equal weights in our KNN procedure, and analyzed the overall prediction accuracy of various model alternatives.

To understand how the randomization we introduced into our retention tree construction can impact prediction accuracy, we first considered the impact of varying the size of the restricted candidate list (RCL) in the greedy randomized tree construction phase. Specifically, we constructed decision trees and evaluated retention trees using RCL size of 3 through 7, and for every RCL size we generated 10 retention trees. The results are shown in Table 3.7, which for each RCL size shows the worst, best, and average mean absolute error (as expressed in objective function 3.6) over the 10 instances when evaluated on the test set. Because the retention trees produced are stochastic, we also calculate the standard deviations for each RCL. These are approximately 0.2%, which indicates the stability of our GRASP retention tree algorithm.

The average error over all the RCLs we tested is 0.94% for the training set, and 1.62% for the test set. Surprisingly, the test set results indicate that, as the RCL increases (i.e., becomes more diverse) from 3 to 7, the mean absolute errors decrease. This illustrates that incorporating diversification strategies into the tree construction by allowing locally suboptimal attributes to occupy a node can lead to better overall retention trees. In addition, our experiments revealed that the procedure discovered high quality solutions rather quickly in the search process. Given
Table 3.7: GRASP Retention Tree Modeling Results (Test Set)

<table>
<thead>
<tr>
<th>Restricted Candidate List (RCL)</th>
<th>Worst Mean Absolute Error</th>
<th>Best Mean Absolute Error</th>
<th>Avg Mean Absolute Error</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>1.68%</td>
<td>1.50%</td>
<td>1.56%</td>
<td>0.06%</td>
</tr>
<tr>
<td>4</td>
<td>1.69%</td>
<td>1.49%</td>
<td>1.62%</td>
<td>0.08%</td>
</tr>
<tr>
<td>5</td>
<td>1.72%</td>
<td>1.34%</td>
<td>1.47%</td>
<td>0.12%</td>
</tr>
<tr>
<td>6</td>
<td>1.77%</td>
<td>1.15%</td>
<td>1.48%</td>
<td>0.20%</td>
</tr>
<tr>
<td>7</td>
<td>1.69%</td>
<td>1.15%</td>
<td>1.34%</td>
<td>0.16%</td>
</tr>
</tbody>
</table>

the maximum of ten candidate trees (each of which were further modified using the local improvement procedure), a high quality solution is usually discovered within the first four iterations and the retention tree rarely improves after that. Though more compact tree structure are generally preferred due to increased ease of interpretation, our results indicated that high-quality retention tree generally have at least 60 nodes.

To further evaluate the performance of our group-based retention trees, we also compared their prediction accuracy with the CHAID decision tree that was constructed using the training set. A graphical illustration of this CHAID tree is partially shown in Figure 3.3. While the overall objective behind this tree construction is to predict the retention probability of individual employees, we can calculate its mean absolute error in predicting employee group retention probabilities using expression 3.6 on the test set. We compared these with the errors that were obtained using the best performing group retention tree when a RCL of 4 was used, which yields a worst-case comparison in that using a RCL of 4 yielded the poorest accuracy in our initial tests. The resulting retention tree is shown in Figure 3.4. A detailed comparison is shown in Table 3.8. For each of the 25 groups in the data set, the table shows the actual retention probabilities of the employee groups in the test set, as well as the estimated group retention probabilities and mean absolute deviations for both the GRASP and CHAID retention trees. The results in the table show that our retention trees produce better estimates for all but one employee group (group 22). Moreover, the overall mean absolute deviation is 6.63% for the CHAID tree and 1.72% for the retention tree.

Further analysis of the departing personnel provides a more comprehensive understanding
Figure 3.3: CHAID Results: Partial Decision Tree Illustration (Test Set).

Note: Each node represents the population (number of employees) that reside under a specific node-attribute pair.

Figure 3.4: Best GRASP Retention Tree for RCL = 4 (first generated tree, Test Set).

Note: A node with a $A\#$ symbol denotes the attribute number, while $e$ denotes a terminal node.
Table 3.8: Group Probabilities Generated by Retention Tree (RCL = 4) vs. CHAID (Test Set)

<table>
<thead>
<tr>
<th>Group Number</th>
<th>Actual Probability</th>
<th>Retention Tree Estimate</th>
<th>CHAID Tree Estimate</th>
<th>AbsDev</th>
<th>CHAID AbsDev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>92.60%</td>
<td>92.11%</td>
<td>86.60%</td>
<td>0.49%</td>
<td>6.00%</td>
</tr>
<tr>
<td>2</td>
<td>93.38%</td>
<td>93.19%</td>
<td>88.48%</td>
<td>0.19%</td>
<td>4.90%</td>
</tr>
<tr>
<td>3</td>
<td>95.06%</td>
<td>94.33%</td>
<td>90.90%</td>
<td>0.73%</td>
<td>4.16%</td>
</tr>
<tr>
<td>4</td>
<td>96.56%</td>
<td>96.36%</td>
<td>93.12%</td>
<td>0.20%</td>
<td>3.44%</td>
</tr>
<tr>
<td>5</td>
<td>90.04%</td>
<td>92.52%</td>
<td>84.52%</td>
<td>2.48%</td>
<td>5.52%</td>
</tr>
<tr>
<td>6</td>
<td>88.86%</td>
<td>88.66%</td>
<td>81.99%</td>
<td>0.20%</td>
<td>6.87%</td>
</tr>
<tr>
<td>7</td>
<td>98.79%</td>
<td>96.30%</td>
<td>94.85%</td>
<td>2.49%</td>
<td>3.94%</td>
</tr>
<tr>
<td>8</td>
<td>92.60%</td>
<td>89.40%</td>
<td>81.33%</td>
<td>3.20%</td>
<td>11.27%</td>
</tr>
<tr>
<td>9</td>
<td>94.99%</td>
<td>95.61%</td>
<td>91.12%</td>
<td>0.62%</td>
<td>3.87%</td>
</tr>
<tr>
<td>10</td>
<td>90.33%</td>
<td>93.15%</td>
<td>84.74%</td>
<td>2.82%</td>
<td>5.59%</td>
</tr>
<tr>
<td>11</td>
<td>94.49%</td>
<td>94.37%</td>
<td>90.22%</td>
<td>0.12%</td>
<td>4.27%</td>
</tr>
<tr>
<td>12</td>
<td>94.21%</td>
<td>93.12%</td>
<td>88.13%</td>
<td>1.09%</td>
<td>6.08%</td>
</tr>
<tr>
<td>13</td>
<td>84.06%</td>
<td>83.07%</td>
<td>76.57%</td>
<td>0.99%</td>
<td>7.49%</td>
</tr>
<tr>
<td>14</td>
<td>85.07%</td>
<td>86.12%</td>
<td>76.86%</td>
<td>1.05%</td>
<td>8.21%</td>
</tr>
<tr>
<td>15</td>
<td>89.77%</td>
<td>93.91%</td>
<td>84.27%</td>
<td>4.14%</td>
<td>5.50%</td>
</tr>
<tr>
<td>16</td>
<td>92.48%</td>
<td>91.95%</td>
<td>84.97%</td>
<td>0.53%</td>
<td>7.51%</td>
</tr>
<tr>
<td>17</td>
<td>83.86%</td>
<td>84.92%</td>
<td>75.40%</td>
<td>1.06%</td>
<td>8.46%</td>
</tr>
<tr>
<td>18</td>
<td>94.44%</td>
<td>93.94%</td>
<td>87.83%</td>
<td>0.50%</td>
<td>6.61%</td>
</tr>
<tr>
<td>19</td>
<td>95.65%</td>
<td>93.24%</td>
<td>88.07%</td>
<td>2.41%</td>
<td>7.58%</td>
</tr>
<tr>
<td>20</td>
<td>85.58%</td>
<td>86.4%</td>
<td>77.55%</td>
<td>0.82%</td>
<td>8.03%</td>
</tr>
<tr>
<td>21</td>
<td>88.4%</td>
<td>88.29%</td>
<td>80.42%</td>
<td>0.11%</td>
<td>7.98%</td>
</tr>
<tr>
<td>22</td>
<td>68.21%</td>
<td>77.14%</td>
<td>59.39%</td>
<td>8.93%</td>
<td>8.82%</td>
</tr>
<tr>
<td>23</td>
<td>85.51%</td>
<td>90.31%</td>
<td>79.32%</td>
<td>4.80%</td>
<td>6.19%</td>
</tr>
<tr>
<td>24</td>
<td>88.49%</td>
<td>86.19%</td>
<td>80.26%</td>
<td>2.30%</td>
<td>8.23%</td>
</tr>
<tr>
<td>25</td>
<td>87.05%</td>
<td>87.81%</td>
<td>77.91%</td>
<td>0.76%</td>
<td>9.14%</td>
</tr>
</tbody>
</table>

Mean - - - 1.72% 6.63%

of company retention behavior. Figure 3.5 shows a Pareto chart (see Montgomery, 1991) with personnel losses and their respective contribution to the cumulative company losses. The figure shows that group 14 (personnel less than 25 years of age), which has a 85% retention rate, represents 12% of the losses. Similarly, group 6 (personnel located outside the U.S.), with a 89% retention rate, represents 11 of the losses. On the other hand, Group 22 (personnel working in design/build projects and with less than 2 years of service), which showed the largest prediction error in the results (Table 3.8), had a 68% retention rate yet presents less than 2% of the losses. Given the
group’s low population size, the company may choose not to target this group’s retention rate. This can be a relevant insight, since the company likely wants to focus on the retention behavior of larger employee groups. Next, we evaluate our KNN approach. In our experiments, we varied the number of nearest neighbors considered by letting K be equal to 3, 4, 5, and 7. In addition, we evaluated the models both with and without the proximity constraint 3.13. The results are in Table 3.9, which for each of the case shows both the mean absolute prediction error on the test set and the weight found by our procedure. Surprisingly, using the proximity constraints 3.13 (as suggested by Atiya(2005)) yields solutions with equal weights. Optimizing the weights without this proximity constraints, however, did yield unequal weights, and led to better performance. In fact, the latter approach produced better results than all other models considered in this chapter. Though perhaps somewhat unintuitive, this is achieved by increasing the weights on more distant neighbors. We note, however, that the computation time require to determine these weight is substantially higher than the time needed to construct the retention trees. We conclude this section with an overall comparison of the various model alternatives. First, Table 3.10 shows the
Table 3.9: KNN Modeling Results (Test Set)

<table>
<thead>
<tr>
<th>Model</th>
<th>K</th>
<th>Mean Absolute Error</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free</td>
<td>3</td>
<td>0.10%</td>
<td>(0.234, 0.766)</td>
</tr>
<tr>
<td>Free</td>
<td>4</td>
<td>0.17%</td>
<td>(0.023, 0.098, 0.877)</td>
</tr>
<tr>
<td>Free</td>
<td>5</td>
<td>0.08%</td>
<td>(0, 0, 0.239, 0.761)</td>
</tr>
<tr>
<td>Free</td>
<td>7</td>
<td>0.12%</td>
<td>(0, 0, 0.151, 0.0920, 0.457, 0.300)</td>
</tr>
<tr>
<td>Constrained</td>
<td>3</td>
<td>1.0%</td>
<td>(0.333, 0.333, 0.333)</td>
</tr>
<tr>
<td>Constrained</td>
<td>4</td>
<td>0.98%</td>
<td>(0.25, 0.25, 0.25, 0.25)</td>
</tr>
<tr>
<td>Constrained</td>
<td>5</td>
<td>0.98%</td>
<td>(0.25, 0.25, 0.25, 0, 0)</td>
</tr>
<tr>
<td>Constrained</td>
<td>7</td>
<td>0.90%</td>
<td>(0.1429, 0.1429, 0.1429, 0.1429, 0.1429, 0.1429, 0.1429)</td>
</tr>
</tbody>
</table>

mean absolute prediction error for the best performing variant of each of the models considered in this section. Clearly, our GRASP retention tree algorithm outperforms the CHAID tree when predicting employee group retention probabilities. In fact, CHAID does not perform much better than a simple naive approach. However, the KNN approach performs better than all other models, in particular for higher values of $K$.

Table 3.10: Comparison of the Best Modeling Results: GRASP Retention Tree, KNN, and CHAID (Test Set)

<table>
<thead>
<tr>
<th>Model</th>
<th>Type</th>
<th>Mean Absolute Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>RetentionTree</td>
<td>RCL = 7</td>
<td>1.34%</td>
</tr>
<tr>
<td>KNN Free</td>
<td>K = 5</td>
<td>0.08%</td>
</tr>
<tr>
<td>KNN Constrained</td>
<td>K = 7</td>
<td>0.90%</td>
</tr>
<tr>
<td>CHAID</td>
<td>-</td>
<td>6.63%</td>
</tr>
<tr>
<td>Naive</td>
<td>-</td>
<td>9.00%</td>
</tr>
</tbody>
</table>

As stated before, our data set is imbalanced in that approximately 91% of company employees stayed with the company. As a result, a naive method (which predicts that all employees stay with the company) would already achieve 91% predictive accuracy. To better understand our models’ performance, we therefore also analyze their ability to predict the small fraction of employees that departed the company.

A commonly used metric to evaluate the minority class prediction is the so-called $F_1$ score (see Goutte and Gaussier, 2005, for a detailed overview), which combines precision and recall measures.
Within our context, recall is defined as the number of employees that the model correctly predicted to leave the company (i.e., true positives), divided by the total number of people that actually left the company (i.e., true positives plus false negatives). Precision, on the other hand, is defined as the number of employees that the model correctly predicted to leave the company (i.e., true positives), divided by the total number of people that the model predicted to leave the company (i.e., true positives plus false positives). The $F_1$ score we use combines these measures using the harmonic mean method, that is, $F_1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$, where equal weight is given to the precision and recall measures. The higher the $F_1$ score, the better the model’s accuracy in predicting the minority class. We note that a potential limitation of this metric for our purposes is that it emphasizes the prediction of individual employees, while our models have been designed to predict the retention probability of employee groups.

Table 3.11 shows the $F_1$ score for the models discussed in this chapter. As stated before in Section 3.5.3, we restrict CHAID to a 5-level tree depth in order to avoid creating small terminal node size populations. Based on the $F_1$ score, our models’ accuracy in predicting the minority class ranges from a rate of 12.3% to 21.8%, while the predictive accuracy of the CHAID tree equals 14.1%. Because recall and precision are weighed equally, the $F_1$ scores are low for all models; yet even if more emphasis were to be placed on the precision metric in calculating $F_1$ scores, all models would produce results no greater than 48%. Overall, it is not surprising that the CHAID tree performs better at predicting the minority class than our best retention tree (RCL = 7) and two of our KNN models, as its objective is to classify individual entities in a data set. Our models’ objective, on the other hand, is to minimize the deviation between the true and estimated employee-group probabilities. We note though that the low $F_1$ score indicates that CHAID performs rather poorly at its primary objective of predicting the retention propensity of individual employees, and using models that are based on determining employee-group retention propensity will also produce individual employee retention classifications of the minority class with almost equivalent accuracy.
Table 3.11: Precision, Recall and $F_1$ Scores (Rates) for All Models (Test Set)

<table>
<thead>
<tr>
<th>Model</th>
<th>Parameter</th>
<th>Average Precision</th>
<th>Average Recall</th>
<th>Average $F_1$ Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>RetentionTree</td>
<td>RCL = 7</td>
<td>0.4718</td>
<td>0.0708</td>
<td>0.1230</td>
</tr>
<tr>
<td>KNN Free</td>
<td>K = 3</td>
<td>0.1844</td>
<td>0.1925</td>
<td>0.1884</td>
</tr>
<tr>
<td>KNN Free</td>
<td>K = 4</td>
<td>0.2182</td>
<td>0.2186</td>
<td>0.2184</td>
</tr>
<tr>
<td>KNN Free</td>
<td>K = 5</td>
<td>0.2182</td>
<td>0.2186</td>
<td>0.2184</td>
</tr>
<tr>
<td>KNN Free</td>
<td>K = 7</td>
<td>0.2073</td>
<td>0.2064</td>
<td>0.2069</td>
</tr>
<tr>
<td>KNN Constrained</td>
<td>K = 3</td>
<td>0.3079</td>
<td>0.1566</td>
<td>0.2077</td>
</tr>
<tr>
<td>KNN Constrained</td>
<td>K = 4</td>
<td>0.4151</td>
<td>0.0801</td>
<td>0.1344</td>
</tr>
<tr>
<td>KNN Constrained</td>
<td>K = 5</td>
<td>0.4151</td>
<td>0.0801</td>
<td>0.1344</td>
</tr>
<tr>
<td>KNN Constrained</td>
<td>K = 7</td>
<td>0.4423</td>
<td>0.1117</td>
<td>0.1784</td>
</tr>
<tr>
<td>CHAID</td>
<td>5-level depth</td>
<td>0.6429</td>
<td>0.7900</td>
<td>0.1406</td>
</tr>
</tbody>
</table>

3.6 Conclusions and Further Research

The main contribution of our research is the introduction of methods to estimate the retention probabilities of certain strategically important employee groups, as opposed to estimating the retention propensity of individual employees. To achieve this, we modify commonly used decision tree and KNN methods, and our research shows that our models produce better results than the well-known CHAID tree method.

There are several possible avenues for further research. From an algorithmic perspective, our randomized tree construction procedure might be improved by optimally determining the size of the restricted candidate list. Another possible improvement could be to extend the local search method by simultaneously selecting a set of internal nodes delete and reconstruct, rather than deleting and reconstructing a single node. Moreover, one could expand the procedure by conducting swaps of internal nodes, or by swapping internal nodes and their respective child nodes. However, the concept of swapping in the context of decision trees can yield numerous complications. As an alternative to using GRASP in the search procedure, one could also use Tabu Search algorithms.

From an application-oriented perspective, it could be useful to predict the time until departure, instead of the departure decision itself. This might enable us to predict not only which employees will leave, but also when they will leave. Ultimately, the lack of available data prior to
2010 was judged to be prohibitive to pursuing this approach. In addition, one could develop sur-
vivability profiles for each employee group to track retention rate changes during a specific number of periods. This type of time series analysis can be valuable in that one can potentially forecast employee retention rates under the assumption that some attributes change over time. Lastly, one could also consider approaches that themselves attempt to determine employee groups that have a high risk of departure.
Chapter 4

Surrogate Data Representations for Feature Selection and Prediction

4.1 Introduction

With the rapid development of computer technology in recent years, public and private organizations now more than ever have an interest in analyzing their data resources. A desire to utilize large data sets for making predictions concerning future events such as projected sales, marketing demographics, and employee retention propensities has brought the development of such predictive models to the forefront.

However, even with the advancement in computer processing capabilities using these models based on an entire data set can be difficult at times, and management’s desire for an immediate response to their “what if” scenarios can make the task of forecasting a daunting task. Given the need to provide accurate answers to questions in real time, reducing the size of data sets without significantly compromising forecasting accuracy is highly desirable. Given such reduced data sets, machine learning algorithms can then assist the data analyst.

For example, an important task in supervised learning is inferring a function from labeled training data (Witten and Frank, 2005). One example of labeled data is a response variable having a binary response. The objective of the training data is to properly define a function that will correctly determine the class labels for unseen examples (instances), based on the features of that instance. That is, by designating a large data set as the training set, one tries to discern a desired output value (response) of unseen instances based on the training data’s input. One way to reduce the computational complexity of learning and predictive algorithms is to reduce the number of
instances and features in a data set necessary to perform quantitative analysis. Our use of the term “surrogate data representations” refers to reducing the number of columns (features), reducing the number of rows (data sampling), or reducing both in unison.

Not using all the features available can save on the cost of measuring features that are not selected. Occasionally, a model with only a subset of features may even perform better than one with all features. In addition, selecting only certain features can provide insights into the most important features of the data for other types of inferential analysis. Hence, the act of selecting a subset of features is a key step to making inferences from data sets with a large number of instances or features in a data set (Navot et al., 2005). At times, however, reducing not only the number of features but also the number of training examples (observations) may be necessary in order to conduct timely and accurate inferential statistics for extremely large data sets.

Similar to feature selection, data sampling reduces the computational complexity of learning and predictive algorithms. Instead of reducing the number of features, data sampling (also referred to as surrogate data selection) reduces the number of instances, making predictive analysis less computationally expensive. One disadvantage of using surrogate data instead of the entire data set is the potential loss of information contained within the excluded instances. However, it is possible to use statistical approaches to minimize the loss of information that results from feature selection and data sampling.

Determining the appropriate sample size and the subset of features needed to make inferences on a data set is standard practice in the field of statistics. Current statistical methods for sampling large data sets usually consist of selecting data from these large data sets and forming new, smaller data sets via a probabilistic design using random sampling methods such as Monte Carlo sampling (Hastings, 1970). However, these methods do not guarantee that the subset of data will be a fitting proxy for the large data source. A random sample may or may not fairly represent the original data (the population) as a whole (Kruskal and Mosteller, 1980). For this reason, we propose deterministic surrogate data sampling to maximize representativeness of the original data. Although deterministic data sampling may introduce bias when performing inferential statistics on
the newly created smaller data set, we believe it is not an issue in our approach if the input data set is “attribute diverse” among observations. We address this potential shortfall in the next section.

The feature selection approach we propose is similar to the common statistical procedure known as principal component analysis (PCA) (see Jolliffe, 2005). Unlike PCA, which uses orthogonal transformations and a covariance matrix to determine a subset of linearly uncorrelated variables, our method does not rely on traditional statistical measures to determine the most relevant features in a data set. In addition, our approach is not restricted to a parametric set of restrictions. For example, a necessary requirement for the principal components to be considered independent is that the data be normally distributed.

Our main contribution is a set of deterministic surrogate data representations that not only reduce the sample size of our data set but also determine its most relevant features for accurate representation of the original data set. Previous research suggests that feature selection ideas based on optimization can be more effective than those based on classical statistical analysis (Eksioglu et al., 2005). Consequently, we develop an optimization-based method that is competitive with classical approaches. Given the surrogate data results and the utilization of a diversity measure among features, we introduce a simple feature selection strategy and two novel quadratic programming approaches to conduct feature selection for predictive modeling. In empirical experiments utilizing a U.S. Army enlisted personnel data set, both of our quadratic programming (QP) modeling approaches accurately identify the most important soldier features that determine a soldier’s propensity to remain in the Army.

The results of our experiment indicate that our methodology improves forecasting accuracy when compared to a traditional probabilistic data sampling method. The loss in predictive accuracy is minimal when compared to analysis conducted on the entire data set. Additionally, combining surrogate data sampling and feature selection performs well with respect to predictive accuracy compared to conducting the two independently. Together, surrogate data sampling and feature selection conserve computational time which allows the data analyst to focus his or her’s computational resources to other analytical tasks such as advanced inferential statistics and better synthesis
of numerical experiment results.

The remainder of this section is organized as follows. First, we present a literature review on data mining approaches related to our methods. Second, we explain our surrogate data sampling method, which involves the creation of so-called data tranches. Third, we introduce the concept of co-diversity and a simple feature selection method. Fourth, we introduce two binary quadratic optimization methods using our notion of co-diversity and demonstrate how surrogate data can be used for feature selection and forecasting. Lastly, we provide a real-world empirical example using employee data from the U.S. Army to compare our methods to traditional methods.

4.2 Literature Review

The surrogate data representations we introduce in this chapter fall within the field of data mining, specifically in pattern recognition and subset selection. Pattern recognition has been studied extensively, and a thorough review of the literature on this subject is beyond the scope of this chapter. We therefore focus on material that is most relevant to our research. When we refer to surrogate data representation, our intent is to select a subset of observations (rows). In the literature, the term surrogate data is most commonly used in the area of signal processing, where the goal is to determine whether an observed time series is deterministic. For discussions on this topic, see Small and Tse (2003).

When we use the term data sampling, we refer to sampling from a large data source. However, in the literature, the use of data sampling primarily refers to sampling a population of people for the purpose of conducting surveys or polls. For an extensive collection of this type of data sampling methods, we refer to Moore and McCabe (2005) and the seminal works of Cochran (1997) and McKay et al. (1979a). Readers unfamiliar with traditional sampling may consult Giunta (2002), which provides a review of data sampling methods including Monte Carlo sampling, Latin hypercube sampling, and orthogonal array sampling, while Guyon and Elisseeff (2003a) provide a thorough introduction to the sample selection (surrogate data) problem.

Related to our work, Heckman (1976) provides a thorough discussion on the sample selection
bias (errors) that occurs from performing deterministic data sampling on an incomplete data set. He explains that sample selection bias refers to the error attributed to non-randomly selecting a subset of a population and making inaccurate inferences from the sample. Citing one example of this type of bias, Heckman states, "Suppose a data analyst wishes to estimate the determinants of wage offers, but has access to wage observations for only those who work. Since people who work are selected non-randomly from the population, estimating the determinants of wages from the subpopulation who work may introduce bias." Although Heckman (1976) identifies this problem with deterministic sample selection, our deterministic methods do not target a particular population of observations and then make inferences on their behavior. For example, we do not try to make inferences on an employee’s retention propensity by using employee records of only those individuals that remain with the company. Hence, it is unlikely that we could potentially introduce bias into our samples if we are cognizant of our data selection strategy.

Ghysels et al. (2004b) propose a Mixed Data Sampling (MIDAS) set of regression models. Their models consider time series data, which is characterized as sampling data at different time intervals. This type of sampling is more appropriate for research in the area of macroeconomics and finance. Similarly, Wooldridge (1995) proposes new methods for testing and correcting for sample selection bias in panel data models. He asserts that his methods are more robust when compared to maximum likelihood methods. Tourassi and Floyd (1997) discuss the use of data sampling in the medical field, while Linhart and Zucchini (1986) derive a sample selection criteria using a multinomial (probability) distribution function.

Additionally, Tibshirani (1996) provides a pseudo-surrogate data method for shrinkage and selection for regression problems. The author does not focus on subsets but defines a continuous shrinking operation that produces coefficients for the regression model and provides an empirical example of his model using prostate cancer data. Kyriazidou (1997) proposes estimators for a sample-selection model from panel data with individual-specific effects. The author addresses the bias problem that may occur whenever one tries to estimate population parameters from a non-random sample. According to Kyriazidou (1997), a non-random sample is defined as one
that contains only individuals that possess a certain set of characteristics, fall within a certain demographic group, or voluntarily agree to be in the sample. Again, our methods do not suffer from this bias since this type of sampling is usually reserved for marketing research or for purposeful demographic surveying.

For a detailed discussion on subset (feature) selection, we refer the reader to Bishop (2006). In the area of exact feature selection methods, Narendra and Fukunaga (1977b) propose an efficient subset enumeration scheme without resorting to an exhaustive search. However, the authors state that branch and bound algorithms in feature selection require monotone objective functions, which can be problematic. In their algorithm, the feature selection monotonicity requirement means that a subset of features should not perform better than any larger set that contains the subset. The monotonicity requirement is too restrictive for our application since our combinatorial optimization algorithm must be able to freely search the entire solution space. For example, in the feature selection methods we propose, we permit a data set with a small number features to outperform a data set with a large number of features. Also, allowing additional features to augment a data set may introduce data “noise” and limit our ability to conduct the most accurate statistical analysis.

In the field of search heuristics, Tahir et al. (2004b) introduce a tabu search feature selection heuristic for improving classification rates in the medical field. Their algorithm uses intermediate-term memory to reduce the number of features extracted from multi-spectral images. In lab experiments, their method has proved effective for the classification of prostate needle biopsies. Yang and Honavar (1998) present a genetic algorithm to tackle the feature selection problem by introducing different cost and performance constraints.

The most relevant research related to our work is presented by Eksioglu et al. (2005). The authors propose a quadratic programming (QP) method for feature selection but use multiple linear regression for the forecasting component of classification. The criteria the authors use for feature selection is based on maximizing the correlation between the response variable and features, and minimizing the correlation between features. Unlike their method of using correlations and the linearization of their QP, we attempt to maximize the diversity between features and minimize
the diversity between each feature and the response variable. We diversify interactions between features without using the correlation statistical measure, but introduce a metric we refer to as “co-diversity.” Furthermore, Eksioglu et al. (2005) use the adjusted $R^2$ measure to represent model accuracy which is relevant when they compare their algorithm’s speed and accuracy to other published methods. In contrast, the focus of our set of models is not a faster model to conduct feature selection, but rather a more accurate way of extracting surrogate data and a subset of features with which to conduct forecasting and classification. We elaborate on Eksioglu et al.’s (2005) method in Section 4.7.4.

Although the above-mentioned strategies in this review have proved effective, none has provided a set of deterministic models that provide sample-size reduction and feature selection reduction for the purposes of classification and forecasting. The modeling ideas in this chapter are motivated by the initial work in Glover and Kochenberger (2010). We now explain the notation we will use throughout this chapter before introducing our models.

### 4.3 Preliminaries

Prior to explaining our tranche generation process, we begin by introducing our notation. The input data can be viewed as an $m \times n$ matrix $X$, where rows correspond to observations and columns to attributes. Let $M = \{1, \ldots, m\}$, and let $N = \{1, \ldots, n\}$. Next, let $X = \{x(i) : i \in M\}$ be a collection of vectors that consists of the entire data set. Therefore, $i$ represents the index (observation identification), while $M$ represents a collection of all observation indices. The vectors $x(i) = (x_1(i), \ldots, x_n(i))$ are observations (rows of data) and assumed to be normalized so that each variable (feature) value $x_j(i), j \in N$, ranges from 0 to 1. Hence, $\min_{i \in M} x_j(i) = 0$ and $\max_{i \in M} x_j(i) = 1$.

Let us explain our modeling assumptions. First, the collection of data, $X$, must represent scalar data. This assumption is essential to our approach, and implies that we work exclusively with continuous data. Hence, our approach is not applicable when the input data has nominal/categorical features. Next, each observation is considered independently of the others. For example, in our
Army personnel data set, each soldier decides to stay or leave the Army based on his/her decision (independent) and not influenced by a peer’s retention decision.

In order to describe our approach, we first explain our idea of a tranche. We define a tranche as a bucket/bin that consists of a collection of $x_j(i)$ (column of data), or $x(i)$ (row of data) vectors some of whose elements fall within certain upper and lower bound values. The upper and lower bound values are determined iteratively, and the higher the number of iterations, the larger the number of bins are created. It is important to note that the upper and lower bound values of a tranche are unique for each feature, that is, branches do not overlap. The concept of a tranche is very important to our models because we conduct feature selection and data sampling from key tranches that we infer contain a rich collection of input data.

In the next section, we begin by explaining the concept of generating data tranches and extracting surrogate data representations for use in classification and clustering.

4.4 Generating Tranches

Prior to conducting subset data and feature selection, we explain how we generate tranches from a large data set. As previously mentioned, tranches are essentially subsets of vectors in which membership is determined based on the upper and lower limits of the tranche. Specifically, let $\hat{M} \subseteq M$ be some subset of observations, and let $L_j(\hat{M}) = \min_{i \in \hat{M}} x_j(i)$ and $U_j(\hat{M}) = \max_{i \in \hat{M}} x_j(i)$ for some attribute $j$; these are the boundaries of the subset $\hat{M}$. Let $C_j(\hat{M})$ denote a “center” of the interval from $L_j(\hat{M})$ to $U_j(\hat{M})$. For example, $C_j(\hat{M})$ may be specified as the average of the lower and upper values of a tranche, $C_j(\hat{M}) = \left( L_j(\hat{M}) + U_j(\hat{M}) \right) / 2$, or can be expressed as the mean of the tranche members, $C_j = \frac{1}{|\hat{M}|} \sum_{i \in \hat{M}} x_j(i)$.

By iteratively partitioning subsets according their center, we generate tranches of data for each feature $j$. This process is illustrated in the Tranche Generation Method shown in Algorithm 4. In the generation process, we define a set $S_j$ that contains each tranche that is generated during an iteration, and let $\hat{\ell}$ denote the maximum number of iterations. For each feature $j$, an iteration of the tranche-generation method begins with the set $S_j$ consisting of a given collection of tranches.
and ends with a set $S'_j$ consisting of the new collection of tranches generated from $S_j$. The set $S'_j$ is produced by partitioning each tranche into two new tranches. The process continues for a small number of iterations.

**Algorithm 4** Tranche Generation Method

```plaintext
for each $j$ do
    Let $\hat{M} \in M$, $S_j = \{M\}$, and $S'_j = \emptyset$
    for Iter = 1 to $\hat{\ell}$ do
        for each $S_j$ do
            Identify $\hat{M}^L = \{i \in \hat{M} : L_j(\hat{M}) \leq x_j(i) \leq C_j(\hat{M})\}$
            Identify $\hat{M}^U = \{i \in \hat{M} : C_j(\hat{M}) < x_j(i) \leq U_j(\hat{M})\}$
            $S'_j = \{\hat{M}^L\} \cup \{\hat{M}^U\} \cup S'_j$
        end for
        Let $S_j = S'_j$ and $S'_j = \emptyset$
    end for
end for
```

To illustrate the tranche generation process, we use the Belle Ayr Coal Mine data set shown in Table 4.4. The Belle Ayr coal mine is located in Gillette, Wyoming. Since a publication by Cronauer et al. (1978) presented the results of the coal mine's process of using solvents for coal liquefaction, the data set is widely used for comparing various regression approaches. The data consists of 26 observations and 7 features.

Using the data shown in Table 4.4 as input into our model, the tranche generator produces the following output by setting the parameter $l$ to 3 (eight tranches):

<table>
<thead>
<tr>
<th>Feature Number: 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>{0, 9, 25}</td>
</tr>
<tr>
<td>{13, 14, 16}</td>
</tr>
<tr>
<td>{10, 11, 23}</td>
</tr>
<tr>
<td>{18, 20, 21, 24, 26}</td>
</tr>
<tr>
<td>{1, 2, 17, 22}</td>
</tr>
<tr>
<td>{4, 12, 15, 19}</td>
</tr>
<tr>
<td>{3, 5, 6}</td>
</tr>
<tr>
<td>{7, 8}</td>
</tr>
</tbody>
</table>
**Table 4.1: Belle Ayr Coal Mine Data**

<table>
<thead>
<tr>
<th>Run Number</th>
<th>$CO_2$ Level</th>
<th>Expansion Time (min)</th>
<th>Temperature ($^\circ$ C)</th>
<th>Dissolution (%)</th>
<th>Oil Yield (g/100g MAF)</th>
<th>Total Coal</th>
<th>Total Solvent</th>
<th>Hydrogen Consumption</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>36.98</td>
<td>5.1</td>
<td>400</td>
<td>51.37</td>
<td>4.24</td>
<td>1484.83</td>
<td>2227.25</td>
<td>2.06</td>
</tr>
<tr>
<td>2</td>
<td>13.74</td>
<td>26.4</td>
<td>400</td>
<td>72.33</td>
<td>30.87</td>
<td>289.94</td>
<td>434.9</td>
<td>1.33</td>
</tr>
<tr>
<td>3</td>
<td>10.08</td>
<td>23.8</td>
<td>400</td>
<td>71.44</td>
<td>33.01</td>
<td>320.79</td>
<td>481.19</td>
<td>0.97</td>
</tr>
<tr>
<td>4</td>
<td>8.53</td>
<td>46.4</td>
<td>400</td>
<td>79.15</td>
<td>44.61</td>
<td>164.76</td>
<td>247.14</td>
<td>0.62</td>
</tr>
<tr>
<td>5</td>
<td>36.42</td>
<td>7.0</td>
<td>450</td>
<td>80.47</td>
<td>33.84</td>
<td>1097.26</td>
<td>1645.89</td>
<td>0.22</td>
</tr>
<tr>
<td>6</td>
<td>26.59</td>
<td>12.6</td>
<td>450</td>
<td>89.9</td>
<td>41.26</td>
<td>605.06</td>
<td>907.59</td>
<td>0.76</td>
</tr>
<tr>
<td>7</td>
<td>19.07</td>
<td>18.9</td>
<td>450</td>
<td>91.48</td>
<td>41.88</td>
<td>405.37</td>
<td>608.05</td>
<td>1.71</td>
</tr>
<tr>
<td>8</td>
<td>5.96</td>
<td>30.2</td>
<td>450</td>
<td>98.6</td>
<td>70.79</td>
<td>253.7</td>
<td>380.55</td>
<td>3.93</td>
</tr>
<tr>
<td>9</td>
<td>15.52</td>
<td>53.8</td>
<td>450</td>
<td>98.05</td>
<td>66.82</td>
<td>142.27</td>
<td>213.4</td>
<td>1.97</td>
</tr>
<tr>
<td>10</td>
<td>56.61</td>
<td>5.6</td>
<td>400</td>
<td>55.69</td>
<td>8.92</td>
<td>1362.24</td>
<td>2043.36</td>
<td>5.08</td>
</tr>
<tr>
<td>11</td>
<td>26.72</td>
<td>15.1</td>
<td>400</td>
<td>66.29</td>
<td>17.98</td>
<td>507.65</td>
<td>761.48</td>
<td>0.6</td>
</tr>
<tr>
<td>12</td>
<td>20.8</td>
<td>20.3</td>
<td>400</td>
<td>58.94</td>
<td>17.79</td>
<td>377.6</td>
<td>566.4</td>
<td>0.9</td>
</tr>
<tr>
<td>13</td>
<td>6.99</td>
<td>48.4</td>
<td>400</td>
<td>74.74</td>
<td>33.94</td>
<td>158.05</td>
<td>237.08</td>
<td>0.63</td>
</tr>
<tr>
<td>14</td>
<td>45.93</td>
<td>5.8</td>
<td>425</td>
<td>63.71</td>
<td>11.95</td>
<td>130.66</td>
<td>1961.49</td>
<td>2.04</td>
</tr>
<tr>
<td>15</td>
<td>43.09</td>
<td>11.2</td>
<td>425</td>
<td>67.14</td>
<td>14.73</td>
<td>682.59</td>
<td>1023.89</td>
<td>1.57</td>
</tr>
<tr>
<td>16</td>
<td>15.79</td>
<td>27.9</td>
<td>425</td>
<td>77.65</td>
<td>34.49</td>
<td>274.2</td>
<td>411.3</td>
<td>2.38</td>
</tr>
<tr>
<td>17</td>
<td>21.6</td>
<td>5.1</td>
<td>450</td>
<td>67.22</td>
<td>14.48</td>
<td>1496.51</td>
<td>2244.77</td>
<td>0.32</td>
</tr>
<tr>
<td>18</td>
<td>35.19</td>
<td>11.7</td>
<td>450</td>
<td>81.48</td>
<td>29.69</td>
<td>652.43</td>
<td>978.64</td>
<td>0.44</td>
</tr>
<tr>
<td>19</td>
<td>26.14</td>
<td>16.7</td>
<td>450</td>
<td>83.88</td>
<td>26.33</td>
<td>458.42</td>
<td>687.62</td>
<td>8.82</td>
</tr>
<tr>
<td>20</td>
<td>8.6</td>
<td>24.8</td>
<td>450</td>
<td>89.38</td>
<td>37.98</td>
<td>312.25</td>
<td>468.38</td>
<td>0.02</td>
</tr>
<tr>
<td>21</td>
<td>11.63</td>
<td>24.9</td>
<td>450</td>
<td>79.77</td>
<td>25.66</td>
<td>307.08</td>
<td>460.62</td>
<td>1.72</td>
</tr>
<tr>
<td>22</td>
<td>9.59</td>
<td>39.5</td>
<td>450</td>
<td>87.93</td>
<td>22.36</td>
<td>193.61</td>
<td>290.42</td>
<td>1.88</td>
</tr>
<tr>
<td>23</td>
<td>4.42</td>
<td>29</td>
<td>450</td>
<td>79.5</td>
<td>31.52</td>
<td>155.96</td>
<td>233.95</td>
<td>1.43</td>
</tr>
<tr>
<td>24</td>
<td>38.89</td>
<td>5.5</td>
<td>460</td>
<td>72.73</td>
<td>17.86</td>
<td>1392.08</td>
<td>2088.12</td>
<td>1.35</td>
</tr>
<tr>
<td>25</td>
<td>11.19</td>
<td>11.5</td>
<td>450</td>
<td>77.88</td>
<td>25.2</td>
<td>663.09</td>
<td>994.63</td>
<td>1.61</td>
</tr>
<tr>
<td>26</td>
<td>75.62</td>
<td>5.2</td>
<td>470</td>
<td>75.5</td>
<td>8.66</td>
<td>1464.11</td>
<td>2196.17</td>
<td>4.78</td>
</tr>
<tr>
<td>27</td>
<td>36.03</td>
<td>10.6</td>
<td>470</td>
<td>83.15</td>
<td>22.39</td>
<td>720.07</td>
<td>1080.11</td>
<td>5.88</td>
</tr>
</tbody>
</table>

1 Data source: Montgomery (1991)
In the output shown for feature 4, each number in a tranche set represents an observation identifier (row of data). The data in brackets represent the lower limit, center, and upper limit of feature 4 in the corresponding tranche. The tranche \(\{18, 20, 21, 24, 26\}\) represents the tranche with the highest cardinality for feature 4. This tranche (consisting of observations), along with the highest cardinality tranches from the remaining features, is the one we are most interested in extracting for future analysis (i.e., surrogate data and feature selection). The reason we are interested in tranches with high cardinalities (higher data density) is that we infer that a tranche with relatively many observations has potentially more valuable information to extract. We do not create tranches for each feature by simply splitting data using equal intervals, because this splitting method assumes that the range of values for each feature is uniformly distributed; in many instances this is not the case. Given the results of the tranche-generating process, we now present ways we can harness tranches for valuable information.

### 4.5 Signatures of Variables and Vectors

As mentioned above, a key means for exploiting the tranches for each data feature occurs by taking account of their cardinalities. We denote the tranches generated for a given variable \(x_j\) by \(M_{jp}\) for \(p = 1, \ldots, p_j^*\). Let \(L_{jp} = L_j(M_{jp})\) and \(U_{jp} = U_j(M_{jp})\) for any tranche \(M_{jp}\), and for convenience, assume the tranches for \(x_j\) are ordered so that \(L_{j1} < U_{j1} < L_{j2} < L_{jp^*} < U_{jp^*}\). Hence, \(L_{j1} = 0\) and \(U_{jp^*} = 1\). We denote the cardinality of a tranche as \(c_{jp} = |M_{jp}|\) for \(p = 1, \ldots, p_j^*\). Consequently, \(c_{jp}\) identifies the number of observations that have a value \(x_j(i)\) within the tranche interval.

We also define the variable-signature for the variable \(x_j\) to be the vector given by

\[
v_j = \left( c_{j1}, \ldots, c_{jp_j^*} \right).
\]

Note that \(v_j\) defines an empirical distribution of the values of \(x_j(i)\) over \(i \in M\), determined relative to the intervals defined by the tranches. The tranches that contain a larger number of vectors \(x(i)\) (i.e., for which larger numbers of \(x(i)\) have values \(x_j(i)\) that fall within the intervals defined by the
tranches) will be those that qualify as more important for the purpose of data mining operations such as classification and clustering. In essence, this signature provides us cardinalities by column.

Similarly, we define a signature for each vector \( x(i) \), \( i \in M \), as follows. This signature provides us cardinalities based on a particular observation, or “row.” For each component \( x_j(i) \) of the vector \( x(i) \), define \( c_j(i) = c_{jp} \) where \( p \) is the index for which \( i \in M_{jp} \), and \( c_{jp} \) represents the cardinality we defined at the beginning of this section. The value \( c_j(i) \) thus gives a measure of how common the value of \( x_j(i) \) is in the entire data set. A larger value of \( c_j(i) \) indicates that \( x_j(i) \) is one of the more common \( x_j \) values. In particular, the value \( c_j(i)/m \) identifies an empirical “tranche probability” for the assignment \( x_j = x_j(i) \), where all values \( x_j(i) \) that lie in the interval for the same tranche \( M_{jp} \) are associated with the same probability, which is the probability applicable to the tranche interval itself. So, we define the vector-signature for the vector \( x(i) \) by

\[
\omega(i) = (c_1(i), c_2(i), \ldots, c_n(i)).
\] (4.2)

Associated with this vector-signature we define an overall cardinality evaluation given by

\[
\delta(i) = \sum_{j \in N} c_j(i). \tag{4.3}
\]

The quantity \( \delta(i) \) gives us a measure of the representativeness of \( x(i) \). The greater the value of \( \delta(i) \), the more representative (or more common) \( x(i) \) is in relation to other vectors in \( X \). One key assumption here is that if one attribute has a higher dispersion than the other attributes for a particular observation, the resulting \( \delta(i) \) will not be skewed.

The question is how to identify surrogate vectors, either generated from the \( x_j(i) \) values in selected tranches or from a subset of the vectors \( x(i) \) in \( X \). We examine a way to extract surrogate data sets (rows of data) in the next section.

### 4.6 Selecting a Surrogate Set \( X^* \) to Replace \( X \)

We now address the goal of choosing a subset \( X^* \) of \( X \), so that \( X \) may be replaced by \( X^* \) as a basis for carrying out various data mining operations by utilizing tranches generated for each
We can select a target number \( m^* \) for the number of observations in \( X^* \) that may be significantly smaller than \( m \) if \( X \) is large.

One possible way to construct a surrogate set of size \( m^* \) is by using the vector-signature (Equation 4.2) metric. For example, to select a subset with \( m^* \) observations, one simply selects the \( m^* \) vectors with the highest \( \delta \) values (row counts). An alternative approach, however, consists of using the variable-signature metric (Equation 4.1) to construct the surrogate set.

To apply this metric, we start with a selected value \( p_0 \). Then, we select those vectors \( x(i) \) for which each of its attribute values \( x_j(i) \) lie in one of the \( p_0 \) largest tranches for \( x_j \). We focus on these largest tranches since they contain values \( x_j(i) \) for \( x_j \) that are more representative (with respect to being shared by larger numbers of vectors \( x(i) \)) of the entire data set.

Let us further explain the index set \( M^* \), such that \( X^* = \{ x(i), i \in M^* \} \). The surrogate data extraction process operates by generating \( X^* \) in relation to \( p_0 \) (as subsequently described), and then, if \( X^* \) does not contain approximately \( m^* \) elements, we revise \( p_0 \) (upward if \( |X^*| < m^* \) and downward if \( |X^*| > m^* \)) and repeat the process, until we obtain a surrogate set \( X^* \) of the desired size.

In particular, for each variable \( x_j \) we identify the largest cutoff value \( c_{jo} \) such that the set \( P_j = \{ p : c_{jp} \geq c_{op} \} \) includes the \( p_0 \) largest cardinality tranches for \( x_j \) (or all \( p^*_j \) tranches if \( p^*_j \leq p_0 \)). Then, we define

\[
M^* = \{ i \in M : \forall j \in N, \exists p \in P_j \text{ s.t. } L_{jp} \leq x_j(i) \leq U_{jp} \}.
\] (4.4)

Suppose, for example, if we set \( p_0 = 2 \) in the Belle Ayre Mine data set. Considering only feature 4, we would select observations in the following tranches.

**Feature Number: 4**

\{8, 20, 21, 24, 26\}

\{1, 2, 17, 22\}

\{4, 12, 15, 19\}
The requirement for membership in $X^*$ may be restrictive if the set $X$ consists of vectors $x(i)$ whose components $x_j(i)$ vary widely in relation to other vectors, or, on the other hand, may be somewhat loose if most of the vectors $x(i)$ contain components $x_j(i)$ that fall within similar tranches for each $j$. We may increase $p_o$ if $|X^*| < m^*$ and decrease $p_o$ if $|X^*| > m^*$, then repeat the process until we obtain a proxy set $X^*$ containing a desired number of elements. The value $m^*$ may be determined by selecting a desired percentage of the original data set, such as 10%, 20%, or 50%.

The process of revising $p_o$ can be significantly accelerated by noting that after generating an $X^*$ that is too large based on a target $m^*$, this set will still generally be substantially smaller than $X$, and we may restrict attention to this $X^*$ in place of $X$ for the next pass. Similarly, if $X^*$ is too small, we may retain all of these elements as members of the final $X^*$ to be generated, and only seek an additional number of elements sufficient to bring the total into the vicinity of $m^*$.

In Section 4.8 we discuss the loss of information (data) that arises by decreasing $m^*$, and evaluate the impact of varying the number of bins we use to extract the surrogate data. Intuitively, if we iteratively increase the number of bins we use to determine a surrogate data set, the closer our statistical and/or forecasting results match the original data set’s results.

At times, it may be necessary to not only extract a subset of data but also to select a subset of features. In the next section we demonstrate how we can exploit our generated tranches using a simple feature selection strategy.

4.7 Exploiting the Tranches for Feature Selection

We now show how to take advantage of the preceding ideas to identify a subset of variables (features) that can be used in place of the full set of attributes $N$. The basic goal is to isolate a reduced collection of variables, $N_o \leq N$, and to draw inferences about the complete system from this reduced collection. The purpose of this variable reduction is to reduce the overall complexity of analyzing a data set with numerous features. At times, reducing redundant features may also improve the overall accuracy in predictive analysis. We first introduce a simple feature selection
strategy that uses the variable signatures previously defined.

4.7.1 Simple Feature Selection Strategies

We begin by looking at feature selection without explicitly accounting for interdependencies between variables. From a first-order consideration perspective we treat a variable $x_j$ as more or less influential — i.e., having more or less potential information content—according to the degree that its signature $v_j = (c_{j1}, \ldots, c_{jp_j})$ approximates or departs from the signature of a uniform distribution. Thus, we can create a measure of a variable’s influence (information content) as follows. Let $f(v)$ be a function that increases at an increasing rate as its argument $v$ becomes more positive or negative, e.g., $f(v) = v^2$ or $|v|^3$. Then, for any attribute $x_j$ we define the influence $\gamma_j$ as

$$\gamma_j = \frac{1}{p_j^*} \sum_{p=1}^{p_j^*} f(c_{jp} - m_{pj}). \tag{4.5}$$

Note that $\gamma_j = 0$ for a completely uniform distribution of $x_j$ values, and $\gamma_j$ becomes increasingly positive as the distribution departs from a uniform distribution. We order the variables according to their influence by defining $j_\ell$ ($\ell = 1, \ldots, N$) such that $\gamma_{j_\ell} \geq \gamma_{j_{\ell'}}$ if $\ell \leq \ell'$.

To determine the reduced index set $N_o$, we select a limiting value $n^* < n$ on the number of variables in the set and a value $\varepsilon > 0$ representing a minimum influence. Then construct $N_o$ as

$$N_o = \{j_\ell : \ell \leq n_o \land \gamma_{j_\ell} \geq \varepsilon\}. \tag{4.6}$$

While intuitive, this approach does not account for interactions among features.

4.7.2 Accounting for Interdependencies

To account for such interactions, we introduce co-diversity measures that may be viewed as a counterpart to statistical correlation measures. Our co-diversity measure involves two different components. Relative to a pair of variables $x_k$ and $x_j$ ($k < j$), we define a forward co-diversity
measure $c_{kj}$ and a reverse co-diversity measure $d_{kj}$ relative to an exponent $q$ as follows

$$c_{kj} = \sum_{i \in M} |x_k(i) - x_j(i)|^q, \quad \text{and}$$

$$d_{kj} = \sum_{i \in M} |1 - x_k(i) - x_j(i)|^q.$$  \hspace{1cm} (4.7) \hspace{1cm} (4.8)

Although the exponent $q$ can be set to any positive integer, we use a value between 1 and 3. To see the motivation for these definitions, note that $c_{kj} = 0$ if $x_j$ and $x_k$ share the same values for each $i \in M$, while $d_{kj} = 0$ if they share opposite values, where $x_k(i) = 1 - x_j(i)$ for each $i \in M$ (recall our assumption that the data is normalized so that $x_j(i)$ ranges from 0 to 1 for each $i \in M$). On the other hand, the larger the difference between $x_k(i)$ and $x_j(i)$ (alternately, $x_k(i)$ and $1 - x_j(i)$), the larger the value of $c_{kj}$ (alternately, $d_{kj}$) will be. The maximum possible value of $c_{kj}$ and $d_{kj}$ is evidently $m$ for $q = 1$. Since $|x_k(i) - x_j(i)|$ must lie between 0 and 1, the value $c_{kj}$ will decrease as $q$ increases. However, the values of $c_{kj}$ will differ more greatly from each other in a relative sense the larger that $q$ becomes.

To use these measures for feature selection, we combine $c_{kj}$ and $d_{kj}$ into an overall co-diversity measure $e_{kj}$ by defining

$$e_{kj} = \min(c_{kj}, d_{kj}).$$  \hspace{1cm} (4.9)

An alternative could be to define $e_{kj}$ as the product of the two component measures rather than the minimum of these measures:

$$e_{kj} = c_{kj} d_{kj}.$$  \hspace{1cm} (4.10)

We refer to $x_k$ and $x_j$ as being more or less related than another pair $x_r$ and $x_s$ according to whether $e_{kj}$ is larger or smaller than $e_{rs}$. This notion is analogous to statistical correlation measures although its precise definition is clearly different.

### 4.7.3 Exploiting Co-Diversity for Feature Selection

We exploit co-diversity by seeking a set of variables that are maximally pairwise diverse (or minimally pairwise related), using a quadratic programming (QP) approach. Specifically, let $y_j$
denote a binary indicator variable that takes the value $y_j = 1$ if we select $x_j$ to be one of the variables in the desired feature set (i.e., select $j$ to be a member of the index set $N_o$), and $y_j = 0$ otherwise. Then we choose non-negative threshold values $T(0)$ and $T(1)$ and select a subset of $n_o$ variables having maximum pairwise diversity using the following binary quadratic optimization problem:

\[(QP1) \text{Maximize } \sum_{k,j \in N: k < j} e_{kj}y_ky_j, \quad (4.11)\]

s.t.

\[
\sum_{j \in N} y_j \leq n_o, \\
y_ky_j = 0 \ \forall k, j \in N : k < j \text{ and } e_{kj} < T(0), \\
y_ky_j = 1 \ \forall k, j \in N : k < j \text{ and } e_{kj} > T(1), \\
y_j \in \{0, 1\}, \ j \in N.
\]

Formulation $(QP1)$ aims to maximize the diversity between features, given parameter values $T(0)$ and $T(1)$ that express minimum and maximum on the co-diversity measures. We explain how we set these parameters later in this section. To understand this formulation, we note that the first constraint ensures that we identify no more features than the number we desire, $n_o$. For the second and third constraints, appropriate values for $T(0)$ and $T(1)$ are determined by means of a diversity-screening method that also provide an initial solution. This method also makes it possible to determine the value $n_o$ adaptively, as a basis for incorporating this value into formulation $(QP1)$. Let us discuss the concept of diversity screening next.

The diversity-screening method constructs an initial solution $y'$ by starting from a null solution and gradually adding components, so that $y'$ is a partial vector defined by components $y'_j$, $j \in N'$, where $N'$ starts empty. $N'$ is partitioned into the subsets $N'(0) = \{ j \in N' : y'_j = 0 \}$ and $N'(1) = \{ j \in N' : y'_j = 1 \}$. Let $N'' = N \setminus N'$ denote the set of “unassigned” indexes. For each unassigned variable $y_j$, $j \in N''$, we create evaluations $\text{Eval}_j(0)$ and $\text{Eval}_j(1)$ by setting $y'_j = 0$ and
1, respectively. Essentially, we want to maximize the contribution of the $e_{kj}$ values to the objective of formulation $(QP1)$ when setting $y'_j = 1$ and to minimize the amount of this contribution that will be lost when setting $y'_j = 0$.

These effects on the objective function value can be analyzed as follows, for an arbitrary unassigned variable $y_j$, $j \in N''$. We use the notation $e_{(kj)}$ to denote $e_{kj}$ if $k < j$ and $e_{jk}$ if $j < k$ to ensure all values $e_{(kj)}$ associated with variable $y_j$ are incorporated in our evaluation measures. Let us explain the three possible cases when we try to determine the interactions among features.

**Case 1.** $k \in N'(0)$ ($y'_k = 0$): The contribution of $e_{(kj)}$ by means of the term $e_{(kj)} y_k y_j$ in the objective function has already been eliminated (given that $y'_k = 0$) and hence setting $y'_j = 0$ or setting $y'_j = 1$ will have no effect. Consequently, variables $y_k$ for $k \in N'(0)$ are disregarded in determining Eval$_j$(0) and Eval$_j$(1).

**Case 2.** $k \in N'(1)$ ($y'_k = 1$): The term $e_{(kj)} y_k y_j$ in the objective function reduces to $e_{(kj)} y_j$; hence setting $y'_j = 1$ gains the contribution $e_{(kj)}$ to the objective and setting $y'_j = 0$ loses this contribution.

**Case 3.** $k \in N''$ ($y_k$ is unassigned): Setting $y'_j = 0$ results in $e_{(kj)} y_k y'_j = 0$ regardless of the value of $y_k$ and hence loses the contribution of $e_{(kj)}$ to the objective. Setting $y'_j = 1$ causes $e_{(kj)} y_k y'_j$ to become $e_{(kj)} y_k$, thus creating Case 2 for the next iteration when $y_k$ takes the role of $y_j$. This effect can be less than fully gaining or losing the contribution $e_{(kj)}$, and so we measure it by a value $u e_{(kj)}$ where $u < 1$.

Combining these three cases we obtain the following measures Eval$_j$(0) and Eval$_j$(1) for $j \in N''$:

$$
\text{Eval}_j(0) = \sum_{k \in N'(1)} e_{(kj)} + \sum_{k \in N''} e_{(kj)}, \quad (4.12)$$

$$
\text{Eval}_j(1) = \sum_{k \in N'(1)} e_{(kj)} + u \sum_{k \in N''} e_{(kj)}, \quad (4.13)
$$

Based on this, the best unassigned variables to set to 0 and 1, respectively, will be $y_j(0)$ and
Determining whether to set \( y_j(0) = 0 \) or to set \( y_j(1) = 1 \) can be done heuristically by reference to the following considerations, involving the number of variables currently assigned values of 1 and 0. Once \(|N'(1)| = n_o\), all unassigned variables \( y_j \) \((j \in N'')\), will automatically be assigned \( y_j' = 0 \). Hence, as \(|N'(1)|\) approaches \( n_o\), setting \( y_j(1) = 1 \) is a more strongly determining choice than setting \( y_j(0) = 0 \), and is therefore preferable. Similarly, once \(|N'(0)| = n - n_o\), then all \( y_j, j \in N''\), will automatically be assigned \( y_j' = 1 \), and consequently setting \( y_j(0) = 0 \) is the more strongly determining choice. This might argue for always choosing the assignment \( y_j(1) = 1 \) if \(|N'(0)| < n - n_o\) and always choosing \( y_j(0) = 0 \) otherwise. However, there may be merit in choosing a less strongly determining assignment for some period, for example until attaining a balance where \(|N'(1)| = |N'(0)|\), and then alternating the choices \( y_j(0) = 0 \) and \( y_j(1) = 1 \) thereafter. Empirical tests incorporating the option of always setting \( y_j(0) = 0 \) or always setting \( y_j(1) = 1 \) versus that of alternating the choices according to the preceding considerations can determine which of the approaches is best. Based on these observations, the *Diversity Screening Method* is described in Algorithm 5.

**Algorithm 5 Diversity Screening Method**

1. Let \( N' = N'(0) = N'(1) = \emptyset \) and \( N'' = N \).
2. Compute \( \text{Eval}_j(0) \) and \( \text{Eval}_j(1) \) for \( j \in N'' \).
3. Choose to set \( y_{j(0)}' = 0 \) or \( y_{j(1)}' = 1 \) for
   \[ j(0) = \arg \min_{j \in N''} \text{Eval}_j(0) \] or
   \[ j(1) = \arg \max_{j \in N''} \text{Eval}_j(1). \]
4. Add \( j^* = j(0) \) to \( N'(0) \) or \( j^* = j(1) \) to \( N'(1) \), as appropriate.
   Update \( N' = N' \cup \{j^*\} \) and \( N'' = N'' \setminus \{j^*\} \).
5. If \( |N'(1)| = n_o \) or \( |N'(0)| = n - n_o \)
   set \( y_j' = 0 \) or \( y_j' = 1 \), respectively, for all remaining \( j \in N'' \) and stop.
   Else, return to Step 1.

As can be seen, the *Diversity Screening Method* can be used to generate solutions for multiple
values of \( n_o \) in a single pass simply by recording the solution obtained at Step 4 for a current \( n_o \) and then continuing to Step 1 (with \( n_o = n_o + 1 \)) instead of stopping. This provides the ability to determine \( n_o \) adaptively, based on considerations such as the current magnitude of \( \text{Eval}_j(1) \). However, in our empirical study, we set \( n_o \) to a fixed number, i.e., the number of desired subset of features. We now explain how we determined parameter values for \( T(0) \) and \( T(1) \).

The solution \( y' \) generated from the Diversity Screening Method gives a basis for obtaining threshold values \( T(0) \) and \( T(1) \). Specifically, we determine \( T^*(0) \) and \( T^*(1) \) as follows,

\[
T^*(0) = \min_{k,j \in \mathbb{N}, k < j \land y'_k y'_j = 1} e_{kj},
\]

\[
T^*(1) = \max_{k,j \in \mathbb{N}, k < j \land y'_k y'_j = 0} e_{kj}.
\]

Then it follows that \( y_k' y_j' = 0 \) for all \( e_{kj} < T^*(0) \) and \( y_k' y_j' = 1 \) for all \( e_{kj} > T^*(1) \). It is important to mention that \( T^*(0) \) and \( T^*(1) \) are safe choices for \( T(0) \) and \( T(1) \) relative to the solution \( y' \), and we can allow increased flexibility by selection \( T(0) \) and \( T(1) \) heuristically to be any values satisfying \( T(0) \leq T^*(0) \) and \( T(1) \geq T^*(1) \).

**4.7.4 Application to Forecasting**

In formulation \( (QP1) \), we only consider the interaction among features in a data set, and disregard the response variable when determining the most relevant features in a data set. We now introduce a quadratic programming method that incorporates the response variable during feature selection for the purpose of forecasting.

The forecasting problem involves predicting the values of a vector \( x_o \) whose empirical values we denote by \( x_o(i), i \in M \), associated with the observations \( x(i), i \in M \). In this case, we want to select variables \( x_j \) for feature selection that most strongly relate to \( x_o \). Hence, we associate a binary variable \( y_o \) with \( x_o \), in order to select pairs \( y_o y_j \) such that \( e_{oj} \) is as small as possible, where \( e_{oj} \) is defined as \( e_{oj} = \min (c_{oj}, d_{oj}) \), and \( c_{oj} \) and \( d_{oj} \) are as defined in Equations 4.7 and 4.8.

Since \( x_o \) is always among the variables included in the feature set, we have \( y_o = 1 \). Hence,
instead of being concerned with pairs $y_o y_j$ and associated diversity values $e_{oj} y_o y_j$, we need only be concerned with variables $y_j$ whose diversity values have the simpler form $e_{oj} y_j$. Thus, we obtain a natural extension of the formulation $(QP1)$ with the objective of simultaneously maximizing the diversity of pairs $x_k$ and $x_j$ and minimizing the diversity of pairs $x_o$ and $x_j$ over the selected variables. We combine these objectives, as is customarily done in multi-objective optimization, by choosing a positive scalar $\psi$ to obtain the formulation:

$\text{(QP2) Maximize } \sum_{k,j \in N: k < j} e_{kj} y_k y_j - \psi \sum_{j \in N} e_{oj} y_j,$

\begin{align*}
\text{s.t.} \\
\sum_{j \in N} y_j & \leq n_o, \\
y_k y_j & = 0 \ \forall \ k, j \in N: k < j \ \text{and } e_{kj} - \psi (e_{ok} + e_{oj}) < T(0), \\
y_k y_j & = 1 \ \forall \ k, j \in N: k < j \ \text{and } e_{kj} - \psi (e_{ok} + e_{oj}) > T(1), \\
y_j & = 0 \ \forall \ j : e_{oj} > T_o(0), \\
y_j & = 1 \ \forall \ j : e_{oj} < T_o(1), \\
y_j & \in \{0, 1\}; j \in N.
\end{align*}

(Eksioglu et al. 2005) proposes the idea of a multi-objective type of binary quadratic formulation for feature selection (for the purpose of forecasting) using traditional statistics measures. As previously mentioned, our $(QP2)$ model does not assume the data follows a specific probability distribution nor does it rely on statistical measures. Let us first discuss the constraints in formulation $(QP2)$. As in formulation $(QP1)$, the first constraint does not change, given that we are targeting a given number of features $n_o$. In the second constraint, we eliminate pairs of $x_j$ and $x_k$ variables with low diversity. In contrast, the third constraint includes pairs of variables with a high level of diversity. The term $\psi(e_{ok} + e_{oj})$ attempts to capture the diversity contribution between the response variable, $x_o$, and the individual variables in the pair.
The parameter $\psi$ in the objective function represents a positive value denoting the amount of weight we place on the diversity between features $x_j$ and $x_o$. Obviously, the higher the $\psi$ value, the higher the emphasis placed on the diversity with the response variable.

Similar to the diversity screening method discussed for formulation (QP1), we include additional thresholds $T_o(0)$ and $T_o(1)$ for the purpose of eliminating those variables $y_j$ from consideration that have an associated diversity measure $e_{oj} > T_o(0)$, or of enforcing their selection if $e_{oj} < T_o(1)$. The assignments $y_j = 0$ and $y_j = 1$ of course permit these variables simply to be removed from the problem formulation.

Again, referring to the Diversity Screening Method previously described, the following analogous method can be applied to the forecasting problem. The same definitions of $\text{Eval}_j(0)$ and $\text{Eval}_j(1)$ continue to apply, by including the index $k = 0$ associated with $y_o$. Only Case 2 from among the three cases previously described is relevant for $k = 0$, since $y_o'$ is fixed at the value 1, and hence the case for $k \in N'(1)$ is the one that applies to the index $k = 0$. For clarity, however, we will retain the index $k = 0$ outside of $N'(1)$.

Noting that the contribution of $e_{oj}$ to the objective is in fact $-\psi e_{oj}$, we therefore modify the preceding definitions of $\text{Eval}_j(0)$ and $\text{Eval}_j(1)$ by rewriting $\sum_{k \in N'(1)} e(k_j)$ in the form $\sum_{k \in N'(1)} e(k_j) - \psi e_{oj}$, thus giving:

\[
\text{Eval}_j(0) = \sum_{k \in N'(1)} e(k_j) - \psi e_{oj} + \sum_{k \in N''} e(k_j), \quad (4.20)
\]

\[
\text{Eval}_j(1) = \sum_{k \in N'(1)} e(k_j) - \psi e_{oj} + u \sum_{k \in N''} e(k_j), \quad (4.21)
\]

As before, the “best” unassigned variables to set to 0 and 1, respectively, will be $y_{j(0)}$ and $y_{j(1)}$, where

\[
j(0) = \arg\min_{j \in N''} (\text{Eval}_j(0)), \quad (4.22)
\]

\[
j(1) = \arg\max_{j \in N''} (\text{Eval}_j(1)). \quad (4.23)
\]

Aside from changes, the Diversity Screening Method for forecasting is identical to the diversity method as previously defined (see Algorithm 5).
Parameter values for $T_o(0)$ and $T_o(1)$ can be determined by the same analysis previously described to determine the values $T(0)$ and $T(1)$, including the contribution to the objective when $y_o$ is involved; we replace $e_{oj}$ by $-\psi e_{oj}$. Specifically, we have the following outcomes to determine the parameter values and follow the same design as we did in determining $T(0)$ and $T(1)$ in Section 4.7.3. We define them as

\[ T^*_o(0) = \min_{j \in N: y_j' = 1} -\psi e_{oj}, \quad (4.26) \]
\[ T^*_o(0) = \max_{j \in N: y_j' = 1} e_{oj}, \quad (4.28) \]
\[ T^*_o(1) = \min_{j \in N: y_j' = 0} e_{oj}, \quad (4.29) \]
\[ T^*_o(1) = \max_{j \in N: y_j' = 0} e_{oj}. \quad (4.29) \]

We again select values $T(0)$ and $T(1)$ such that $T(0) \geq T^*_o(0)$ and $T(1) \leq T^*_o(1)$. It also follows that $y_j' = 0$ for all $-\psi e_{oj} < T^*_o(0)$ and $y_j' = 1$ for all $-\psi e_{oj} > T^*_o(1)$. Equivalently, we remove the value $\psi$ by instead defining

\[ T^*_o(0) = \max_{j \in N: y_j' = 1} e_{oj}, \quad (4.28) \]
\[ T^*_o(1) = \min_{j \in N: y_j' = 0} e_{oj}. \quad (4.29) \]

Then we have $y_j' = 0$ for all $e_{oj} > T^*_o(0)$ and $y_j' = 1$ for all $e_{oj} < T^*_o(1)$. In numerical experiments, one may allow flexibility to parameter settings by selecting $T_o(0)$ and $T_o(1)$ to be any values satisfying $T_o(0) \geq T^*_o(0)$ and $T_o(1) \leq T^*_o(1)$.

**4.8 Numerical Experiments**

To evaluate our models, we perform several experiments. The overall objective of these experiments is to assess whether these methods outperform random sampling techniques, and a traditional statistically-based feature selection approach. We utilize a U.S. Army proprietary data set with monthly historical enlisted personnel records for fiscal years 2007 to 2010, which has 280,049 observations.
In these numerical experiments, the purpose of the forecasting model is to determine a soldier’s propensity to remain in the Army. In this data set, we limit ourselves to six features and one binary response variable. The features we use are Experience (years), Age (years), Pay Grade (1 through 9), Armed Forces Qualification Test (1 to 100), Time in Grade (months), and Service Term (1 to 6 years). These features are designated as $x_1$ to $x_6$, respectively. First, we discuss features extraction based on the signature definitions we previously discussed. As a first step, we provide the minimum, maximum, and mean values of each of these features in Table 4.2.

Table 4.2: Dataset Description

<table>
<thead>
<tr>
<th>Column/Feature</th>
<th>Name</th>
<th>Minimum</th>
<th>Mean</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Stay</td>
<td>0</td>
<td>0.7692</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>Experience</td>
<td>0.083</td>
<td>5.7538</td>
<td>44.25</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>0</td>
<td>26.7215</td>
<td>64</td>
</tr>
<tr>
<td>3</td>
<td>Pay Grade</td>
<td>1</td>
<td>4.6461</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>AFQT Test Score</td>
<td>0</td>
<td>57.6579</td>
<td>99</td>
</tr>
<tr>
<td>5</td>
<td>Time In Grade</td>
<td>1</td>
<td>22.0257</td>
<td>486</td>
</tr>
<tr>
<td>6</td>
<td>Service Term</td>
<td>1</td>
<td>3.7431</td>
<td>9</td>
</tr>
</tbody>
</table>

We use split-validation to train and test a forecasting model’s predictive accuracy. To train our models, we randomly designate 80% of the data as a training set (224,039 records), and the remaining 20% (56,010 records) a test set. In this data, 76.95% and 76.81% of the training and test set personnel are reenlistees (remain in the Army), respectively. On account of the large size of our training set, we will not provide examples of the sample tranches that we generate using our algorithm. Using a tranche limit of 8 as a parameter setting, a sample of the variable-signatures allows us to make some interesting observations. A sample of the variable-signatures is as follows:

$\gamma_1 = (34780, 34139, 32991, 29658, 28460, 24281, 24225, 15507)$

$\gamma_2 = (42273, 36438, 30196, 28650, 27726, 21618, 21276, 15864)$

$\gamma_3 = (88040, 69098, 39925, 18085, 6439, 2454, 0, 0)$

$\gamma_4 = (36807, 34330, 29212, 28726, 27211, 26692, 20875, 20188)$

$\gamma_5 = (38614, 36051, 34615, 30032, 29098, 23884, 20702, 11045)$
\[ \gamma_6 = (99298, 85929, 18066, 16428, 3149, 1171, 0, 0) \]

The results above indicate the cardinalities for the eight tranches of each feature; the highest cardinality tranche is placed first in the order. By glancing at the resulting \( \gamma \) values, one can see that Pay Grade (\( \gamma_3 \)) and Service Term (\( \gamma_6 \)) have a significantly higher cardinality in its largest tranche (first value in each signature) compared to the other tranches.

For the other signature metric, vector-signature (\( \omega \)), we present the statistical results of the top five (with respect to their vector-signatures). Recall that for vector signatures, the features are normalized. The \( \omega(i) \) is in the form \((x_1, x_2, x_3, x_4, x_5, x_6, \delta(i))\), where \( x_j \) denotes the value of feature \( j \), \( x_0 \) denotes the response value, and \( \delta(i) \) represents the vector cardinality for observation \( i \). The five best vector signatures are as follows:

\[
\begin{align*}
\omega(4685) &= (0.05660, 0.34375, 0.375, 0.56566, 0.00617, 0.25, 1.0, 339812) \\
\omega(8329) &= (0.05472, 0.34375, 0.375, 0.50505, 0.00206, 0.25, 1.0, 339812) \\
\omega(8848) &= (0.06604, 0.34375, 0.375, 0.53535, 0.00823, 0.25, 0.0, 339812) \\
\omega(10199) &= (0.05660, 0.34375, 0.375, 0.50505, 0.00206, 0.25, 1.0, 339812) \\
\omega(10724) &= (0.06038, 0.35938, 0.375, 0.50505, 0.01029, 0.25, 1.0, 339812).
\end{align*}
\]

The \( \delta(i) \) values shown above for the best vector signatures are similar across observations since the values for each feature are not very well dispersed. Given a data set with each feature having more diverse continuous values, we are confident that our model will produce vector signatures with varying overall cardinalities among vector-signatures.

With respect to the influence measure, \( \gamma_j \), based on each feature’s variable signature we are able to execute a simple feature selection. The influences, based on the training set data, are as follows:

\[ \gamma = (1.16 \times 10^{10}, 8.22 \times 10^9, 5.84 \times 10^8, 5.13 \times 10^8, 2.96 \times 10^8, 2.34 \times 10^8) \]

The influence values above represent the following features: \((x_1, x_2, x_4, x_5, x_3, \text{and } x_6)\). The
order of influence results suggest that if we set an influence cutoff of $5.0 \times 10^8$, we would select the variables $x_1$, $x_2$, $x_4$, and $x_5$ to conduct our retention forecasting ($n_0 = 4$).

Now, if we perform a more explicit feature selection strategy where we account for interdependencies between features, we discover a different set of features. We produce net co-diversity measures by using the formula $e_{kj} = \min(c_{kj}, d_{kj})$. In determining forward and reverse diversity, we set the $q$ parameter to 2 and make the necessary adjustments to our calculation to ensure that our co-diversity measure coincides with our previous discussion on re-scaling the co-diversity metric. Recall that if the result of $e_{kj}$ is larger than $e_{rs}$, then the pair $x_k$ and $x_j$ is more diverse (less related) than the pair $x_r$ and $x_s$. For our training set, the resulting co-diversity matrix is shown in Table 4.3.

Table 4.3: Net Co-Diversity Measure (Training Set)

<table>
<thead>
<tr>
<th></th>
<th>$x_1$</th>
<th>$x_2$</th>
<th>$x_3$</th>
<th>$x_4$</th>
<th>$x_5$</th>
<th>$x_6$</th>
<th>$x_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0.00</td>
<td>9.9993×$10^8$</td>
<td>1.2076×$10^9$</td>
<td>1.6789×$10^9$</td>
<td>3.9276×$10^8$</td>
<td>9.7260×$10^8$</td>
<td>5.8907×$10^9$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>-</td>
<td>0.00</td>
<td>3.3327×$10^7$</td>
<td>1.8470×$10^8$</td>
<td>2.1644×$10^9$</td>
<td>9.9492×$10^7$</td>
<td>3.4076×$10^9$</td>
</tr>
<tr>
<td>$x_3$</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>2.2071×$10^8$</td>
<td>2.4773×$10^9$</td>
<td>1.4268×$10^8$</td>
<td>3.5908×$10^9$</td>
</tr>
<tr>
<td>$x_4$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>3.0373×$10^9$</td>
<td>3.7016×$10^8$</td>
<td>4.4406×$10^9$</td>
</tr>
<tr>
<td>$x_5$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>2.0155×$10^9$</td>
<td>7.4499×$10^9$</td>
</tr>
<tr>
<td>$x_6$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
<td>3.7834×$10^{10}$</td>
</tr>
<tr>
<td>$x_0$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.00</td>
</tr>
</tbody>
</table>

The results from Table 4.3 allow us to gather some preliminary insights about the features before implementing our feature selection methods. For the relation between features and the response variable, one would prefer to include variables in the $(QP2)$ model that are less diverse (more related). Observing our co-diversity results, $x_2$ and $x_3$ are the most desirable variables in this preliminary analysis if we decide to include only two features. We now demonstrate through our experiments ways we can exploit co-diversity measures for feature selection.

When we employed the diversity screening method based on our co-diversity results for our $(QP1)$ model, we determined non-threshold values of $2.21\times10^8$ and $2.16\times10^9$ for $T(0)$ and $T(1)$, respectively. Recall that the goal for our $(QP1)$ model is to select a subset of $n_0$ variables having
maximum pairwise diversity. Using the $T(0)$ and $T(1)$ parameter settings and targeting between three and four features, we solve our QP models using the IBM CPLEX optimization software. The solution to the $(QP1)$ model considers the variables $x_3$, $x_4$, and $x_5$ to be the most relevant and adds $x_1$ if we select four features.

As for our formulation $(QP2)$, we first determine the values for $T(0)$, $T(1)$, $T_o(0)$, and $T_o(1)$ relying heavily on the results of our diversity screening method. The “safe” settings for these parameters are $2.07\times10^{10}$, $1.43\times10^{10}$, $1.49\times10^{10}$ and $7.57\times10^9$, respectively. In order to simplify our experiments, we designate in our diversity-screening algorithm parameter settings of $u = 0.5$ and $\psi = 2$ (weight on co-diversity measure $e_{oj}$). Based on these settings and targeting three to four features, our formulation $(QP2)$ identifies the variables $x_1$, $x_2$, $x_3$, and $x_5$ as having the most explanatory power. Table 4.5 shows the different results generated by formulations $(QP1)$ and $(QP2)$. These results differ since we add a term to formulation $(QP2)$ which emphasizes minimizing co-diversity between the features and the response variable. When comparing the selection of three features among simple feature selection, formulation $(QP1)$, and formulation $(QP2)$, all three techniques identify $x_5$, Time in Grade, as the most important feature. Also, both formulations $(QP1)$ and $(QP2)$ consider Pay Grade ($x_3$) an important feature.

Prior to conducting employee retention forecasting using our three feature selection methods, we create two surrogate data sets and two random sample data sets from the overall training set. We form these surrogate data sets based on the cardinalities of the variable-signatures in the tranche-generation process. We extract surrogate data samples that are 14.1% and 78.4% of the original data set. A summary of the data sets we created along with their associated $m^*$ values are shown in Table 4.4.

Given the data sets shown in Table 4.4, we make use of commercially available software and conduct experiments using classification and regression tree (CRT) forecasting analysis. We limit our analysis to CRT to provide a forecasting example of the capabilities of our surrogate data representations. In using the collection of Army data sets, we include all features or use only the variables selected by our simple feature selection (Simple), formulation $(QP1)$, and formula-
Table 4.4: Summary of Surrogate, Random and Entire Training Data Set

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Number of Observations ($m^*$)</th>
<th>Percentage of Entire Training Set (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surrogate14</td>
<td>31,686</td>
<td>14.1</td>
</tr>
<tr>
<td>Random14</td>
<td>31,686</td>
<td>14.1</td>
</tr>
<tr>
<td>Surrogate78</td>
<td>175,558</td>
<td>78.4</td>
</tr>
<tr>
<td>Random78</td>
<td>175,558</td>
<td>78.4</td>
</tr>
<tr>
<td>Training</td>
<td>224,041</td>
<td>100.0</td>
</tr>
</tbody>
</table>

Our scenario naming scheme identifies modeling and data combinations in the form $\text{DataSet\_FeatureSelection}$. Thus, if we only use a subset of variables in the forecast, the description at the end of the scenario name denotes the number of features we use in the predictive analysis or the feature selection method used (along with number of features selected). For example, the scenario name $\text{Surrogate14\_QP1–4}$ indicates that the forecasting scenario employs only four features selected by the ($QP_1$) model using the 14.1% surrogate data set. Table 4.5 (next page) displays a summary of each scenario’s performance using the same test set. In our analysis, we limit our feature subset selection to three and four features since our data set only contains six features.

The results from our CRT forecasting analysis show that the surrogate data set scenarios do very well and perform better at classifying reenlisting Army personnel than all but two random subsets. The bold-faced percentages in Table 4.5 under the Overall Classification Accuracy column represent the scenario that performs the best within each feature selection category. When we use all the features for forecasting, employing our surrogate data sets perform better in classifying the test set instances than did using the entire data set. Overall, the random (Monte Carlo) sample selection methods performs the worst.

In our simple feature selection approach using only four features, our surrogate data samples produces better results (accuracy greater than 79%) than all the other data sets. Also, the data sets we employ in the four-feature approach loses approximately 2% accuracy in comparison to the same data sets when all six features are used for forecasting. Quite surprising, the random samples perform better than all the other data sets when the results of formulation ($QP_1$) (feature selection method) are used, possibly due to intentionally excluding the response variable and limiting our
### Table 4.5: CRT: Comparison of Data Set Performance using Test Set

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Features Included in The Model</th>
<th>Overall Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training_All</td>
<td>( (x_1, x_2, x_3, x_4, x_5, x_6) )</td>
<td>81.2</td>
</tr>
<tr>
<td>Surrogate78_All</td>
<td>( (x_1, x_2, x_3, x_4, x_5, x_6) )</td>
<td>78.0</td>
</tr>
<tr>
<td>Random78_All</td>
<td>( (x_1, x_2, x_3, x_4, x_5, x_6) )</td>
<td>77.3</td>
</tr>
<tr>
<td>Surrogate14_All</td>
<td>( (x_1, x_2, x_3, x_4, x_5, x_6) )</td>
<td>81.6</td>
</tr>
<tr>
<td>Random14_All</td>
<td>( (x_1, x_2, x_3, x_4, x_5, x_6) )</td>
<td>78.5</td>
</tr>
<tr>
<td>Training_Simple</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>78.8</td>
</tr>
<tr>
<td>Surrogate78_Simple</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.6</td>
</tr>
<tr>
<td>Random78_Simple</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>76.4</td>
</tr>
<tr>
<td>Surrogate14_Simple</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.1</td>
</tr>
<tr>
<td>Random14_Simple</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>78.4</td>
</tr>
<tr>
<td>Training_QP1-3</td>
<td>( (x_3, x_4, x_5) )</td>
<td>77.3</td>
</tr>
<tr>
<td>Surrogate78_QP1-3</td>
<td>( (x_3, x_4, x_5) )</td>
<td>77.0</td>
</tr>
<tr>
<td>Random78_QP1-3</td>
<td>( (x_3, x_4, x_5) )</td>
<td>77.2</td>
</tr>
<tr>
<td>Surrogate14_QP1-3</td>
<td>( (x_3, x_4, x_5) )</td>
<td>76.4</td>
</tr>
<tr>
<td>Random14_QP1-3</td>
<td>( (x_3, x_4, x_5) )</td>
<td>77.2</td>
</tr>
<tr>
<td>Training_QP2-3</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>78.6</td>
</tr>
<tr>
<td>Surrogate78_QP2-3</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.7</td>
</tr>
<tr>
<td>Random78_QP2-3</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>77.9</td>
</tr>
<tr>
<td>Surrogate14_QP2-3</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.1</td>
</tr>
<tr>
<td>Random14_QP2-3</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>78.4</td>
</tr>
<tr>
<td>Training_QP1-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.1</td>
</tr>
<tr>
<td>Surrogate78_QP1-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.4</td>
</tr>
<tr>
<td>Random78_QP1-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>76.6</td>
</tr>
<tr>
<td>Surrogate14_QP1-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.1</td>
</tr>
<tr>
<td>Random14_QP1-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>78.6</td>
</tr>
<tr>
<td>Training_QP2-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.0</td>
</tr>
<tr>
<td>Surrogate78_QP2-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.8</td>
</tr>
<tr>
<td>Random78_QP2-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>78.1</td>
</tr>
<tr>
<td>Surrogate14_QP2-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>79.2</td>
</tr>
<tr>
<td>Random14_QP2-4</td>
<td>( (x_1, x_2, x_3, x_4, x_5) )</td>
<td>78.6</td>
</tr>
</tbody>
</table>
feature selection to three variables. However, forecasting accuracy using the surrogate data samples with the features selected by formulation (QP2) exceeds those we obtain using the entire data set or a random sample. Additionally, the surrogate data samples using the features generated by QP2–3 perform better than those generated by the QP1–3 model and by the simple feature selection method. Using the features produced by our QP2–4 model, our Surrogate14 and Surrogate78 data samples produce slightly better predictive accuracy than our QP1–4 models, thus demonstrating that consideration of the response variable is an absolute necessity when performing forecasting analysis.

One can infer that random selection methods do not perform better when using a subset of features determined by formulations (QP1) and (QP2) for the data set we use. For the most part, as the amount of the surrogate data increases, its overall classification accuracy also increases. A very interesting result of this empirical analysis reveals that the smaller surrogate data set, Surrogate14, outperforms the other data sets in forecasting when all the features in the data sample are included. Overall, the Surrogate data sets did well representing the larger (original) data set for the purpose of military retention forecasting.

In sum, although the differences in predictive accuracy between modeling scenarios are not significant, a classification accuracy difference of only 1% for a large data set, such as five million records, is rather substantial. Furthermore, depending on the composition of the input data set, there are other forecasting analysis tools besides CRT that may better illustrate the benefits of using our surrogate data and feature selection tools. Hence, we have shown that employing intelligent feature selection and surrogate data extraction is essential to conducting the most accurate personnel and cost forecasts for an organization. Based on the results of our numerical experiments, we advise data analysts to incorporate the (QP2) model (i.e., with response variable) in their surrogate data representations to achieve the highest predictive power.
4.9 Conclusion

Given the rapid growth of analytics as a means for extracting valuable information from large data sets, it is of the utmost importance to find ways of reducing the vast amount of data so as to make statistical inferences. We have shown that our surrogate data representations offer a highly effective way achieving this goal. Not only does it yield a proxy that can usefully serve in place of the larger data source, but it is extremely effective when compared to standard stochastic data sampling methods. The deterministic underpinnings of the approach permit it to be easily modified depending on the desired size of the surrogate data source. Our second approach, the QP feature selection model, is both intuitive and capable of being applied to numerous data (scalar) types. The QP feature selection model can also serve to identify the most relevant features of a data set.

Although quadratic programming is presented in previous research (Eksioglu et al., 2005) using the correlation statistical measure, the incorporation of a diversity-screening method in our approach strengthens the overall subset-selection process in a manner that makes it highly competitive with traditional subset-selection methods. Furthermore, our models do not require the data to follow a specific, parametric probability distribution. Also, unlike many traditional exact methods, our model does not require monotonicity of the objective function to make accurate feature selections.

Our forecasting binary quadratic programming model (QP2) demonstrates its effectiveness in determining the response variable given a randomly selected test set. The model yields results rivaling those of traditional methods and presents a new way of looking at forecasting. Our experimental outcomes show the usefulness of this model for making inferences about employee retention propensity.

In the future, the use of a tabu search (TS) metaheuristic would afford a useful enhancement of the surrogate data selection algorithm. By taking advantage of the tranches, we may go a step farther and not only select a subset $N_o$ of variables for consideration but additionally identify ranges of values for these variables using TS that suffice for conducting the analyses we intend to perform.
For this purpose, we may choose to make use of a Sequential Fan Candidate list strategy of tabu search. In the present context, this procedure is a variant of beam search enhanced by drawing on principles from TS multi-start methods.

Additionally, to emphasize the relative differences among the $e_{(kj)}$ values in the definitions of $\text{Eval}_j(0)$ and $\text{Eval}_j(1)$, we may replace these values by $\xi_{(kj)} = e^{q}_{(kj)}$, where $q$ is a selected exponent with a value that may be different than that in the definitions of $c_{kj}$ and $d_{kj}$ (though still in the range of 1 to 3) in order to more fully accentuate differences among $e_{(kj)}$ values. This consideration may provide a means of refining the Diversity Screening method to increase its effectiveness for quickly obtaining near optimal solutions to formulation ($QP1$).

The fact that our methods can be employed in mathematical modeling for pattern recognition and response-variable forecasting heightens the relevance of future research into enhancing their ability to perform feature selection and forecasting. The contribution of the present research represents one in a series of steps toward developing more complex mathematical models to aid in making statistical inferences and handling big data issues more effectively.
Bibliography


Bennett, Kristin P. 1992. *Decision tree construction via linear programming*. Center for Parallel Optimization, Computer Sciences Department, University of Wisconsin.


G1, Army Personnel. 2010a. Enlisted Active Army Military Manpower Plan (AAMMP). Headquarters, Department of the Army.

G1, Army Personnel. 2010b. Monthly Program Update Briefing (PUB). Eg0710f,eg0810e,eg0910e ed. Headquarters, Department of the Army.


Heckman, J. J. 1976. The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, vol. 5. NBER, 475–492.


Madden, S. 2012. From databases to big data. IEEE Internet Computing 16(3).


Appendix A

Further Data Cleanup

The original inventory file is reduced without altering the number of records by eliminating fields believed to be either irrelevant or unlikely to provide value our model prior to merging with transaction and bonus data. The number of fields in the column-condensed file is reduced from 255 to 36 variables as shown in Table A.1. Also, we further reduced the number of records by only considering first term Soldiers (Reenlistment Quantity = 0). Based on this filter, our inventory file was reduced to 6,600,352 records.

The next procedure involved manipulating the transaction and bonus data files as shown in Table A.2. We provide a detailed description of the data fields here to ensure thorough understanding of the stay/leave transactions. We reduced the size of the transaction data set by including only reenlistment eligible ETS losses (LETS) and reenlistments (LIMR). The transaction file decreased to 322,084 transactions.

Further reduction of the inventory file was achieved by including only individuals that appear in the transaction file (Reenlistment or Voluntary Loss); this reduced the inventory file to 2,837,040 personnel. The inventory and transaction data sets were then merged by social security number and date (month and year merge). The merged file that was created consists of a data set that contains records of inventory and transaction data for first term soldiers only. Our merged file consists of 131,288 unique first term personnel with military experience between 17 months and 6 years (first termer by definition). The bonus file (with bonus amounts) we received from HRC consists of data from October 2000 through October 2010. We condensed the data based on the limited timeline in
Table A.1: Variable Candidates to Serve as Explanatory Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>Variable Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADULT_DEP_QY</td>
<td>Interval</td>
<td>Number of adult dependents.</td>
</tr>
<tr>
<td>AFQT_PCNT_QY</td>
<td>Interval</td>
<td>Armed Forces Qualification Test score range: 0 to 99.</td>
</tr>
<tr>
<td>AFQT_GRP</td>
<td>Categorical</td>
<td>Armed Forces Qualification Test (AFQT) score grouping.</td>
</tr>
<tr>
<td>BIRTH_DT</td>
<td>Interval</td>
<td>Date of birth.</td>
</tr>
<tr>
<td>BASD_DT</td>
<td>Interval</td>
<td>Basic Active Service date. Date entered the Army on Active Duty.</td>
</tr>
<tr>
<td>BIRTH_DT</td>
<td>Interval</td>
<td>Date of birth.</td>
</tr>
<tr>
<td>BIRTH_ST_CD</td>
<td>Categorical</td>
<td>U.S. state of birth.</td>
</tr>
<tr>
<td>CMF_CD</td>
<td>Categorical</td>
<td>Army Career Management Field consists of group of MOS's.</td>
</tr>
<tr>
<td>CNTRY_BIRTH_CD</td>
<td>Categorical</td>
<td>Country of birth.</td>
</tr>
<tr>
<td>CIV_ED_QY</td>
<td>Interval</td>
<td>Number of years of civilian education (high school = 12)</td>
</tr>
<tr>
<td>CTZN_COUNTRY_CD</td>
<td>Categorical</td>
<td>Country of citizenship.</td>
</tr>
<tr>
<td>DEPL_DT</td>
<td>Interval</td>
<td>Last combat deployment date.</td>
</tr>
<tr>
<td>DOB_DT</td>
<td>Interval</td>
<td>Date of current rank.</td>
</tr>
<tr>
<td>EP_SSN</td>
<td>Categorical</td>
<td>Enlisted personnel service number.</td>
</tr>
<tr>
<td>ETS_DT</td>
<td>Interval</td>
<td>Expiration Term of Service (ETS) date.</td>
</tr>
<tr>
<td>EXPER</td>
<td>Interval</td>
<td>Experience level (years).</td>
</tr>
<tr>
<td>FISCAL_YEAR_ID</td>
<td>Interval</td>
<td>Fiscal year (FY) for current record.</td>
</tr>
<tr>
<td>GI_BILL_STATUS</td>
<td>Categorical</td>
<td>GI Bill eligibility code.</td>
</tr>
<tr>
<td>MARST_CD</td>
<td>Categorical</td>
<td>Marital Status. Married, Divorced, Separated or Single.</td>
</tr>
<tr>
<td>METS_QY</td>
<td>Interval</td>
<td>Months until ETS.</td>
</tr>
<tr>
<td>MINOR_DEP_QY</td>
<td>Interval</td>
<td>Number of minor children.</td>
</tr>
<tr>
<td>MSV_QY</td>
<td>Interval</td>
<td>Soldier experience level in months.</td>
</tr>
<tr>
<td>NUM_LONG_TOURS</td>
<td>Interval</td>
<td>Number of long overseas tours.</td>
</tr>
<tr>
<td>NUM_SHRT_TOURS</td>
<td>Interval</td>
<td>Number of short overseas tours.</td>
</tr>
<tr>
<td>PAY_GRADE_ID</td>
<td>Interval</td>
<td>Enlisted pay grade (1-9).</td>
</tr>
<tr>
<td>PMOS_CD</td>
<td>Categorical</td>
<td>Primary Military Occupation Specialty (PMOS).</td>
</tr>
<tr>
<td>RANK</td>
<td>Categorical</td>
<td>Military rank (Private through Sergeant Major).</td>
</tr>
<tr>
<td>REDCAT_CD</td>
<td>Categorical</td>
<td>Racial Code (White, Hispanic, Black, Asian, Native Amer., and Other).</td>
</tr>
<tr>
<td>REENL_QY</td>
<td>Interval</td>
<td>Number of reenlistments.</td>
</tr>
<tr>
<td>SEXCATEGORY_CD</td>
<td>Categorical</td>
<td>Gender.</td>
</tr>
<tr>
<td>SPD</td>
<td>Categorical</td>
<td>Separation Personnel Code (SPD).</td>
</tr>
<tr>
<td>STRENGTH_CAT_CD</td>
<td>Categorical</td>
<td>Strength Category: Training or Deployable Unit of assignment.</td>
</tr>
<tr>
<td>SVC_TERM_CD</td>
<td>Interval</td>
<td>Service Term Code. Current contract length in months.</td>
</tr>
<tr>
<td>THS_STATUS_CD</td>
<td>Categorical</td>
<td>Trainee, Holdee (medical or prisoner) or Student.</td>
</tr>
<tr>
<td>TIG_QY</td>
<td>Interval</td>
<td>Amount of months in current grade.</td>
</tr>
<tr>
<td>YM_DT</td>
<td>Interval</td>
<td>Year-Month date of record.</td>
</tr>
</tbody>
</table>
Table A.2: TAPDB Transaction and Bonus Fields

<table>
<thead>
<tr>
<th>Variable</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOS</td>
<td>Additional Obligation of Service (AOS). True length of new contract.</td>
</tr>
<tr>
<td>BONUS</td>
<td>Maximum retention Bonus amount received or offered during reenlistment period in $1000's.</td>
</tr>
<tr>
<td>EXT_MO</td>
<td>Calculation of extension months pertaining to Soldier affected by Stop Loss policy.</td>
</tr>
<tr>
<td>GAIN_TYPE_CD</td>
<td>Description of specific type of gain to active duty (Prior Service, Non-Prior Service, Immediate Reenlistment).</td>
</tr>
<tr>
<td>LOSS_TYPE_CD</td>
<td>Description of specific type of loss from active duty (Expiration of Term of Service, Retirement, Misconduct, Physical Disability, Dropped From Rolls, Entry Level Separation, Hardship or Parenthood, Pregnancy, Unfit, Unsatisfactory Performance, Reduction In Force, Early Release, Early Retirement, Immediate Reenlistment, Other).</td>
</tr>
<tr>
<td>MONTH</td>
<td>Month of transaction (MM).</td>
</tr>
<tr>
<td>PARTITION_ID</td>
<td>Numerical value assigned to each monthly data segment.</td>
</tr>
<tr>
<td>SSN_ID</td>
<td>Social Security Number pertaining to the Soldier who executed the transaction.</td>
</tr>
<tr>
<td>SPD_CD</td>
<td>Three character Separation Program Designator code that identify reasons for, and types of, separation from active duty.</td>
</tr>
<tr>
<td>TRANS_CAT_CD</td>
<td>Major category type of transaction (Gain, Loss, Extension, Promotion, Demotion).</td>
</tr>
<tr>
<td>TRANS_DT</td>
<td>Actual date of transaction (DDMMYYYY).</td>
</tr>
<tr>
<td>YEAR</td>
<td>Year of transaction (YYYY).</td>
</tr>
</tbody>
</table>

our merged inventory transaction file, while also only keeping records of first term reenlistments. Our bonus file was reduced from 587,864 to 94,015 individuals. Next, the maximum bonus data field was added to every record in the remaining condensed and consolidated data set by merging the base data set with the constructed bonus data file obtained from HRC.

For the leaving personnel group (ETS), a random sample (27.4% of first term reenlistments are for 4 years) was selected based on the proportion of 4-year reenlistments compared to all reenlistments. One important note is that we excluded reenlistments that occurred for soldiers held beyond their original contractual obligation (temporary stop loss Army policy). Following the last step, we performed several cleaning and validation checks to further condense our regression data file. Lastly, given the largest career field, Infantry, and the most popular reenlistment option,
Option 2 (Table A.3), we limit our retention rate analysis to a manageable 1,475 personnel that have re-enlisted in their short career. As previously mentioned, we used the bonus amount a re-enlisting soldier received as their bonus value for our empirical analysis.

Table A.3: Reenlistment Options for 4-Year Reenlistees

<table>
<thead>
<tr>
<th>Description</th>
<th>Option Number</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Needs of the Army</td>
<td>1</td>
<td>10.06</td>
</tr>
<tr>
<td>Station Stabilization</td>
<td>2</td>
<td>22.69</td>
</tr>
<tr>
<td>MOS Change (training)</td>
<td>3</td>
<td>15.63</td>
</tr>
<tr>
<td>Overseas (non-combat) Assignment</td>
<td>4</td>
<td>19.23</td>
</tr>
<tr>
<td>U.S. Station of Choice</td>
<td>5</td>
<td>32.39</td>
</tr>
</tbody>
</table>

Leaving personnel were also assigned a random Additional Obligation of Service (AOS) value based on the AOS distributions (Table A.4) provided by the 4-year reenlistment bonus file. AOS is a variable that we seek to determine its value in predicting reenlistment behavior since it measures the total “actual” contractual time a incumbent soldier adds to his/her commitment. The minimum commitment according to AR 601-280 (2011) is 24 months.
Table A.4: Additional Obligation of Service (AOS) Distribution for 4-Year Re-enlistees

<table>
<thead>
<tr>
<th>AOS</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>1763</td>
<td>8.49</td>
<td>1763</td>
<td>8.49</td>
</tr>
<tr>
<td>25</td>
<td>1265</td>
<td>6.09</td>
<td>3028</td>
<td>14.59</td>
</tr>
<tr>
<td>26</td>
<td>1032</td>
<td>4.97</td>
<td>4060</td>
<td>19.56</td>
</tr>
<tr>
<td>27</td>
<td>1002</td>
<td>4.83</td>
<td>5062</td>
<td>24.39</td>
</tr>
<tr>
<td>28</td>
<td>899</td>
<td>4.33</td>
<td>5961</td>
<td>28.72</td>
</tr>
<tr>
<td>29</td>
<td>916</td>
<td>4.41</td>
<td>6877</td>
<td>33.13</td>
</tr>
<tr>
<td>30</td>
<td>963</td>
<td>4.64</td>
<td>7840</td>
<td>37.77</td>
</tr>
<tr>
<td>31</td>
<td>1043</td>
<td>5.03</td>
<td>8883</td>
<td>42.80</td>
</tr>
<tr>
<td>32</td>
<td>1115</td>
<td>5.37</td>
<td>9998</td>
<td>48.17</td>
</tr>
<tr>
<td>33</td>
<td>1044</td>
<td>5.03</td>
<td>11042</td>
<td>53.20</td>
</tr>
<tr>
<td>34</td>
<td>973</td>
<td>4.69</td>
<td>12015</td>
<td>57.89</td>
</tr>
<tr>
<td>35</td>
<td>1176</td>
<td>5.67</td>
<td>13191</td>
<td>63.56</td>
</tr>
<tr>
<td>36</td>
<td>1094</td>
<td>5.27</td>
<td>14285</td>
<td>68.83</td>
</tr>
<tr>
<td>37</td>
<td>977</td>
<td>4.71</td>
<td>15262</td>
<td>73.53</td>
</tr>
<tr>
<td>38</td>
<td>945</td>
<td>4.55</td>
<td>16207</td>
<td>78.09</td>
</tr>
<tr>
<td>39</td>
<td>849</td>
<td>4.09</td>
<td>17056</td>
<td>82.18</td>
</tr>
<tr>
<td>40</td>
<td>799</td>
<td>3.85</td>
<td>17855</td>
<td>86.03</td>
</tr>
<tr>
<td>41</td>
<td>697</td>
<td>3.36</td>
<td>18552</td>
<td>89.39</td>
</tr>
<tr>
<td>42</td>
<td>592</td>
<td>2.85</td>
<td>19144</td>
<td>92.24</td>
</tr>
<tr>
<td>43</td>
<td>523</td>
<td>2.52</td>
<td>19667</td>
<td>94.76</td>
</tr>
<tr>
<td>44</td>
<td>367</td>
<td>1.77</td>
<td>20034</td>
<td>96.53</td>
</tr>
<tr>
<td>45</td>
<td>329</td>
<td>1.59</td>
<td>20363</td>
<td>98.11</td>
</tr>
<tr>
<td>46</td>
<td>287</td>
<td>1.38</td>
<td>20650</td>
<td>99.49</td>
</tr>
<tr>
<td>47</td>
<td>103</td>
<td>0.50</td>
<td>20753</td>
<td>99.99</td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td>0.01</td>
<td>20755</td>
<td>100.00</td>
</tr>
</tbody>
</table>
Appendix B

Additive Retention Case

Using the same ideas as Petruzzi and Dada (1999), we start by defining the retention function as,

$$\chi(b, \epsilon) = y(b) + \epsilon$$

and restricting $Q$ and $q$ to be both $\geq 0$, the cost function becomes:

$$\Pi(q, b) = c_u [Q - q - y(b) - \epsilon]^{+} + c_r q + b \min\{Q - q, y(b) + \epsilon\}$$

$$= c_u [Q - q - y(b) - \epsilon]^{+} + c_r q + b(Q - q) - b(Q - q - y(b) - \epsilon)^{+}$$

$$= (c_u - b)[Q - q - y(b) - \epsilon]^{+} + (c_r - b)q + bQ. \quad (B.1)$$

To interpret this new equation, one should see that the 'c_u - b' term as the cost of not reenlisting enough soldiers, while one can see the 'c_r - b' term as the cost of recruiting relative to the bonus level, $b$, that is being offered. Both $b^*$ and $q^*$ minimize expected cost. To make the analysis simpler, we must define $z = Q - q - y(b)$ or $q = Q - y(b) - z$. Also, let $\Gamma(z) = \int_A^z [z - u] f(u) du$, the Expected Cost function is then defined as,

$$E[\Pi(z, b)] = \int_A^z (c_u - b)[z - u] f(u) du + c_r (Q - y(b) - z) + b(z + y(b))$$

$$= (c_u - b)\Gamma(z) + (b - c_r)(y(b) + z) + c_r Q. \quad (B.2)$$

Hence, the objective of our model is to minimize expected cost:

$$\text{Min}_{z; \; b \geq 0} \quad E[\Pi(z, b)]. \quad (B.3)$$

In order to further analyze the objective function's structure, we must analyze its first and second order partial derivatives with respect to $z$ and $b$. Using Leibniz’s rule to differentiate an integral,
we arrive at the following expressions with respect to z:

\[
\frac{\partial E[\Pi(z, b)]}{\partial z} = \frac{\partial}{\partial z} \left((c_u - b)\Gamma(z) + (b - c_r)(y(b) + z) + c_r Q\right)
\]

\[
0 = (c_u - b) F(z) + c_r \beta (-1) + b
\]

\[
0 = (c_u - b) F(z) + (b - c_r)
\]

\[
z = F^{-1}\left(\frac{c_r - b}{c_u - b}\right).
\]  \(\text{(B.4)}\)

Now, we arrive at the following equations with respect to b:

\[
\frac{\partial E[\Pi(z, b)]}{\partial b} = \frac{\partial}{\partial b} \left(\int_{-\infty}^{z} (c_u - b)[z - u] f(u) \, du + (b - c_r)(y(b) + z) + c_r Q\right)
\]

\[
\frac{\partial E[\Pi(z, b)]}{\partial b} = (-1) \Gamma(z) + [(\alpha + b \beta) + z] + (b - c_r) \beta.
\]  \(\text{(B.5)}\)

**Lemma 1.** For a fixed z, the optimal bonus is determined uniquely as a function of z:

\[
b^* = b(z) = \frac{c_r \beta + \Gamma(z) - \alpha - z}{2 \beta}
\]

**Proof.** Using equation (7),

\[
0 = [(\alpha + b \beta) + z] + (b - c_r) \beta - \Gamma(z)
\]

\[
-2 \beta b = \alpha + z - c_r \beta - \Gamma(z)
\]

\[
\Rightarrow b = \frac{c_r \beta + \Gamma(z) - \alpha - z}{2 \beta}.
\]  \(\Box\)

Continuing to the second order differential equations, the resulting equations are:

\[
\frac{\partial^2 E[\Pi(z, b)]}{\partial z^2} = (c_u - b) f(z),
\]  \(\text{(B.6)}\)

\[
\frac{\partial^2 E[\Pi(z, b)]}{\partial b^2} = 2 \beta,
\]  \(\text{(B.7)}\)

\[
\frac{\partial^2 E[\Pi(z, b)]}{\partial z \partial b} = \frac{\partial^2 E[\Pi(z, b)]}{\partial b \partial z} = 1 - F(z).
\]  \(\text{(B.8)}\)
The Additive Expected Cost function is convex ∀b and z. Hence, we define and prove the following theorem.

**Theorem 1.** The single-period optimal recruiting and bonus policy for the additive demand case is to recruit \( q^* = Q - y(b^*) - z^* \) soldiers at the unit bonus level \( b^* \), where \( b^* \) is specified by Lemma 1 and \( z^* \) is determined according to the following:

(a) If \( F(\cdot) \) is an arbitrary distribution function, then an exhaustive search over all values of \( z \) in the region \([A,B]\) will determine \( z^* \).

(b) If \( F(\cdot) \) is a distribution function satisfying the condition \( 2r(z)^2 + dr(z)/dz > 0 \) for \( A <= z <= B \), where \( r(\cdot) \equiv f(\cdot)/(1 - F(\cdot)) \) is the hazard rate, then \( z^* \) is the smallest \( z \) in the region \([A,B]\) that satisfies \( dE[\Pi(z,b(z))] / dz = 0 \).

(c) If the condition for (b) is met AND \( c_u - c_r > 0 \), then \( z^* \) is the unique \( z \) in the region \([A,B]\) that satisfies \( dE[\Pi(z,b(z))] / dz = 0 \).

**Proof.** Using the chain rule and our first order results for \( z \) and \( b \) (Lemma 1):

First, let us define the following derivatives of the multiplicative bonus function \( b(z) \) and \( dE[\Pi(z,b(z))] / dz \):

\[
\frac{db(z)}{dz} = -1/(2\beta)(1 - F(z))
\]

\[
\frac{d^2b(z)}{dz^2} = 1/(2\beta)(f(z)) = 1/(2\beta)(1 - F(z))r(z) = -\frac{db(z)}{dz}r(z)
\]

\[
\frac{dE[\Pi(z,b(z))]}{dz} = (c_u - b(z))F(z) + (b(z) - c_r).
\]

Let us now define, \( R(z) \equiv \frac{dE[\Pi(z,b(z))]}{dz} \), substitute for \( db(z)/dz \), we arrive at the following equation,

\[
\frac{dR(z)}{dz} = f(z)[c_u - b(z)] + F(z)(-\frac{db(z)}{dz}) + \frac{db(z)}{dz} + \frac{db(z)}{dz} = f(z)[c_u - b(z)] - \frac{F(z)db(z)}{f(z)} + \frac{db(z)/dz}{f(z)}
\]

\[
= f(z)\left[2c_u\beta - c_r\beta - \Gamma(z) + \alpha + z + \frac{F(z)(1 - F(z))}{r(z)(1 - F(z))} - \frac{1 - F(z)}{r(z)(1 - F(z))}\right]
\]

\[
= f(z)\left(2c_u\beta - c_r\beta - \Gamma(z) + \alpha + z - \frac{1 - F(z)}{r(z)}\right).
\]

First, we take the derivative of \( \frac{dR(z)}{dz} \) using the chain rule, use the definition of \( r(\cdot) \) and we arrive
at the following second order equation,

\[
\frac{d^2 R(z)}{dz^2} = f(z) \left( \frac{dR(z)/dz}{f(z)} \right) \frac{df(z)}{dz} + f(z) \left[ \frac{1 - F(z)}{2\beta} - \frac{r(z)(-f(z))}{r(z)^2} - \frac{(1 - F(z))dr(z)/dz}{r(z)^2} \right]
\]

Finally, substituting the function \( f(z) = (1 - F(z))r(z) \) and setting \( dR(z)/dz = 0 \) to arrive at a local optimum point,

\[
\frac{d^2 R(z)}{dz^2} = f(z) \left( \frac{dR(z)/dz}{f(z)} \right) \frac{df(z)}{dz} + f(z) \left[ \frac{1 - F(z)}{2\beta} + \frac{dr(z)/dz}{r(z)^2} \right]
\]

\[\Rightarrow 2r(z)^2 + \frac{dr(z)}{dz} > 0.\]

Theorem 1 tells us that if \( F(\cdot) \) is a cumulative distribution function that satisfies the condition \( 2r(z)^2 + (dr(z)/dz) > 0 \), then we can conclude that \( R(z) \) is either a monotone or unimodal function. This also implies that \( R(z) = dE[\Pi(z, b(z))] / dz \) has at most two roots. One can easily see that \( R(A) = b(A) - c_r < 0 \). So, if \( R(z) \) has only one root, it indicates a change of sign for \( R(z) \) from negative to positive, and therefore corresponds to a local minimum of \( E[\Pi(z, b(z))] \). If it has two roots, the smaller of the two corresponds to a local minimum and the larger of the two corresponds to a local maximum of \( E[\Pi(z, b(z))] \). In either case, \( E[\Pi(z, b(z))] \) has only one local minimum, identified either as the unique value of \( z \) that satisfies \( R(z) = dE[\Pi(z, b(z))] / dz = 0 \) or as the smaller of two values of \( z \) that satisfies \( R(z) = 0 \). And since \( E[\Pi(z, b(z))] \) is unimodal if \( R(z) \) has one root (assuming \( 2r(z)^2 + dr(z)/dz > 0 \)), a sufficient condition for unimodality of \( E[\Pi(z, b(z))] \) is \( R(B) > 0 \). Likewise, \( 2\beta R(A) > 0 \), where:

\[
2\beta R(B) = (c_u - b(B))F(B) + b(B) - c_r
\]

\[= (c_u - b(B)) + b(B) - c_r
\]

\[= c_u - c_r. \]
B.0.0.1 Additive Reformulation and Numerical Examples

The results presented above allow us to simplify (3) to an optimization problem over the single variable $b$ by first solving for the optimal value $q$ as a function of $b$ and then substituting the result back into $E[\Pi(q, b)]$. This method described above was utilized by Petruzzi et al. (1999), but first introduced by Whitin (1955). By implementing this two-stage process and inserting equation (3), the expected cost objective function becomes:

$$
\min_{b \geq 0} E[\Pi(b)] = \min E \left[ \int_{-\infty}^{F^{-1}(\frac{cr-b}{cu-b})} (cu-b)[F^{-1}\left(\frac{cr-b}{cu-b}\right) - u] f(u) \, du \right]
$$

$$(B.9)$$

$$(c_r - b) \left( Q - y(b) - F^{-1}\left(\frac{cr-b}{cu-b}\right) \right) + b Q.$$

The equation above serves the following meaning. One must consider the $'cr - b'$ term as “opportunity cost.” If one is to recruit a new soldier at the cost of recruiting $'cr'$, the Army could have attained another soldier at the $b$ price instead.

In order to simplify the expected value function, we define the following variables for $w$ and $b$:

$$w = \frac{cr - b}{cu - b}, \quad w \in [0, cr/cu], \quad (B.10)$$

$$b = \frac{cr - wcu}{1 - w} = \frac{cr - cu}{1 - w} + cu, \quad b \in [0, cr], \quad (B.11)$$

Given the above definitions of $w$ and $b$, our expected cost function (equation 4) with respect to $w$ is:

$$= E\left[ \frac{cr - cu}{w - 1}(F^{-1}(w) - \epsilon)^+ \right] + [(c_r - cu) - (c_r - cu)\left( Q - y\left(\frac{cr - cu}{1 - w} + cu\right) - F^{-1}(w) \right) + \left(\frac{cr - cu}{1 - w} + cu\right)Q$$

$$= cuQ + (c_r - cu)\left\{ E\left[ \frac{1}{w - 1}(F^{-1}(w) - \epsilon)^+ \right] - \frac{w}{1 - w}[Q - y\left(\frac{cr - cu}{1 - w} + cu\right) - F^{-1}(w)] + \frac{Q}{1 - w} \right\}. \quad (B.12)$$

We conduct numerical and simulation experiments using the previously defined and equations in this section using the definition that $y(b) = \alpha + \beta b^a$ and the Normal cumulative distribution function (CDF). Also, not included in this paper, we conducted experiments using the Exponential CDF,
and confirmed that our theoretical results match simulation experiments. The results from our Normality-assumed theoretical and simulated experiments are shown in Figures B.1-B.4.

![Figure B.1: Derived E(q,b) for Additive Case: \( c_r/c_u = 0.5 \)](image1)

![Figure B.2: Simulated E(q,b) for Additive Case: \( c_r/c_u = 0.5 \)](image2)

![Figure B.3: Derived E(q,b) for Additive Case: \( c_r/c_u = 0.75 \)](image3)

![Figure B.4: Simulated E(q,b) for Additive Case: \( c_r/c_u = 0.75 \)](image4)

As is evident from Figures B.1-B.4, our theoretical results are validated by the simulated experiments. Visually, one can see that the expected cost function is indeed convex given it follows the conditions specified in Theorem 1. Furthermore, one can observe that the expected cost increases as the number of recruits and bonus levels are increased.

In addition, the expected cost function is convex when the parameters follow \( b \leq c_r \leq c_u \) bounds. As the \( c_r/c_u \) ratio increases from 0.25 to 0.99 the expected cost increases rapidly as \( c_r \)
approaches $c_u$. This result tells us that as the cost of retention approaches the underage cost, it becomes difficult to determine the most cost effective option, retention or recruiting. For this reason, we establish the bounds for recruiting cost, underage cost, and bonus amounts in order to produce realistic results in balancing a particular workforce (equation B.1).

Continuing in our analysis of the Normal CDF expected cost function behavior in terms of the variable $w$, one can see from Figures B.5-B.6 that indeed the function is convex and one can determine the optimal bonus value $b$ given the minimum value of $w$. Also, holding the $\alpha$ value constant, as the additive slope increases, the optimum value of $w$ increases. One reason for this increase is because of the large incremental increase between recruiting and underage cost. Similarly, as the cost of recruiting approaches the underage cost, it is most beneficial to offer a retention bonus to compete against recruiting. Almost identical numerical results were validated assuming a Exponential CDF.

Figure B.5: $E(w)$ for Additive Retention Case: $c_r/c_u = 0.25, \alpha = 30$

Figure B.6: $E(w)$ for Additive Retention Case: $c_r/c_u = 0.75, \alpha = 30$
Appendix C

Multiplicative Retention Case

Although the additive retention case provides us with some valuable insights into the behavior of retention as we vary the ratio of recruiting costs and underage costs, a more realistic representation of this behavior is better illustrated with a multiplicative retention representation.

Before presenting the model, we first present our modeling notation to address the model’s stochastic retention behavior and multiplicative case parameters. As explained in Chapter 2, \( \epsilon \) denotes a Normal random variable. Let \( \chi(b, \epsilon) = y(b)\epsilon \), where \( y(b) = ab^\beta \), and \( z \equiv \frac{Q-y}{y(b)} \). We use the transformation to the variable \( z \) to make the derivation of the expected cost computational tractable.

It follows from Equation A.1 that the single period cost function becomes:

\[
\Pi(z, b) = c_u[z y(b) - y(b)\epsilon^+] + c_r (Q - zy(b)) + b \min \{ z y(b), y(b)\epsilon \}
\]

\[
= c_u[z y(b) - y(b)\epsilon^+] + c_r (Q - zy(b)) + b(zy(b)) - b(zy(b) - y(b)\epsilon)^+
\]

\[
= (c_u - b)[zy(b) - y(b)\epsilon]^+ + (b - c_r) zy(b) + c_r Q. \tag{C.1}
\]

The multiplicative expected cost function becomes:

\[
E[\Pi(z, b)] = (c_u - b) \int_{-\infty}^{z} [zy(b) - y(b)\epsilon] f(u) \, du + (b - c_r) zy(b) + c_r Q. \tag{C.2}
\]

Also, the first order differential equation is as follows:

\[
\frac{\partial E[\Pi(z, b)]}{\partial z} = (c_u - b) \int_{-\infty}^{z} y(b) f(u) \, du + (b - c_r) y(b)
\]

\[
= (c_u - b)y(b)F(z) + (b - c_r) y(b)
\]
\[ y(b)[(c_u - b)F(z) + (b - c_r)]. \tag{C.3} \]

For the first stage, we can solve for \( z^* \) using Leibniz’s rule as follows and our previously defined function, \( \Gamma(z) \):

\[ 0 = (c_u - b)F(z) + b - c_r \]
\[ z^*(b) = F^{-1}\left(\frac{c_r - b}{c_u - b}\right). \tag{C.4} \]

\[
\frac{\partial E[\Pi(z,b)]}{\partial b} = (c_u - b)\frac{dy(b)}{db} \Gamma(z) - (1)y(b)\Gamma(z) + z\left[(b - c_r)\frac{dy(b)}{db} + y(b)\right]
\]
\[
= \Gamma(z)[(c_u - b)(\alpha\beta b^{\beta-1}) - \alpha b^\beta] + z[(b - c_r) \alpha \beta b^{\beta-1} + \alpha b^\beta]. \tag{C.5} \]

We arrive at the following Lemma and first order differential equations for the decision variable \( b \):

**Lemma 2.** For a fixed \( z \), the optimal price is determined uniquely as a function of \( z \):

\[ b^* \equiv b(z) = \left(\frac{\beta}{\beta+1}\right)\frac{c_r z - c_u \Gamma(z)}{z - \Gamma(z)}. \]

**Proof.** Similar to the additive retention case by using the partial derivative of the expected cost with respect to \( b \) (equation 17),

\[ 0 = \alpha b^\beta \left[\frac{c_u \Gamma(z)^\beta}{b} - \Gamma(z)^\beta - \Gamma(z) + \beta z + \frac{c_r z \beta}{b} \right] \]
\[ \Gamma(z)(\beta + 1) - z(\beta + 1) = \frac{c_r z \beta - c_u \Gamma(z)^\beta}{b} \]
\[ b^* = \left(\frac{\beta}{\beta + 1}\right)\frac{c_r z - c_u \Gamma(z)}{z - \Gamma(z)}. \tag{C.6} \]

Continuing to the second order differential equations for the multiplicative demand case, the resulting equations are:

\[
\frac{\partial^2 E[\Pi(z,b)]}{\partial z^2} = y(b) (c_u - b) f(z), \tag{C.7} \]

\[
\frac{\partial^2 E[\Pi(z,b)]}{\partial b^2} = \alpha \beta b^{\beta-2} \left\{ \Gamma(z)((c_u - b)(\beta - 1) - 2 b) + z((b - c_r)(\beta - 1) + 2 b) \right\}, \tag{C.8} \]

\[
\frac{\partial^2 E[\Pi(z,b)]}{\partial z \partial b} = \frac{\partial^2 E[\Pi(z,b)]}{\partial b \partial z} = \alpha \beta b^{\beta-1} \left[ (c_u F(z) - c_r) + b(1 - F(z)) \right] + \alpha \beta \left[ 1 - F(z) \right]. \tag{C.9} \]
The Multiplicative Expected Cost function is convex $\forall b$ and $z$. Hence, we define and prove the following theorem in a similar approach that we used to prove Theorem 1.

**Theorem 2.** The single-period optimal stocking and bonus policy for the multiplicative demand case is to recruit $q^* = y(b^*)z^*$ personnel, and to offer a retention bonus of $b^*$, where $b^*$ is specified by Lemma 2 and $z^*$ is determined according to the following:

(a) If $F(\cdot)$ is an arbitrary distribution function, then an exhaustive search over all values of $z$ in the region $[A, B]$ will determine $z^*$.

(b) If $F(\cdot)$ is a distribution function satisfying the condition $2r(z)^2 + dr(z)/dz > 0$ for $A <= z <= B$, and $r(z) > \frac{2(zF(z) - \Gamma(z))}{z(\Gamma(z))}$, then $z^*$ is the unique $z$ in the region $[A, B]$ that satisfies $dE[\Pi(z, b(z))]/dz = 0$.

**Proof.** Using the chain rule and our first order results for $z$ and $b$ (Lemma 2), we first calculate the following derivatives of the multiplicative bonus function $b(z)$ and $dE[\Pi(z, b(z))]/dz$:

$$
\frac{db(z)}{dz} = \left(\frac{\beta}{\beta + 1}\right)\frac{(c_r - c_u F(z)) (z - \Gamma(z)) - (1 - F(z)) (c_r z - c_u \Gamma(z))}{(z - \Gamma(z))^2}
$$

$$
= \left(\frac{\beta}{\beta + 1}\right) \frac{c_r z - c_r \Gamma(z) - c_u z F(z) + c_u \Gamma(z) F(z) - c_r z + c_u \Gamma(z) + F(z) c_r z - c_u F(z) \Gamma(z)}{(z - \Gamma(z))^2}
$$

$$
= \left(\frac{\beta}{\beta + 1}\right) \frac{c_r - c_u}{{z - \Gamma(z))^2}} \frac{z F(z) - \Gamma(z)}
$$

$$
\frac{d^2 b(z)}{dz^2} = \left(\frac{\beta}{\beta + 1}\right) (c_r - c_u) \frac{(z F(z) + z f(z) - F(z)) (z - \Gamma(z))^2 - 2 (z F(z) - \Gamma(z)) (z - \Gamma(z)) (1 - F(z))}{(z - \Gamma(z))^4}
$$

$$
= \left(\frac{\beta}{\beta + 1}\right) (c_r - c_u) (1 - F(z)) \frac{z r(z)}{(z - \Gamma(z))^2} - 2 \left(\frac{\beta}{\beta + 1}\right) (c_r - c_u) \frac{z F(z) - \Gamma(z)) (1 - F(z))}{(z - \Gamma(z))^2} \frac{db(z)}{dz}
$$

$$
= \left(\frac{\beta}{\beta + 1}\right) (c_r - c_u) \frac{z r(z)}{(z - \Gamma(z))^2} + 2 \left(\frac{1 - F(z)}{z - \Gamma(z)}\right) \frac{db(z)}{dz}
$$

$$
\frac{dE[\Pi(z, b(z))]}{dz} = y(b(z)) \left[(c_u - b(z)) F(z) + (b(z) - c_r)\right].
$$

Next, let us define, $R(z) \equiv (c_u - b(z)) F(z) + (b(z) - c_r)$. The reason we can define $R(z)$ as we do, is because $y(b(z)) > 0$ for all values of $z$ (except the boundary B). $R(z)$ represents the behavior of $E[\Pi(z, b(z))]$ such that the expected value function is increasing for values of $z$ that satisfy $R(z) > 0$, and decreasing for values that satisfy $R(z) < 0$. Also, $R(z)$ has a local optimum point
for any \( z \) that satisfies \( R(z) = 0 \). Hence, we can analyze \( R(z) \) given it is sufficient for determining the shape of \( E[\Pi(z, b(z))] \). One must first see that \( R(A) < 0 \) and \( R(B) > 0 \):

\[
R(A) = (c_u - b(A)) F(A) + (b(A) - c_r) = (c_u - b(A)) \cdot 0 + b(A) - c_r = b(A) - c_r < 0,
\]

\[
R(B) = (c_u - b(B)) F(B) + (b(B) - c_r) = (c_u - b(B)) \cdot 1 + b(B) - c_r = c_u - c_r > 0.
\]

Next, we arrive at the following equation by calculating the first derivative of \( R(z) \) using the chain rule and the definition of \( r(\cdot) \) to observe the shape of \( R(z) \) with respect to \( z \),

\[
\frac{dR(z)}{dz} = (c_u - b(z)) f(z) - \frac{db(z)}{dz} F(z) + \frac{db(z)}{dz}
\]

\[
= (c_u - b(z)) r(z) (1 - F(z)) - \frac{db(z)}{dz} F(z) + \frac{db(z)}{dz}
\]

\[
= (1 - F(z)) [(c_u - b(z)) r(z) + \frac{db(z)}{dz}].
\]

Next, we rearrange the terms in the above equation to solve for \( \frac{db(z)}{dz} \),

\[
\frac{db(z)}{dz} = \frac{dR(z)}{dz} \cdot \frac{1}{1 - F(z)} + (b(z) - c_u) r(z).
\]

Again, using the chain rule on \( \frac{dR(z)}{dz} \) and the definition of \( r(\cdot) \) and \( \frac{db(z)}{dz} \), we arrive at the following second order derivative,

\[
\frac{d^2 R(z)}{dz^2} = -f(z) [(c_u - b(z)) r(z) + \frac{db(z)}{dz}] + (1 - F(z)) [(c_u - b(z)) \frac{dr(z)}{dz} - r(z) \frac{db(z)}{dz} + \frac{d^2 b(z)}{dz^2}]
\]

\[
= -r(z) (1 - F(z)) [(c_u - b(z)) r(z) + \frac{db(z)}{dz}] + (1 - F(z)) [(c_u - b(z)) \frac{dr(z)}{dz} - r(z) \frac{db(z)}{dz} + \frac{d^2 b(z)}{dz^2}].
\]

By setting \( \frac{dR(z)}{dz} = 0 \), substituting in the definition of \( \frac{db(z)}{dz} \) in terms of \( \frac{dR(z)}{dz} \) results in,

\[
\frac{d^2 R(z)}{dz^2} \bigg|_{\frac{dR(z)}{dz} = 0} = (1 - F(z)) [(c_u - b(z)) \frac{dr(z)}{dz} - r(z) \frac{db(z)}{dz} + \frac{d^2 b(z)}{dz^2}]
\]

\[
= (1 - F(z)) \left[ (c_u - b(z)) \frac{dr(z)}{dz} - r(z) \left[ \frac{dR(z)}{dz} \cdot \frac{1}{1 - F(z)} + (c_u - b(z)) r(z) \right] + \frac{d^2 b(z)}{dz^2} \right]
\]

\[
= (1 - F(z)) \left[ (c_u - b(z)) \frac{dr(z)}{dz} + (c_u - b(z)) r(z)^2 + \frac{d^2 b(z)}{dz^2} \right]
\]

\[
= (1 - F(z)) \left[ (c_u - b(z)) \left( \frac{dr(z)}{dz} + r(z)^2 \right) + \frac{d^2 b(z)}{dz^2} \right].
\]

Finally, by substituting for \( \frac{d^2 b(z)}{dz^2} \) we arrive at,

\[
\frac{d^2 R(z)}{dz^2} \bigg|_{\frac{dR(z)}{dz} = 0} = (1 - F(z)) \left[ (c_u - b(z)) \left( \frac{dr(z)}{dz} + r(z)^2 \right) + \left( \frac{\beta}{\beta + 1} \right) (c_r - c_u) \frac{z r(z)}{(\Gamma(z) - z)^2} \right]
\]
\[-2\left(\frac{1 - F(z)}{\Gamma(z) - z} \frac{db(z)}{dz}\right)\]
\[= (1 - F(z)) \left[ (c_u - b(z)) \left( \frac{dr(z)}{dz} + r(z)^2 \right) + \left( \frac{\beta}{\beta + 1} (c_r - c_u) \frac{zr(z)}{z - \Gamma(z)} \right) \right.\]
\[\left. - 2\frac{1 - F(z)}{z - \Gamma(z)} \left( \frac{\beta}{\beta + 1} (c_r - c_u) \frac{zF(z) - \Gamma(z)}{(z - \Gamma(z))^2} \right) \right]\]
\[\Rightarrow \frac{d^2R(z)}{dz^2} > 0\]

If \( r(z)^2 + \frac{dr(z)}{dz} > 0 \)

and

\[r(z) > \frac{2(zF(z) - \Gamma(z))}{z(z - \Gamma(z))}.
\]

In determining the shape of the final \( d^2R(z)/dz^2 \) equation, one can see that \( 1 - F(z) > 0 \), and \( c_u > b(z) \) as initially defined, and \( zF(z) - \Gamma(z) > 0 \) given the definition of \( \Gamma(z) \). Also, \( c_r - c_u < 0 \), \( z - \Gamma(z) > 0 \), and \( zr(z) > 0 \). Also, the entire second order differential equation, \( d^2R(z)/dz^2 \) is greater than zero. The final inequality is derived from the \( d^2b(z)/dz^2 \) term and must be satisfied to ensure convexity of \( R(z) \). Similar to Petruzzi and Dada’s result for their concave cost function, the last inequality also signifies that our defined function, \( R(z) \) is unimodal in \( z \), first decreasing and then increasing. Given the last two inequalities, \( E[\Pi(z, b(z))] \) reaches its minimum at the unique value of \( z \neq B \) that satisfies,

\[dE[\Pi(z, b(z))] / dz = y(b(z))R(z) = 0. \quad \square\]

### C.0.0.2 Multiplicative Reformulation and Numerical Examples

To gain further insights into the multiplicative model, let us define \( w \), as \( w = (c_r - b)/(c_u - b) \), \( b = (c_r - c_u)/(1 - w) + c_u \), and let \( z^*(b) = F^{-1}(w) \). The resulting multiplicative expected value function becomes:

\[E[\Pi(w)] = \frac{c_u - c_r}{1 - w} y(\frac{c_r - c_u}{1 - w} + c_u) \int_{-\infty}^{F^{-1}(w)} (F^{-1}(w) - u) f(u)d(u)\]
\begin{equation}
+ (c_u - c_r)(\frac{-w}{1-w}) F^{-1}(w) \ y(\frac{c_r - c_u}{1-w} + c_u) + c_r Q. \tag{C.10}
\end{equation}

We conduct numerical and simulation experiments using the previously defined equations in this Appendix using the definition that \(y(b) = ab^b\) and the Normal cumulative distribution function (CDF). Again, not included in this dissertation, we conducted experiments using the Exponential CDF, and confirmed that our theoretical results match simulation experiments. The results from our Normality assumed theoretical and simulated experiment are shown below:

![Figure C.1: Derived E(q,b) for Multiplicative Case: \(c_r/c_u = 0.5\)](image1)

![Figure C.2: Simulated E(q,b) for Multiplicative Case: \(c_r/c_u = 0.5\)](image2)

![Figure C.3: Derived E(q,b) for Multiplicative Case: \(c_r/c_u = 0.75\)](image3)

![Figure C.4: Simulated E(q,b) for Multiplicative Case: \(c_r/c_u = 0.75\)](image4)

Figures C.1-C.4 essentially validate our theoretical results that the expected cost function is
convex if the conditions listed in Theorem 2 are not violated. Furthermore, one can observe that
the expected cost increases as the number of recruits and bonus levels are increased, which is an
intuitive result.

Continuing in our analysis of the multiplicative model of the expected cost function \((E(w))\)
one can see from Figures C.5-C.6 that indeed the function is convex and one can determine the
optimal bonus value \(b\) given the minimum value of \(w\). One can see that this model improves the
manner in which retention and recruiting behaves under different underage, recruiting and bonus
levels. Although we only provide results for our Normal CDF numerical experiments, we arrive at
nearly identical expected cost “shape” results using an Exponential CDF.

![Figure C.5: E(w) for Multiplicative Retention Case: \(c_r/c_u = 0.25, \alpha = 30\)](image)

![Figure C.6: E(w) for Multiplicative Retention Case: \(c_r/c_u = 0.75, \alpha = 30\)](image)

As we show in the previous figures, the multiplicative expected cost function is confirmed to
be convex. For small \(w\) ratios the cost function increases linearly as the ratio increases towards
0.25. Interpreting Figure C.5, it is evident that the expected cost plateaus at an approximate 0.2
ratio. One can infer from the expected cost function’s behavior shown in Figures C.5-C.6, that
if one considers recruiting and retention to behave in a multiplicative manner, the underage cost
must be set much higher than the recruiting cost. If the underage cost fails the conditions set
forth in Theorem 2, then the expected cost for retention and recruiting behaves erratically. Given
this erratic behavior for a multiplicative retention demand assumption and exponential cumulative
distribution function, in our empirical analysis we analyzed the expected cost function assuming a more realistic and appropriate Binomial Distribution.