National Defence Défense nationale

Canadian Forces College

Collège des Forces Canadiennes

MACHINE LEARNING ANALYSIS OF FEMALE VOLUNTARY ATTRITION IN THE CANADIAN ARMED FORCES

Maj Jason Estrela

JCSP 44

Master of Defence Studies

PCEMI 44

Maîtrise en études de la défense

Canada

CANADIAN FORCES COLLEGE – COLLÈGE DES FORCES CANADIENNES
JCSP 44 – PCEMI 44
2017 – 2018

MASTER OF DEFENCE STUDIES – MAÎTRISE EN ÉTUDES DE LA DÉFENSE

MACHINE LEARNING ANALYSIS OF FEMALE VOLUNTARY ATTRITION IN THE CANADIAN ARMED FORCES

Maj Jason Estrela

Word Count: 17,421

Compte de mots: 17,421

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Attrition amongst females in the Canadian Armed Forces (CAF) is an important contemporary subject as the organization seeks to achieve a mandated 25% female workforce. CAF human resources research organizations have used several methods to forecast attrition as part of workforce planning and policy development. The intent of this paper is to compliment such research by applying different Machine Learning (ML) techniques to demographic and occupational data contained in the CAF Human Resources Management System (HRMS) in order to predict likely voluntary releases amongst the female population. To that end, three models were trained and tested: a logistic regression (logit) model that included only main effects variables; a logit model that included both main effects and all pairwise interactions; and a model trained using extreme gradient boosted (xgboost) trees. To build the models, missing data was addressed using multiple imputation and regularization was used to avoid overfitting.

The study found that the boosted tree model significantly outperformed predictions from the other two models, achieving 91.0% total accuracy, 88.9% balanced accuracy, and a Kappa score of .794 on the test data under the default decision threshold. Analysis was also performed on the impacts of adjusting the decision threshold based on the relative costs of False Positives (FP) versus False Negatives (FN) or maximizing Cohen's Kappa coefficient. Discussion also included cursory analysis of variable importance and the applicability of the models to CAF attrition forecasting. Ultimately, it is hoped that this study will inform future work on applying ML to human resources management and other data-driven applications in the CAF.

MACHINE LEARNING ANALYSIS OF FEMALE VOLUNTARY ATTRITION IN THE CANADIAN ARMED FORCES

INTRODUCTION

The recruitment and retention of females in the Canadian Armed Forces (CAF) has emerged as a significant human resources issue as it seeks to become a more diverse and inclusive military force. In 2011, updating a previous report to the Office of the Chief Human Resources Officer of Canada, the CAF committed to "diversity as a fundamental value for the CAF" and sought to increase the representation of women in the force to 25%.[1] To that end, the CAF has been trying to increase the proportion of women by 1% per year from an original start-state of 14% representation. Despite these intentions, a 2016 review by the Office of the Auditor General (OAG) of Canada found that the CAF fell far short of achieving these numbers and was deficient in strategy to attain these goals.[2]

In an action plan responding to the 2016 OAG report, the CAF detailed a set of initiatives to rectify these strategic deficiencies. While a number of these initiatives target increasing recruitment, there is scant detail pertaining to retention, with the only mention being that the "CAF Retention Strategy will include policies and programs that are tailored to women members in particular."[3] Implied in this statement are the requirements for better understanding of attrition amongst CAF females and for analysis

---

[1]Department of National Defence, *Canadian Forces Employment Equity Report 2011-2012* (Ottawa: DND Canada, October 2012): 2, 4, http://publications.gc.ca/collections/collection_2014/mdn-dnd/D3-31-2012-eng.pdf.

[2]Office of the Auditor General of Canada, "Canadian Armed Forces Recruitment and Retention – National Defence," last accessed 10 March 2018, http://www.oag-bvg.gc.ca/internet/English/parl_oag_201611_05_e_41834.html#.

[3]House of Commons of Canada, *Detailed Action Plan for OAG Report Recommendations* (Ottawa: HoC Canada, n.d): 4, http://www.ourcommons.ca/Content/Committee/421/PACP/WebDoc/WD8148750/Action_Plans/43-DepOfNatDef-e.pdf.

tools to aid in planning and policy implementation. In doing so, the CAF will be better postured to develop, implement, and assess female retention policies with the ultimate intent of reducing unnecessary attrition that, along with recruitment initiatives, will be critical to achieving the mandated target of a 25% female workforce.

Attrition is of particular importance in workforce discussions due to the associated costs, such as those incurred to mitigate the loss of corporate knowledge and replace the employee. For example, applying Bliss' categorizations of employee turnover costs to a military context reveal the following impacts on the CAF:[4]

- *Lost Personnel and Productivity Costs:* filling or mitigating vacancies, lost productivity, lost corporate knowledge, sunk costs in employee training and development, and the cost of personnel needed to process release administration.
- *Recruitment and New Hire Costs:* recruiting staff, intake administrative personnel, and advertising.
- *Training Costs:* all costs associated with orientation and training for new employees. This constitutes a particular concern to the CAF given limited capacity in the training system and the fact that militaries cannot take in direct hires for higher-level positions.

As a result, this study will focus on retention and attrition by exploring the use of Machine Learning (ML) techniques to build predictive models from data captured in the

[4]William G. Bliss, "Calculating the Costs of Employee Turnover," *Fairfield County Business Journal* 40, no. 32 (2001): 12. All bulleted information was taken from this reference.

Human Resources Management System (HRMS). With that intent, this study will answer the following questions:

- What variables collected in the CAF HRMS function as predictors for voluntary release amongst females in the CAF?
- How well can ML be used to predict the risk of voluntary release for females in the CAF and which techniques work best?

Given the breadth of factors affecting retention, this study will be bounded in order to ensure manageable scope. It will be restricted to analyzing demographic and occupational variables collected in HRMS as opposed to social, psychological, and behavioural variables that are not readily available and are best collected via focussed research methods. Given the differences between employment in the Regular Force (Reg F; full-time) and Reserve Force (Res F; full-time and part-time), this analysis will focus on female attrition in the Reg F only. Also, this study will be limited to voluntary releases assuming that the primary intent of future CAF HR policies in this area will focus on avoiding preventable voluntary attrition versus other types of attrition, such as that associated with medical or disciplinary issues. One of the strengths of data science and ML are that tools can be built to enable front-end analysts to continuously analyze data collected over time, which can then be used to quickly update predictions. With that in mind, this project will produce code that could feasibly work with different samples as the data in HRMS changes from year-to-year. It is infeasible to apply every ML method to this project given the multitude of available techniques, thus modelling will be limited to supervised learning comparing two different model types that are appropriate for the

problem to be solved: logistic (logit) regression and extreme gradient boosted (xgboost) trees.

To begin, a literature review will be conducted in order to establish baseline theoretical knowledge and explore the applicability of previous work to the current study. The literature review will identify previously identified relationships that factor into predictor variable selection – i.e. variables that have been identified in previous work will be candidates for model building. This step is performed to reduce the possibility of overfitting; however, certain variables or categories were also introduced as candidates based on logical reasoning from the author's domain knowledge. Specific focus in the literature review will be paid to work performed by Defence Research & Development Canada (DRDC) and Director General Military Personnel Research and Analysis (DGMPRA); the latter being the CAF's official HR research body that focuses on forecasting attrition and retention amongst other efforts.

After the literature review, this paper will present the methodology used to clean the collected data set, impute missing data, build each model, and assess model performance. Comparisons of each model will be presented with ensuing discussion focussed on assessing metrics on model performance from a predictive perspective while also including discussion of trade-offs amongst metrics, modifying decision thresholds, and some limited discussion of variable importance. Finally, the study will conclude by commenting on the applicability of the analysis and tools to the CAF as well as future work in the area.

LITERATURE REVIEW

Chief Military Personnel (CMP) General Reports on the CAF

Chief Military Personnel (CMP), as the Chief Human Resources Officer in the CAF, commissions DGMPRA to produce two major annual studies that are important in this discussion: the Annual Report on Reg F Personnel, which provides context on the overall demographics of the CAF; and the Annual Report on Reg F Attrition, which provides information on trends in CAF attrition over time. Reports on Attrition from 2014/2015 and Personnel from 2015/2016 were obtained with the highlights of each presented in the remainder of this section.

First, it is important to note the characteristics of the CAF population and how it has changed over time. Events that influence the overall population in certain periods will have downstream effects on the demographic make-up of the CAF and thus, can influence attrition. The most influential of these changes involves the CAF Force Reduction Plan (FRP). Initiated in 1991, the FRP was intended to reduce the size of the CAF population from 88, 800 to a targeted strength of 60, 000 personnel over the course of several years.[5] Due to the FRP, the Trained Effective Strength (TES) of the CAF not only declined from Fiscal Year (FY) 1990/1991 to FY 1996/1997 during implementation, but continued to decline until a recruiting surge was authorized in FY 2000/2001, which led to a re-expansion of the Forces that was completed in 2009/2010.[6] Figure 2.1 shows the end of this period of decline and subsequent surge to the current state.

---

[5]Department of National Defence, *Audit of Force Reduction Program* (Ottawa: Director General Audits, January 1997): 1, http://www.forces.gc.ca/assets/FORCES_Internet/docs/en/about-reports-pubs-audit-eval/705529.pdf.

[6]Department of National Defence, *Annual Report on Regular Force Personnel 2015/2016* (Ottawa: Director General Military Personnel Research and Analysis, October 2017): 2.

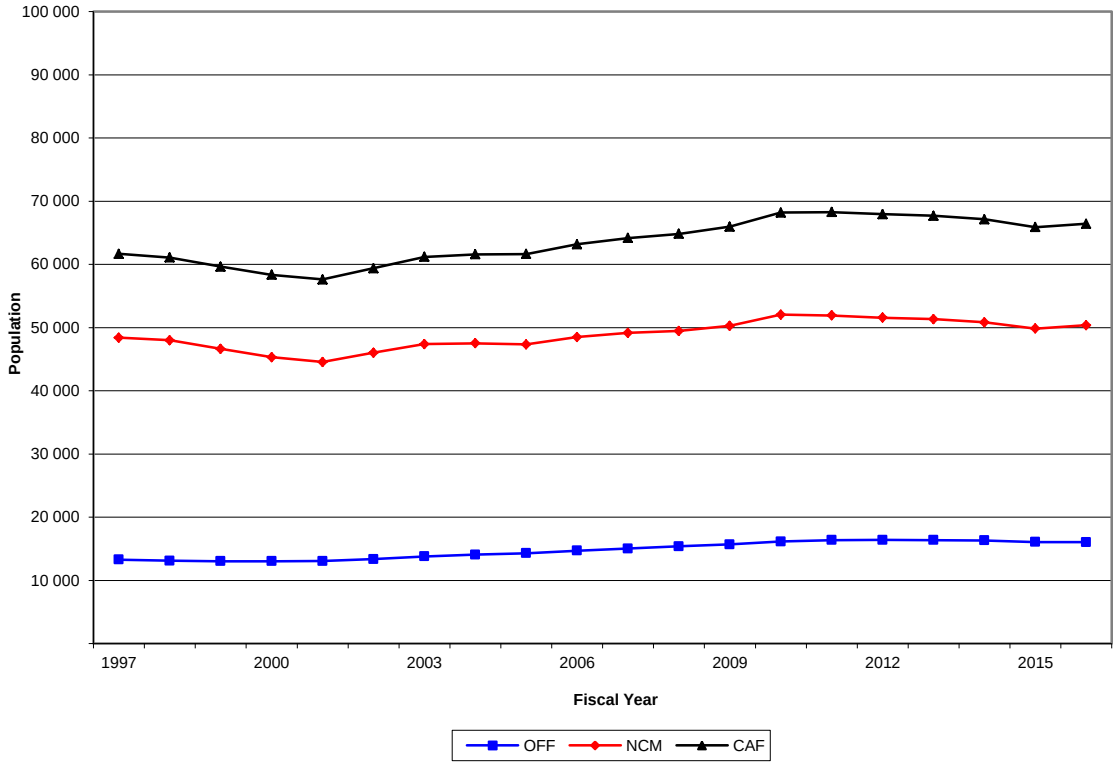

Figure 2.1 – Regular Force Population as of 31 March 2016

Source: Department of National Defence, Annual Report on Reg F Personnel, 3.

The FRP and minor surge that followed it are likely to have impacted today's population profile, evident in Figures 2.2 and 2.3. Due to the rapid outflow of personnel, there is a valley in the population profile between 20-24 years of service (YOS) for both officers and non-commissioned members (NCM). Contrarily, the amount of personnel with "less than 15 YOS exceeds the proportion with 15 or more YOS, particularly for NCMs, which can be attributed to the force expansion from 2001 to 2010."[7] With the end of the recruiting surge, the proportion of personnel with lower YOS decreased for several years, with the trend appearing to reverse as of 2015/2016.[8]

[7] *Ibid.*, 5.
[8] *Ibid.*

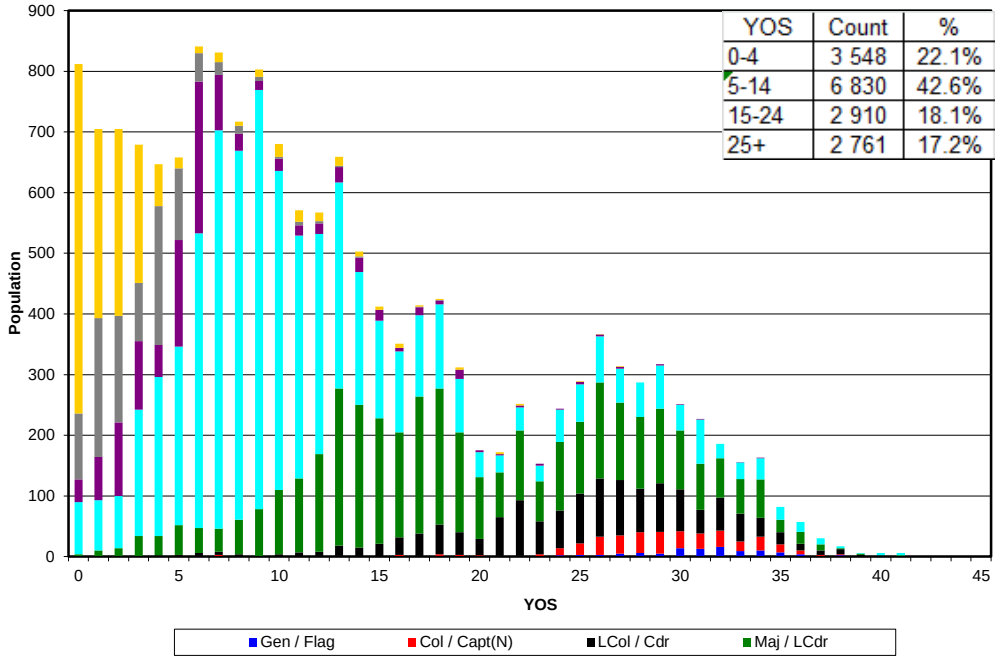

Figure 2.2 – Officer Population Profile by YOS and Rank as of 31 March 2016

Source: Department of National Defence, Annual Report on Reg F Personnel, 5.

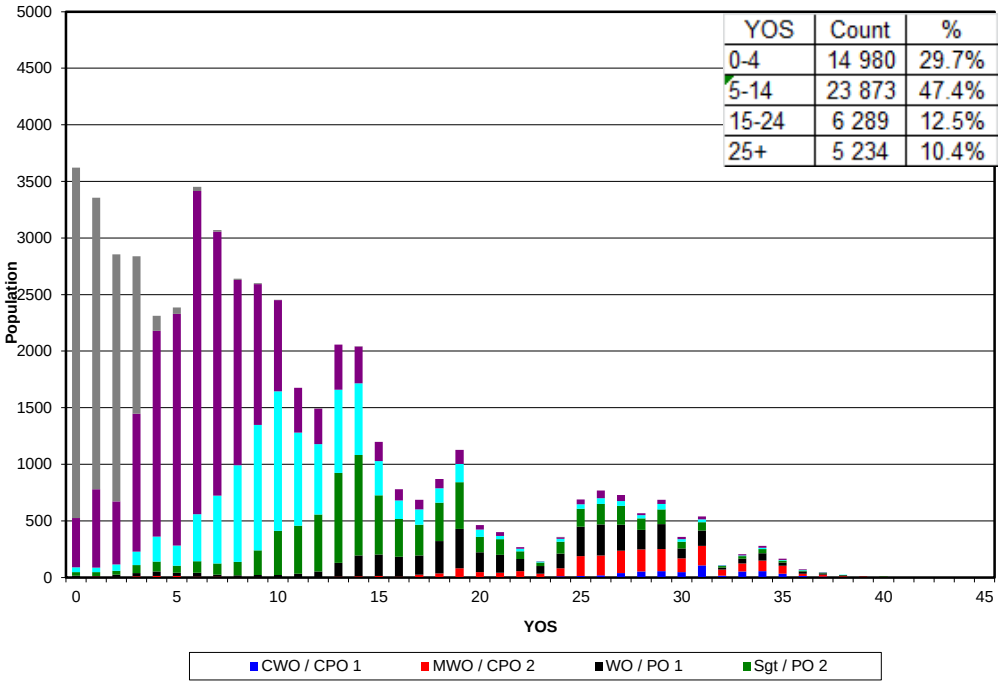

Figure 2.3 – NCM Population Profile by YOS and Rank as of 31 March 2016

Source: Department of National Defence, Annual Report on Reg F Personnel, 6.

The FRP has also impacted other demographics. For example, the Report on Personnel observes that this period saw a general increase in the amount time spent in rank (TIR) for each member while the expansion period had the opposite effect; it created new vacancies that had to be filled through increased promotion.[9] The major implications for the present study are how the FRP and post-FRP expansion will impact the distribution of associated variables as well as how they might skew results. For example, it is expected that personnel may be more likely to terminate during a period where they have 20-24 YOS due to the fact that they are eligible for an immediate annuity; however, we may see less occurrences of this event in later years in the sample due to the smaller population that currently makes up this segment since a lot of personnel that would be part of this demographic terminated at a higher rate during FRP.

Another interesting aspect of the CAF population stems from its linguistic make-up in both first and second official languages, noting that the latter has an impact on career progression, particularly for officers and senior NCMs. As the Report on Personnel notes, Francophone officers and NCMs were much more likely to undergo second-language testing and meet the higher bilingualism requirements asked of CAF members in leadership positions.[10] In particular, Francophones were much more likely to meet the highest required profiles at a rate "almost 6 times higher for officers and 14 times higher for NCMs than amongst Anglophones."[11] As a result, these imbalances in the population merit analysis for potential impact on attrition because, at certain ranks, it is impractical

[9] *Ibid.,* 11.
[10] *Ibid.,* 14.
[11] *Ibid.*

to get promoted or to be part of succession planning without first attaining the required second-language proficiency.

The Report on Personnel also provides statistics on gender make-up in the CAF. It notes that the percentage of women in the CAF Reg F has been growing at various rates since the late 1990s (Figure 2.4), achieving a level of 17.3% among officers and 13.4% among NCMs for an overall workforce percentage of 14.4% as of FY 2015/1016.[12]

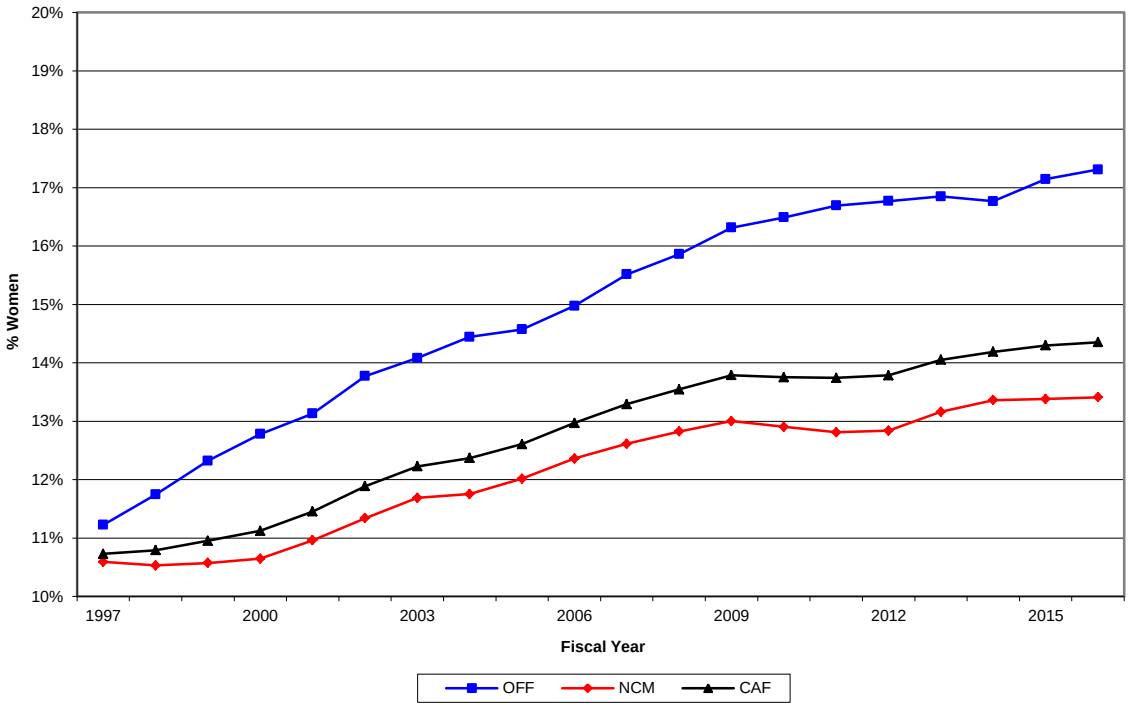

Figure 2.4 – Representation of Women in the CAF as of 31 March 2016

Source: Department of National Defence, Annual Report on Reg F Personnel, 19.

As part of employment equity reporting, the Report on Personnel applies the "80% Rule" to promotion rates in order to look for disparity between designated and non-designated groups (NDG). Roughly stated, "[w]hen the size of the gap between a

[12]Department of National Defence, *Annual Report on Regular Force Personnel 2015/2016*…, 17.

designated group's promotion rate and the NDG promotion rate reaches 20%, this is a warning sign that barriers may exist, which have caused a gap of this magnitude."[13] As it pertains to women in the CAF, no promotion barriers have been found to exist. Female officer and NCM promotion rates have kept within this "80% Rule" since 2000/2001 having closely mirrored NDG promotion rates and even surpassing them in several years.[14] It is important to note that the report does not make mention of how it controlled for comparison of different population sizes between genders (i.e. using standardized residuals), so it is assumed that these are raw promotion rates. Regardless, it provides an indicator that we should not expect promotion barriers to affect female attrition; however, standard career motivators and expectations when it comes to promotion may still apply.

To complement the statistics on the overall CAF population, the CAF Annual Report on Attrition yields some interesting observations. The report from 2014/2015 indicates that "a number of demographic variables have been shown to affect attrition, including age, gender, and YOS", with the latter variable, and Terms of Service (TOS) by extension, being the most influential.[15] Historical attrition rates for the overall population are found at Figure 2.5. The graph shows the expected outlier years due to the FRP, but also indicates higher-than-normal attrition between 2007 and 2009 as well as an increasing rate in later years. No explanation is given for the higher attrition rates outside of the FRP years. As such, these could be the result of CAF population changes experienced during the recruiting surge in the early 2000s or could even be hypothesized to be linked to some other major event such as the changes in the CAF mission in

[13]*Ibid.,* 39.
[14]*Ibid.,* 40.
[15]Department of National Defence, *Annual Report on Regular Force Attrition 2014/2015* (Ottawa: Director General Military Personnel Research and Analysis, June 2016), 1.

Afghanistan that began occurring in 2010. Specific focus on these issues is outside of the scope of the present study but is presented for completeness, noting that the information provided covers all types of releases, not just those that are voluntary.

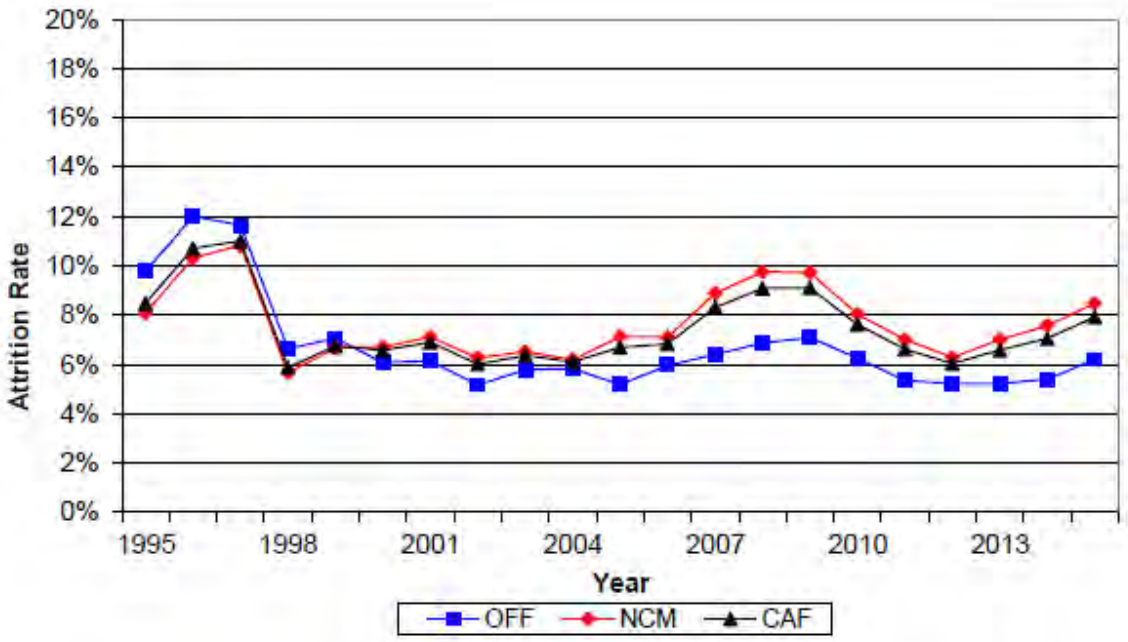

Figure 2.5 – Historical CAF Attrition Rates since 1995

Source: Department of National Defence, Annual Report on Reg F Attrition, 2.

Attrition rates by YOS for officers and NCMs are presented as Figures 2.6 and 2.7, which identify "likely exit points" based on career stage and TOS. For example, attrition tends to be "higher during initial training and after obligatory service has been completed", which corresponds to the first year of service, the ~9 year range for officers, and 3-6 years for NCMs.[16] The report further explains that at "20 YOS, personnel under the old TOS are eligible for an immediate annuity; therefore, larger proportions leave at

[16] *Ibid.*, 5.

20 YOS and beyond" with attrition beyond 35 YOS being exceedingly high given pension maximization and the lower amount of personnel serving at these levels.[17]

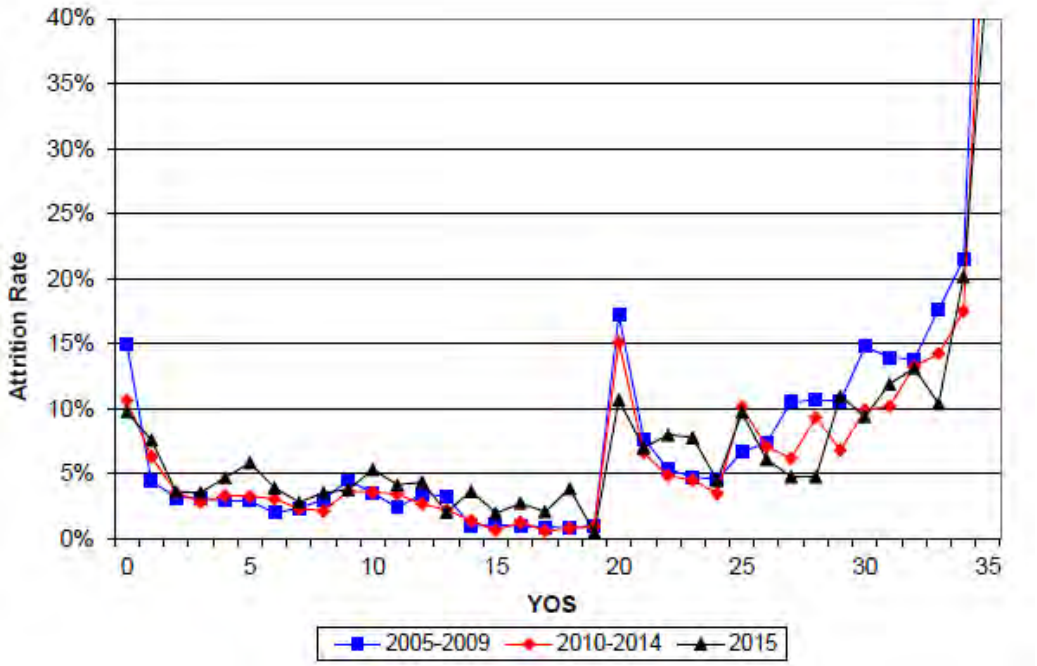

Figure 2.6 – Attrition Rate by YOS for Officers

Source: Department of National Defence, Annual Report on Reg F Attrition, 6.

[17] *Ibid.*

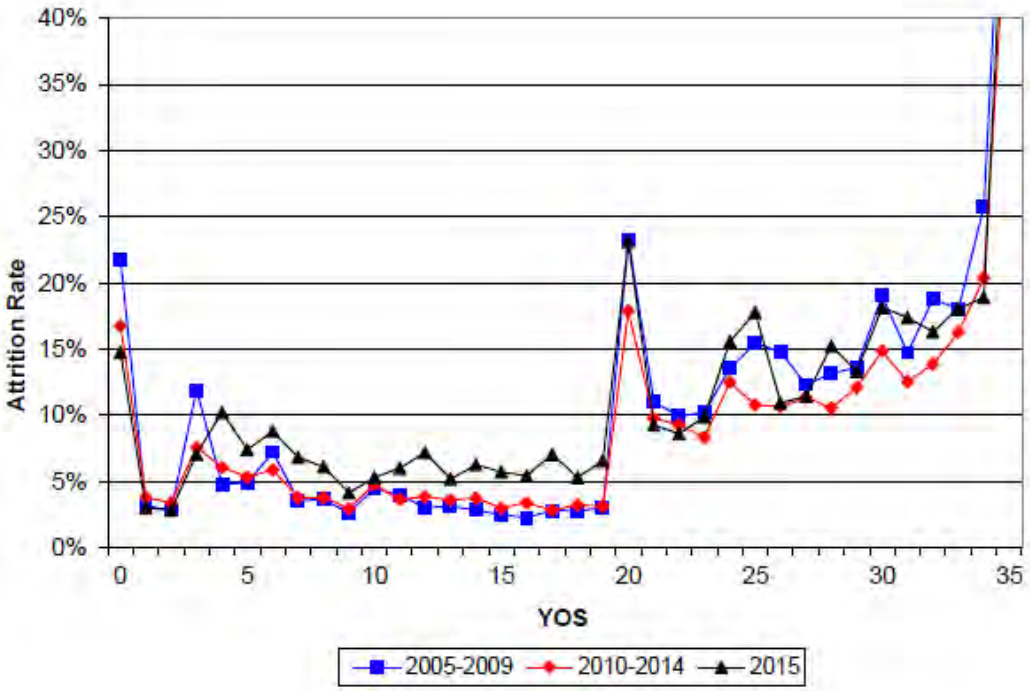

Figure 2.7 – Attrition Rate by YOS for NCMs

Source: Department of National Defence, Annual Report on Reg F Attrition, 6.

The differences in attrition rates during different time periods are also revealing. Period breaks depicted in Figures 2.6 and 2.7 are attributed to the changes in YOS-based attrition patterns due to changes in TOS that occurred in 2005.[18] For example, "the peak at 3 YOS in the NCM profile has decreased and attrition rates from 4 to 9 YOS have increased, reflecting the change from a Basic Engagement (BE) of 3 YOS to a Variable Initial Engagement (VIE) of 3 to 9 YOS."[19] A peak in attrition in 25 YOS is also beginning to form for both officers and NCMs; however, the report does not view this as stemming from the change in immediate annuity entitlements from 20 to 25 YOS given that many members were "grandfathered" under the old plans.[20] As a result, the effects of

[18] *Ibid.*
[19] *Ibid.*
[20] *Ibid.*

the longer-term changes to TOS are not expected to be revealed for several years. Ultimately, these observations justify the inclusion of YOS in attrition analysis and indicate that the engagement or contract type under which the member is serving could also be a significant variable.

The Report on Attrition also provides historical information on release rates based on the type of release, which corresponds to whether the release was compulsory, medical, voluntary, etc. Noting that the present study is only concerned with voluntary releases, some observations can be pulled from Figures 2.8 and 2.9, which correspond to release rate by type of release for officers and NCMs respectively. For officers, it is observed that there was a steady increase in the voluntary release rate in the late 90s, achieving a peak in 1999 after the FRP period before experiencing a decline that bottomed out in 2005. After this time, there was again a steady increase in the rate, culminating in 2009 before again declining. In both of these periods, the peaks were seen at ~4% with valleys at ~2%. The report does not provide potential explanations for either trend, but it could be hypothesized that the first peak stemmed from effects of the FRP, while the latter could again be the result of the recruiting surge. For NCMs, a markedly different pattern is observed with a relatively constant but slightly increasing voluntary release rate seen until the period between 2007 and 2009, where a sudden spike in releases occurs before returning back to the previous trend. Again, no hypothesis or explanation is given for this spike, but clearly the cause of the increased attrition seen during this period was exacerbated for NCMs.

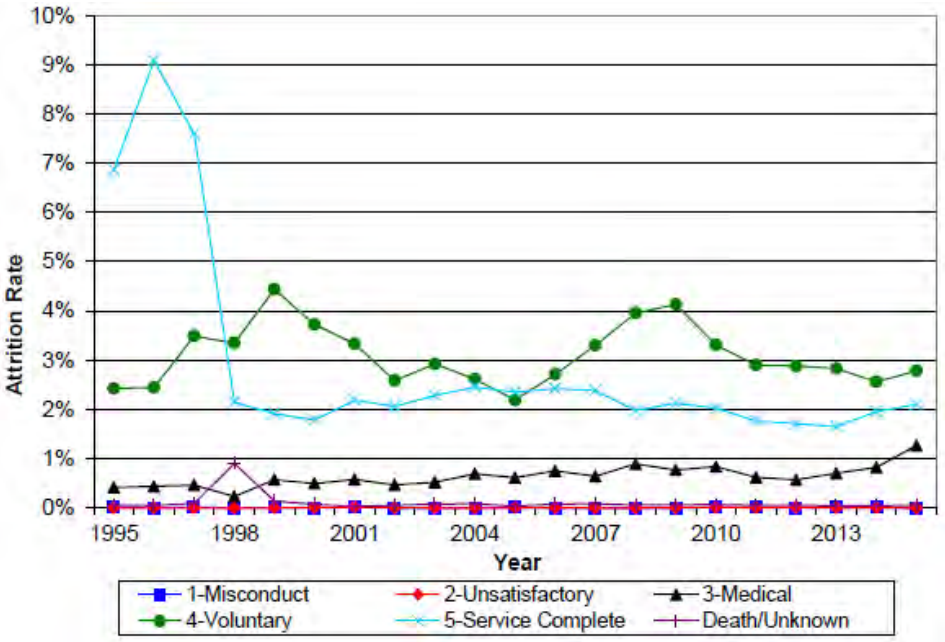

Figure 2.8 – Attrition Rate by Reason for Officers

Source: Department of National Defence, Annual Report on Reg F Attrition, 14.

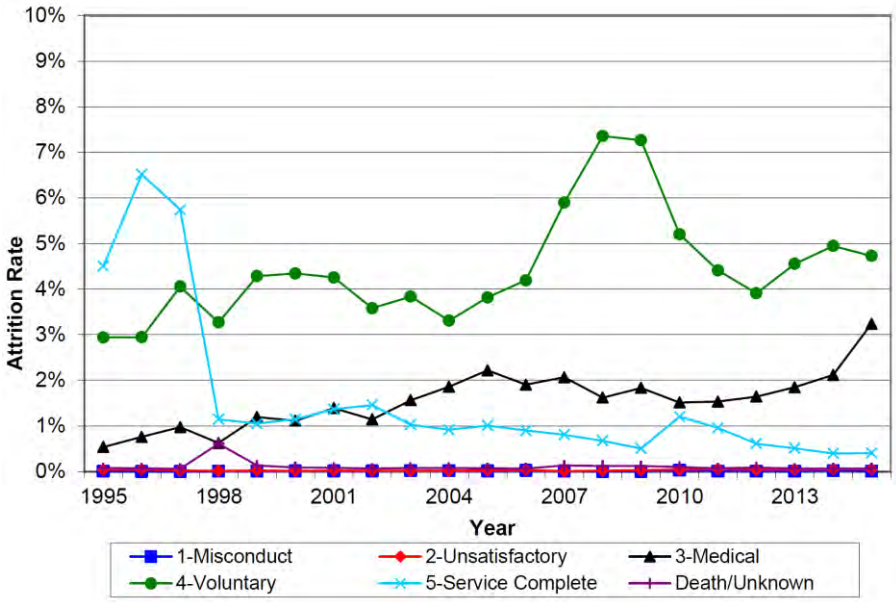

Figure 2.9 – Attrition Rate by Reason for NCMs

Source: Department of National Defence, Annual Report on Reg F Attrition, 15.

In terms of environmental trends, the report concludes that environment (Land, Sea, or Air) does not function as a significant predictor of attrition; however, it does note that over the past decade, the Land environment, and Land NCMs in particular, have experienced the highest levels of attrition aside from a spike seen with Air NCMs in recent years.[21] While the report doesn't indicate a relationship between environment and attrition, the higher rate of attrition amongst Land personnel and recent trends amongst RCAF NCMs, indicate that it is reasonable to include in attrition modelling. Workplace conditions may be a factor in identifying risk groups so it is feasible that environment bears further study, particularly as the report does not break attrition out by release type for each environment. Another consideration is the existence of trades that are assigned to an environment but are managed by Assistant CMP (Asst CMP) as joint trades that either span environments or are not environment-specific. This is not explicitly covered in the report but was looked at during cleaning of the study's data set, discussed in the next chapter.

As it pertains to rank, attrition was seen to be highest amongst General/Flag Officers and Chief Warrant Officers/Chief Petty Officers 1st Class; however, this is expected as these are very senior ranks, which would have correspondingly higher YOS and annuity entitlements.[22] Conversely, the report notes lower attrition amongst junior officer and tradesmen ranks.[23] Ultimately, rank may not function as a significant predictor of attrition outside of its relationship to YOS and TOS; however, it should be

[21]Department of National Defence, *Annual Report on Regular Force Attrition 2014/2015*…, 21.
[22]*Ibid.,* 22.
[23]*Ibid.*

included in the study as further analysis may reveal some effects of mid-career transition or other job-related factors.

Age reveals a similar attrition pattern to YOS as it is reflective of life and career stage for CAF members. Higher attrition rates are observed for both officers and NCMs under the age of 20, which likely corresponds to the high rates of first-year attrition. After this point for officers, attrition quickly drops below 5% and remains relatively level until the age of 36 when it begins to increase as officers begin to retire.[24] NCMs follow a similar pattern aside from a steadily increasing rate after the age of 20, which may correspond to regular attrition patterns for those that are not seeking a long-term career in the military. Ultimately, age may be a difficult variable of study given this non-linear relationship, but given the strong relationship observed at higher ages, it will be considered in modelling.

In terms of First Official Language (FOL), the Report indicates some inconsistency in its relationship to attrition:

> For officers, attrition rates have been slightly higher for Francophones than for Anglophones since FY 2007/2008. For NCMs, attrition rates have been slightly lower for Francophones than for Anglophones since FY 2006/2007; however, in the past 5 years, the difference between the two rates has widened.[25]

It had been expected that language-based attrition would result from the impact of second-language proficiency on career progression. Given the fact that Francophone officers and NCMs were much more likely to have a higher second-language profile than Anglophones, the observations from the Report on Attrition do not consistently follow these expected relationships. It is possible that other factors come into play. For

[24] *Ibid.*
[25] *Ibid,* 26.

example, it is possible that there are increased external opportunities (i.e. in the Public Service) that are available to those with higher second-language proficiency, which may lead to some attrition or potentially that language-based dissatisfiers are different for each group. Ultimately, this provides an indication that FOL and language proficiency may be important variables in the present study.

Changes in policy pertaining to education have also had an impact on attrition. The Report on Attrition hypothesizes that the move towards a degreed officer corps has likely led to an increased attrition rate for officers with "Less than High School", which had been previously a low-attrition group.[26] For NCMs, this group had consistently been a high-attrition group in the past but has since been supplanted in the past few years where NCMs with a Master's degree or higher, have demonstrated one of the highest attrition rates.[27] These changes likely stem from trends in the wider Canadian population, since education level has been increasing for all members of the CAF, trending towards increased education amongst new recruits in both the officer and NCM populations.[28]

Differences in marital status are also seen to impact attrition. Interestingly, CAF statistics show that "attrition amongst members in common-law relationships is typically the lowest", while perhaps unexpectedly, it is highest amongst those that are separated or divorced.[29] It might be expected that given the rigours of military life, attrition would be lowest amongst single members; however, for some unobserved reason, attrition in this cohort is higher compared to common-law individuals but lower compared to married

---

[26]*Ibid.,* 27.

[27]Department of National Defence*, Annual Report on Regular Force Attrition 2014/2015…,* 27.

[28]Department of National Defence, *Annual Report on Regular Force Personnel 2015/2016* (Ottawa: Director General Military Personnel Research and Analysis, October 2017), 21.

[29]Department of National Defence, *Annual Report on Regular Force Attrition 2014/2015* (Ottawa: Director General Military Personnel Research and Analysis, June 2016), 38.

individuals. Thus, marital status provides an interesting variable for study, but there may be unexplained differences between similar categories (i.e. between common-law and married personnel).

The Report on Attrition also conducted a cursory look at attrition by province of residence. They found that the rates were higher in Alberta, particularly for NCMs, which is expected considering that jobs in the oil and gas industries are likely to entice members to leave in search of higher pay.[30] Conversely, provinces with lower or similar attrition rates to the overall population included Ontario, New Brunswick, and Nova Scotia.[31] While these results are of interest to the present study, they do not reveal much in the way of relationships between geolocation and attrition, given the plethora of CAF postings across Canada. It is possible that there may be a relationship between location and attrition based on regional and metropolitan economic factors as well as the potential for "retirement postings", where a given location may be a popular area for retired members based on family life and post-service opportunities. As a result, there is merit to more granular investigation of location-based factors in attrition modelling rather than simply looking at province of residence.

It is also useful to review the comparison of females against NDGs to see if there are any gender-related patterns of attrition. Figures 2.10 and 2.11 show the attrition rates for males and females. These graphs show that attrition patterns between women and men are actually similar, particularly for NCMs. There are several points in the graph where the two curves cross over each other, noting several years where female attrition was lower than that of males. Consequently, it would not be expected that a larger study

[30] *Ibid.,* 42.
[31] *Ibid.*

of the entire CAF population would reveal gender differences; however, models representing the two groups may still show differences in what variables predict the likelihood of termination, particularly since these graphs represent overall attrition, not just voluntary releases.

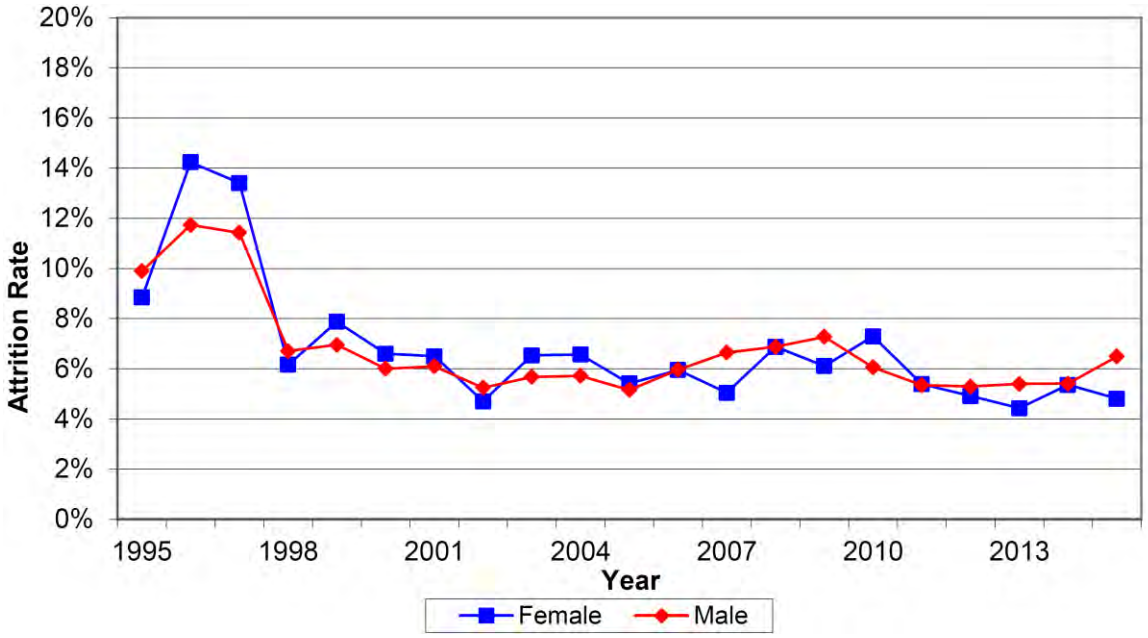

Figure 2.10 – Officer Attrition Rates by Gender

Source: Department of National Defence, Annual Report on Reg F Attrition, 34.

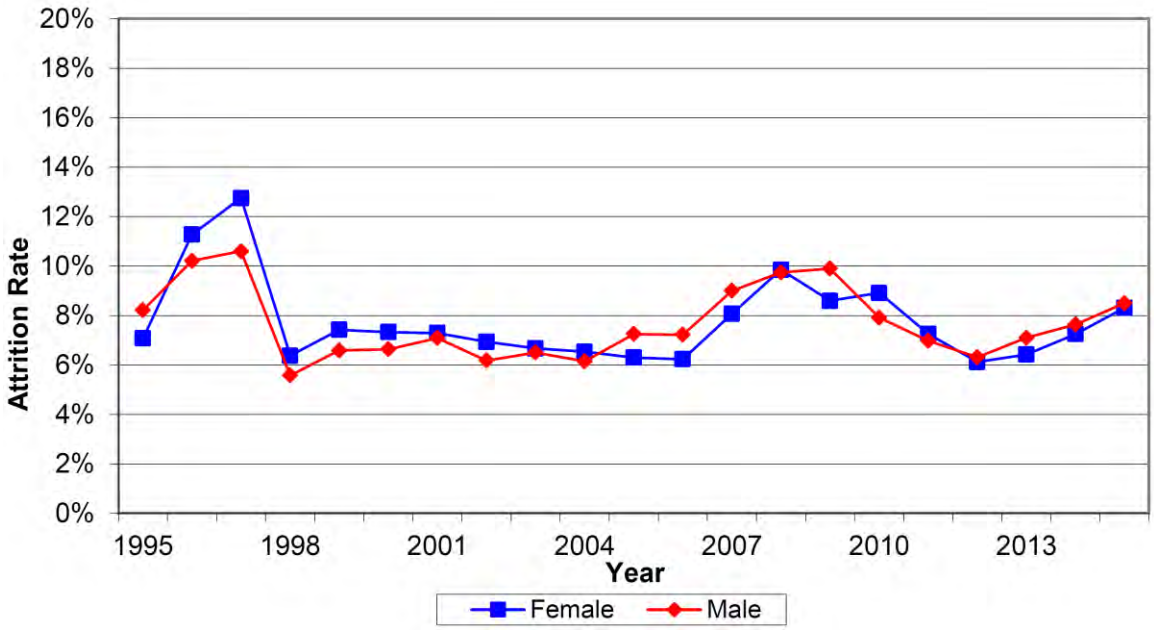

Figure 2.11 – NCM Attrition Rates by Gender

Source: Department of National Defence, Annual Report on Reg F Attrition, 34.

Despite these observations, the Report on Attrition notes some differences when placing women into groups based on YOS. Comparing the weighted average attrition rate over the previous five years for officers and NCMs in ten-year increments (Figures 2.12 and 2.13), attrition was higher amongst females for all but the 0-9 YOS grouping even though overall attrition was slightly higher for males.

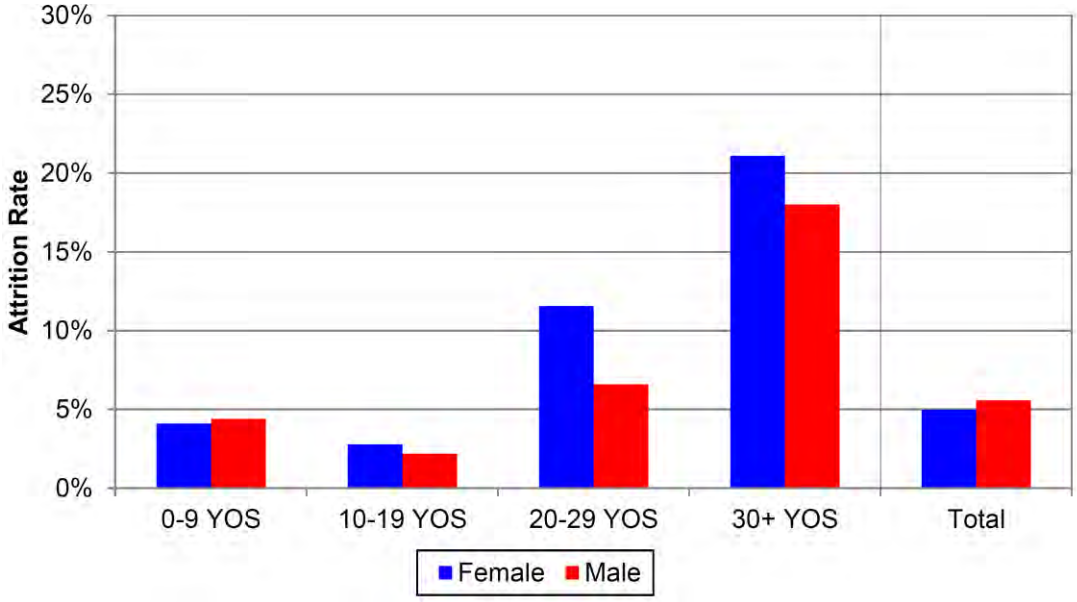

Figure 2.12 – Officer Attrition Rates by YOS Groups and Gender (2011-2015)

Source: Department of National Defence, Annual Report on Reg F Attrition, 35.

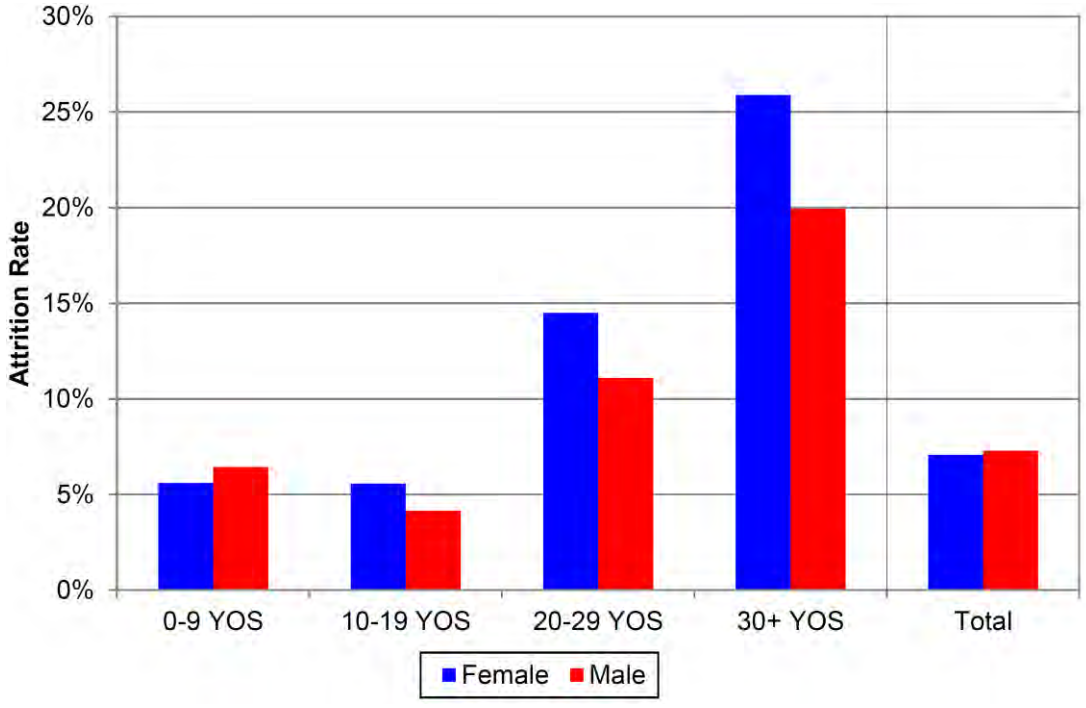

Figure 2.13 – NCM Attrition Rates by YOS Groups and Gender (2011-2015)

Source: Department of National Defence, Annual Report on Reg F Attrition, 35.

Inspecting these results more closely by looking at weighted average attrition rates by YOS without using groupings reinforces the previous results. Figures 2.14 and 2.15 show that for officers, attrition is similar for males and females between 0 and 15 YOS and generally higher after that, while for NCMs, the attrition rate amongst females is generally higher beyond 10 YOS. No possible explanation is given for these observations as the report notes that CAF Retention Survey Results do not reveal any explicit gender-based dissatisfiers.[32]

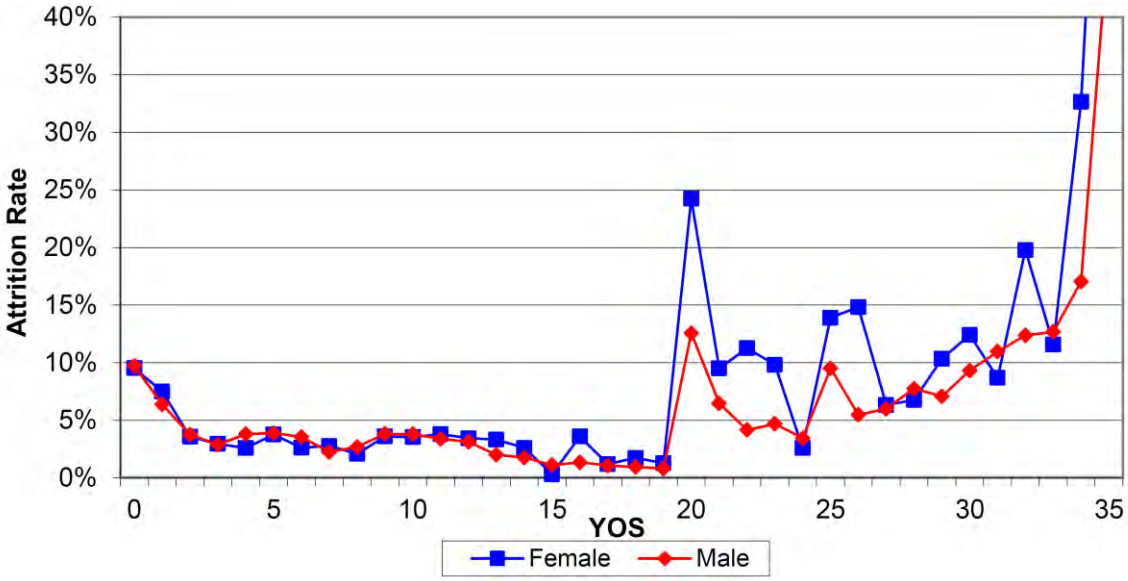

Figure 2.14 – Officer Attrition Rates by YOS and Gender (2011-2015)

Source: Department of National Defence, Annual Report on Reg F Attrition, 36.

[32] *Ibid.*, 47-50.

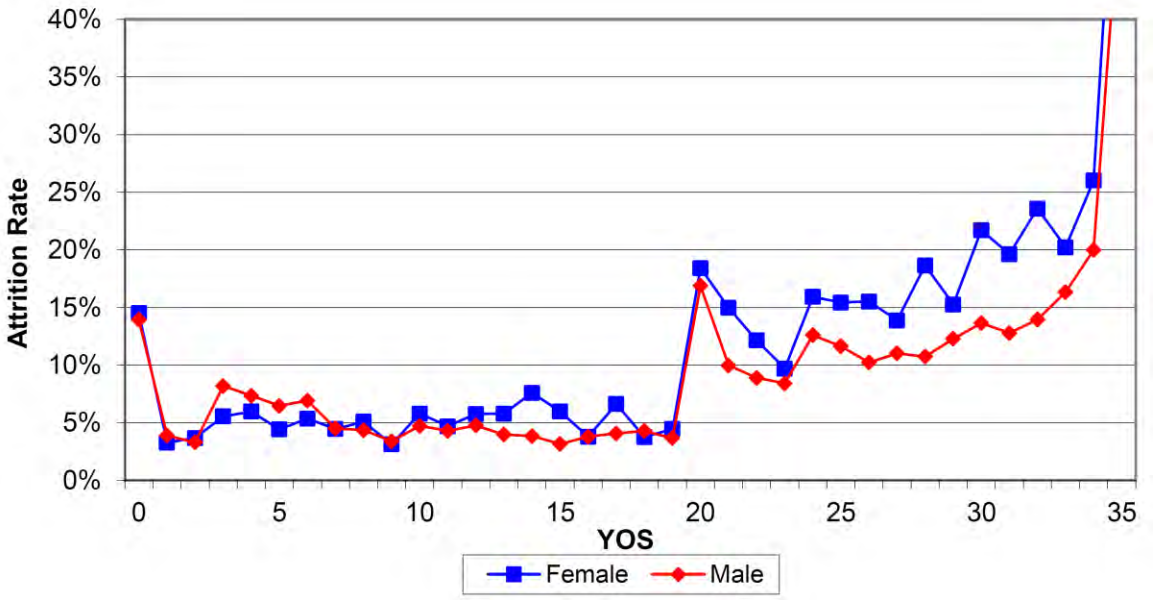

Figure 2.15 – NCM Attrition Rates by YOS and Gender (2011-2015)

Source: Department of National Defence, Annual Report on Reg F Attrition, 36.

DRDC/DGMPRA Focused Research on Attrition Topics

There has been a plethora of other attrition-focussed research conducted by DRDC and DGMPRA. Most early DND-sponsored research has been focussed on reporting and forecasting bulk attrition rates and population profiles in order to anticipate future demand for recruiting and intake plans. Okazawa laid a mathematical base for forecasting CAF attrition with a technical paper in 2007 where he introduced a more accurate mechanism for calculating attrition rates using least squares regression rather than traditional weighted average methods.[33] A methodological improvement on this

[33]Stephen Okazawa, *Measuring Attrition Rates and Forecasting Attrition Volume,* DRDC-CORA-TM-2007-02 (Ottawa: Centre for Operational Research and Analysis, January 2007): v, http://cradpdf.drdc.gc.ca/PDFS/unc66/p527519.pdf.

approach was introduced by Fang and Okazawa in 2008, but again, focusses on bulk attrition forecasting based on historical rates and YOS.[34] While this research provides useful measurement and forecasting for strategic planning, it does not identify individual risk factors outside of YOS nor does it provide individual predictions on a specific member's likelihood to release.

Otis and Straver published a more comprehensive review of CAF retention and attrition studies in 2008. Of interest to this study is their identification of the CAF Attrition Monitoring System, which is a tool used by DGMPRA to conduct continuous monitoring of total and occupational trends for both voluntary and involuntary attrition. Of particular note is the tracking of several variables including rank, age, YOS, gender, release reason (type), and FOL, indicating that these variables merit enough consideration for inclusion in the present study.[35]

Reviewing the findings of CAF Quality of Life (QoL) Surveys, Otis and Straver noted "less satisfaction with [members'] income/standard of living and career domains . . ." with pay emerging as a significant issue, even though these findings contradicted previous research.[36] Based on this conflicting result, they concluded that ". . . it could be that increase in pay is always desirable, but is not often the reason why personnel choose to leave."[37] That said, these results indicate that pay may be a factor in voluntary attrition

---

[34]Manchun Fang and Stephen Okazawa, *Forecasting Attrition Volume: A Methodological Development*, DGMPRA TM 2009-025, (Ottawa: Defence Research and Development Canada and Director General Military Personnel Research & Analysis, December 2009), http://cradpdf.drdc-rddc.gc.ca/PDFS/unc126/p532444_A1b.pdf.

[35]Nancy Otis and Michelle Straver, *Review of Attrition and Retention Research for the Canadian Forces,* DRDC CORA 2008-030 (Ottawa: Defence Research and Development Canada Centre for Operational Research and Analysis, October 2008): 5, http://cradpdf.drdc.gc.ca/PDFS/unc78/p530400.pdf.

[36]*Ibid.,* 12.

[37]*Ibid.*

and while further details are not provided, it is possible that pay could have an effect where competition exists with higher paying jobs in the private sector. As such, CAF members with attractive leadership experience or specific skills that are in demand may be at higher risk of termination, meaning that the effects of rank and occupation should be examined. Occupation and occupational groupings are of particular interest as Otis and Straver indicate a "disproportionately large" amount of research focussed on high-demand trades such as pilots, hard sea occupations, and Military Police, although it is likely that other trades should receive similar treatment due to CAF shortages and external demand.[38]

Other miscellaneous findings from Otis and Straver's literature review are pertinent to this discussion. They identified that exit surveys administered between 2005 and 2007 showed that postings and deficient CAF career management caused dissatisfaction.[39] Variables related to these areas are difficult to extract from data in HRMS but it is considered that TIR, geolocation, or similar variables may be useful proxies, albeit they possess limitations in that changes in these variables are not tracked by HRMS over time as they would be in a proper panel data set. Otis and Straver also discuss language-related attrition factors, where survey findings have indicated dissatisfaction with access to second language training.[40] It is not indicated if this dissatisfaction occurs amongst both Francophone and Anglophone populations and while this is not necessarily reflected in the second language population and attrition behaviour

[38]*Ibid.,* 26.
[39]*Ibid.,* 14.
[40]*Ibid.,* 15.

discussed previously, it merits further analysis considering the lack of research on language-related attrition.[41]

Otis and Straver also presented past research on retention factors pertaining to psychological, social, and economic areas as they relate to organizational commitment. These factors are outside of the scope of the current study, but are complimented by demographics-based research since these factors are likely shaped by what different groups perceive to be as workplace satisfiers and dissatisfiers. Ultimately, review of gender-based research is the most pertinent here. Summarizing the similarities and differences for each gender, Otis and Straver conclude that "while men and women cited many of the same reasons for leaving the CAF, women tended to cite social or familial reasons more often, while men tended to cite work-related reasons more often."[42] As such, demographic-based analysis for female attrition is appropriate and may reveal, at least partially, such social or familial reasons. For example, the number of dependent children and proximity to a family services provided at major Canadian Forces Bases (CFB) may function as predictors.

Sokri, a DRDC researcher, conducted a socio-economic analysis of attrition for NCMs using survival analysis that provides some consideration and context. He found that "higher education, being male, and being single were associated with higher risk of attrition whereas older age at hire and higher rank reduce attrition."[43] Some of these results may be counter-intuitive but reinforce some previous observations; given the

---

[41]Otis and Straver, *Review of Attrition and Retention Research for the Canadian Forces . . .*, 22

[42]*Ibid.,* 21.

[43] Abderrhamane Sokri, *A Socio-economic Analysis of Military Attrition: The Case of Non-Commissioned Members of the Canadian Armed Forces*, DRDC CORA TM 2013-088 (Ottawa: Defence Research and Development Canada – Centre for Operational Research, June 2013): iv, http://cradpdf.drdc-rddc.gc.ca/PDFS/unc263/p537714_A1b.pdf.

stress of service on families it would be conceived that single members and males would experience lower attrition and one would think that higher ranks would have other economic opportunities open to them. That said, this reveals the possibility that other underlying factors may moderate expected effects. For example, attrition in higher ranks may be moderated by the impact of organizational commitment. Sokri also found evidence indicating that, expectedly, lower unemployment in the Canadian economy resulted in a higher attrition hazard.[44] While outside of the scope of this study, it may provide an explanation for periods of higher attrition rates and volumes.

Lee, McCreary, and Villeneuve applied a logistic regression analysis to study attrition amongst CAF recruits during basic training. They found that "recruits who were released from basic training were more likely to be NCM candidates, to have dependents, and to have had an annual household income of less than $20,000 before basic training . . . [while] in contrast to previous findings, age and sex were found not to be associated with basic training attrition."[45] While they indicate that their results on age and sex are contradictory, they are in keeping with what is observed in terms of first-year attrition discussed previously in Figures 2.14 and 2.15.

Penney provides some research focussed on public servant attrition in DND, including a study of attrition in the Assistant Deputy Minister – Materiel Group (ADM(Mat)) using survival analysis techniques (2016) and a model of voluntary attrition of civilians in DND (2013) using random effects methods. While not military-focussed, these studies can be examined to provide some insights into military attrition assuming

---

[44] *Ibid.*

[45] Jennifer E.C. Lee, Donald R. McCreary, and Martin Villeneuve, "Prospective Multifactorial Analysis of Canadian Forces Basic Training Attrition," *Military Medicine* 176, no. 7 (July 2011): 781, https://search-proquest-com.cfc.idm.oclc.org/docview/876042799.

some similarity between the military and public service. In the 2013 study of attrition amongst all DND civilians, Penney identified a decreased likelihood of attrition for higher Pensionable Years of Service, changes in command / leadership, and residing in the provinces of Quebec, Newfoundland, or Saskatchewan.[46] Conversely, an increased likelihood of attrition was associated with French speakers, those in particular occupational groups, those residing in Alberta, and those with higher salaries.[47] Penney describes this last result as unexpected but explainable by the "hypothesis that salary is correlated with ability, and higher ability renders individuals either more attractive to, or more likely to accept, outside employment opportunities".[48]

Penney's 2016 study builds upon his previous work, using survival analysis techniques to look at public servants in the Procurement Group (PG) occupation, one of the previously identified higher-risk groups. He found that higher age and pensionable years of service (PYOS) at the start of an individual's job increased likelihood to attrite while this likelihood was actually slightly less for females. Interestingly, Penney also found a significant interaction between starting PYOS and being female: "the interaction between starting PYOS and female indicates that the positive effect of starting PYOS on the attrition hazard is slightly more pronounced for women than it is for men . . . [meaning that] females who begin their PG careers with less experience are less likely to attrite, while those who begin with more experience are more likely to attrite."[49] While

---

[46]Christopher Penney, *Modelling Voluntary Attrition of Civilian Personnel in the Department of National Defence,* DRDC-CORA-CR-2013-063 (Ottawa: Centre for Operational Research and Analysis, April 2013).

[47]*Ibid.*

[48]*Ibid.,* iv.

[49]Christopher E. Penney, *A Survival Analysis of ADM (Materiel) Workforce Attrition*, DRDC-RDDC-2016-R162 (Ottawa: Defence Research and Development Canada – Centre for Operational Research, October 2016): 12, http://cradpdf.drdc-rddc.gc.ca/PDFS/unc247/p804591_A1b.pdf.

this is interesting in that it identifies a possible implication of Time in Occupation (TIO), it also highlights the importance of looking at interactions between variables and how they may improve predictive modelling.

Going further in secondary analysis, Penney identifies a relationship between level (or rank in the military), whereby "increasing in level is associated with an increasingly greater proportional downward shift in the hazard function", whereas higher salaries produce a "lower proportional likelihood of attrition over time."[50] Penney also found that Francophones were "more likely than others to leave for other employment opportunities", reflecting agreement with other past research.[51] Penney's phrasing of the latter effect of FOL on attrition is interesting, as it may provide some explanation for previous results. Issues with access to second-language training and associated lower second-language testing results amongst Anglophones could be seen as a dissatisfier or overall, lower opportunities for advancement within the military; however, second-language requirements are common across all areas of the government and could be attractive to private sector employers. As a result, this reinforces the hypothesis that more opportunities may exist for Francophones outside of the military, which may cause increased attrition.

More recently, DGMPRA has looked at using ML techniques to look at individual attrition forecasting. Calitoiu has recently submitted draft research targeting both medical and voluntary attrition for the entire population of the CAF from 2006 to 2015. While the focus of our study is on voluntary attrition, both studies provide pertinent details, particularly considering their use of logistic regression. Calitoiu identifies three

[50] *Ibid.,* 17.
[51] *Ibid.*

components to attrition based on predictability and degree of change over time, including: "predictable trended, predictable untrended, and unpredictable (rare events)."[52] In this regard, voluntary and medical attrition are posited as predictable and trended, thus, appropriate for causal forecasting using predictive ML techniques such as logistic regression.[53]

While previous methods have used least squares regression or weighted averages to calculate bulk attrition rates, Calitoiu identifies that these approaches are not always appropriate for studying attrition since, using voluntary attrition as an example, the decision to terminate follows the Bernoulli Distribution.[54] That is to say, the employment status (i.e. "terminated") for an individual can be characterized as having probability p of being true and probability $1 - p$ of being false (i.e. "active"). As a result, ML models that focus on predicting binary outcomes, such as logit or tree-based modelling, are appropriate for these types of problems.

For his voluntary attrition model, Calitoiu explored 47 variables collected in HRMS but deselected some due to issues with "redundancy, consistency, and variability", resulting in 20 variables as candidates for the logit model.[55] These variables included: Age, YOS, TIR, TIO, Rank, Gender, Engagement Code, Qualification Code, Highest Degree of Education, Marital Status, Affiliation Code, Environment, Manning, First Official Language, and Second Language Proficiency (SLP). Categorical variables from the above were transformed before model training using simple dummy coding

---

[52]Dragos Calitoiu, *Logistic Regression Model for Medical Attrition,* DRDC-RDDC-2016-RXXX [DRAFT] (Ottawa: Director General Military Personnel Research and Analysis, November 2016): 3.

[53]*Ibid.,* 3, 4.

[54]*Ibid.,* 5.

[55]Dragos Calitoiu, *Logistic Regression Model for Voluntary Attrition,* DRDC-RDDC-2016-RXXX [DRAFT] (Ottawa: Director General Military Personnel Research and Analysis, November 2016): 1.

based on whether attrition in a category was higher than the overall attrition rate. The models were built using stepwise feature selection with χ^2 goodness-of-fit providing the stepwise parameter. These study design choices were not used in the present study but are highlighted for completeness.

Once the model was fit, Calitoiu used decile analysis to evaluate performance, whereby the probabilities of attrition were grouped into deciles and then evaluated based on what percentage of actual attrition fell into each decile. His model was able to effectively identify high-risk individuals, with more than 31.83% of actual voluntary releases falling into the top decile.[56] To increase model accuracy, he applied Box-Cox transformations to numerical variables. While not required as pre-processing steps in logistic regression as it is not based on the same assumptions as that for linear regression (particularly normality), Box-Cox transformations were applied in order to yield "a more stable and powerful solution."[57] Applying these transformations improved the performance as measured by the percentage of releases in the top decile from 31.83% to 33.25%.[58] While not all of the techniques and ML design decisions used by Calitoiu were adopted, this research nonetheless provides insight into initial variables for analysis and post-prediction model adjustments as well as providing a useful comparative study.

Research External to the CAF

In addition to the CAF-specific research discussed thus far, it is important to also look externally to see if other research, particularly that from US military sources, can

---

[56]*Ibid.,* 7.

[57]Dragos Calitoiu, *Logistic Regression Model for Medical Attrition,* DRDC-RDDC-2016-RXXX [DRAFT] (Ottawa: Director General Military Personnel Research and Analysis, November 2016): 17.

[58]Dragos Calitoiu, *Logistic Regression Model for Voluntary Attrition,* DRDC-RDDC-2016-RXXX [DRAFT] (Ottawa: Director General Military Personnel Research and Analysis, November 2016): 1.

apply some context or considerations for the present study of CAF female attrition. Terse results from such studies are presented in the following paragraphs and include an overview of gender-based attrition from military, public service, and other sectors.

Davis, a researcher focused on female attrition in the CAF, is included in this section due to her contribution to a joint Canada-US publication on gender-based attrition outside of that performed by DRDC and DGMPRA. Using focus-groups of women that had terminated from the CAF, she found that "nine of the twenty-three women interviewed identified family-related circumstances which were directly linked to their exits from the organization. . . [with six of those leaving] as a result of circumstances affecting family stability, primarily in terms of geographical location."[59] This is a very small sample, but nonetheless provides some important direct, post-termination feedback. Further to these results, Davis also found that "without exception, the servicewomen who left for family reasons had children, or were expecting a child in the near future" indicating the importance of this variable in studying gender-based attrition as well as a limitation in data collection, since expecting mothers are not identified in data and thus, may provide a hidden effect amongst those in the "without children" category.[60]

Pollack et al, in studying attrition in US Marine Corps Recruits, found that age, race/ethnicity, and education significantly predicted attrition, with higher discharge rates for older, non-Hispanic, and more educated recruits.[61] Following this research, they

---

[59]Karen D. Davis, "Understanding Women's Exit from the Canadian Forces: Implications for Integration?," in *Wives and Warriors: Women and the Military in the United States and Canada,* ed. Laurie Weinstein and Christie C. White (Westport, CT: Bergin & Garvey, 1997): 191.

[60]*Ibid.*

[61] Lance M. Pollack, Cherrie B. Boyer, Kelli Betsinger, and Mary-Ann Shafer, "Predictors of One-Year Attrition in Female Marine Corps Recruits," *Military Medicine* 174, no. 4 (April 2009): 386, https://search.proquest.com/docview/217051260?accountid=9867.

looked at graduates of recruit training after their first YOS, finding that female Marines who were married at intake had the highest rate of attrition of any group examined.[62] A separate study by Wolfe et al looked at the impact of gender and prior experience with trauma on Marine Corps recruits. They found that attrition in this group was higher for females than males and was particularly higher amongst females that reported a previous interpersonal trauma. This shows that gender as a variable may be significant but also shows the limitations of strictly demographic studies, since the existence of hidden moderator variables may impact the output.[63]

In a doctoral thesis, Hayes studied whether or not demographics could predict turnover intention amongst full-time employees in Texas, finding that "a statistically significant relationship existed between an employee's age, income and turnover intention . . . [but] did not exist between an employee's education, gender or, length of tenure and turnover intention."[64] However, she notes: "It is possible that the predictor variables of education, gender, and length of tenure could have significant effects if a researcher replicated the study using different educational groupings, or using moderating variables."[65] For example, 84% of respondents in her sample had attended college, meaning that "results of a future study may have different indications if there is equal dispersion among educational groups."[66] The important consideration here is that the specific demographics of the population being studied and the distributions of component

---

[62] *Ibid*.

[63] Jessica Wolfe et al, "Gender and Trauma as Predictors of Military Attrition: A Study of Marine Corps Recruits," *Military Medicine* 170, no. 12 (December 2005): 1039, https://search.proquest.com/docview/217065022?accountid=9867.

[64] Tracy Hayes, "Demographic Characteristics Predicting Employee Turnover Intentions," Order No. 3728489, Walden University, 2015, 96, https://search-proquest-com.cfc.idm.oclc.org/docview/1734042628?accountid=9867.

[65] *Ibid.,* 95.

[66] *Ibid.*

variables can influence results; it would be extremely unlikely to find a universally applicable model that does not change over time or based on the introduction of new information.

Moynihan and Landuyt built a logistic regression model to study turnover intention amongst public servants in US State Government, looking at individual and organizational factors as explanatory variables of attrition. Again, drawing some parallels between employment in the public service and the military, their findings can further add to our domain knowledge. With respect to gender, they found that women were less likely to state turnover intention with an odds ratio of .770 and decrease in probability of 2.7% for this group, which contradicts some previous research they reviewed (although, these findings are consistent with some of the research discussed previously).[67] While direct gender-based comparisons will not be made in the current study, this provides another indicator of female attrition behaviours in general, as these results are explained by the authors to likely result from the particular attractiveness of public service employment for women, a sector in which they are over-represented.[68]

Increased age and length of service are shown by the authors to decrease turnover intention. They attribute this to the effects of the life stability hypothesis: "employees who have reached a certain measure of stability in their life, and who have pressing economic and familial concerns, are less likely to seek the changes brought about by seeking a new job", which is also consistent with some human capital theories regarding the difficulty in employment changes for tenured employees given their firm-specific

---

[67] Donald P. Moynihan and Noel Landuyt, "Explaining Turnover Intention in State Government: Examining the Roles of Gender, Life Cycle, and Loyalty," *Review of Public Personnel Administration* 28, no. 2 (2008): 130, https://doi.org/10.1177/0734371X08315771.

[68] *Ibid.,* 132.

knowledge and skills.[69] Not surprisingly, geographic stability was also found to decrease turnover intention, although at lower levels of statistical significance in the model coefficient. While the significance of the coefficient raises questions about this finding, in general, it supports the hypothesis that staying in one area "indicates reluctance to change, which tends to limit the search for new employment opportunities and the intention to exit the workplace."[70] Economic and familial constraints form a part of these stability factors; however, Monahan and Landuyt found that while family-friendly policies were associated with lower turnover intention, the statistical significance of the model coefficient was also lower (p-value greater than 5% but less than 10%).[71]

Moynihan and Landuyt also examined the impacts of public-private pay disparity. They found that the impact of perceived pay disparity was particularly pronounced at higher rank and pay levels, finding a direct relationship between salary and increased turnover intention, which also contradicted some previous public sector research.[72] The authors hypothesize that this effect is possibly explained in how pay perceptions are built. They posit that public sector employees are more likely to perceive fair pay relative to the private sector, noting that higher levels of public sector pay correspond to those with the high seniority or experience that are in-demand externally.[73]

While the effects of pension eligibility are not discussed, another possible explanation for this effect is found in other findings in the research. The authors found that increased education, likely associated with higher rank levels and salary, had a strong

---

[69]*Ibid.,* 131.
[70]*Ibid.*
[71]*Ibid.,* 134.
[72]*Ibid.,* 130.
[73]Monahan and Landuyt, "Explaining Turnover Intention in State Government: Examining the Roles of Gender, Life Cycle, and Loyalty" . . ., 133.

impact on increased turnover intention. Comparing those with a college degree or higher to those with less educational attainment showed a concomitant increase in predicted probability of turnover of 5%, the strongest effect seen amongst all variables.[74] Given the consistency of this result with past findings, it may provide some generalities in other sectors. Contrarily, the impact of merit promotions had the opposite effect to salary and education. The authors observe that the receipt of a merit promotion with the past two years resulted in a 4% decrease in predicted probability of turnover versus those who had not been promoted, indicating that advancement and organizational commitment may moderate some of their other observed results.[75]

Pitts, Marvel, and Fernendez conducted a study with similar aims for US federal employees. Most pertinent from their examination of demographic variables were the effects of age and agency tenure on turnover intentions. For age, they found that "federal government employees in the middle age ranges are more likely than employees in the youngest age range to seek new employment opportunities within the federal government but less likely to leave the federal government for an outside position."[76] This may be the result of their desire to search for new experience while maintaining some stability or could be influenced by pension considerations. Longer agency tenure, on the other hand, showed a consistent negative effect on turnover consistent with the life stability hypothesis, making "an individual less likely to intend to leave his or her agency and less

[74] *Ibid.,* 130, 133.
[75] *Ibid.,* 134.
[76] David Pitts, John Marvel, and Sergio Fernandez, "So Hard to Say Goodbye? Turnover Intention among U.S. Federal Employees," Public Administration Review 71, no. 5 (2011): 756, http://onlinelibrary.wiley.com.cfc.idm.oclc.org/doi/10.1111/j.1540-6210.2011.02414.x/abstract.

likely to intend to leave the federal government entirely."[77] As they note, this provides policy options for the organization to "consider ways to *either* exploit the likelihood of experienced workers to remain in the organization *or* better incentivize inexperienced employees to stay."[78]

These results are consistent with Ryan, Healy, and Sullivan's findings showing that increased tenure had a negative effect on turnover intention amongst academics; however, they also look at other factors that are relevant to the present study.[79] Amongst demographic and occupational variables, their logit model identified that being in a "soft-pure" discipline and having "higher research productivity" were predictors of leaving for another institution while being in a "hard-applied" discipline, not having a spouse or partner, and familial considerations functioned as predictors of leaving academia altogether.[80] While these results indicate that familial and marital status are expectedly significant considerations in turnover, it also reinforces the notion that occupation type, relative to external opportunities, plays a role in turnover decisions and could be applicable in a military context, particularly for those in hard technical fields.

Commentary and Conclusions on Literature Review

This literature review has presented a plethora of statistics, research, and reports pertaining to attrition. It is immediately apparent that while some findings are consistent, there were also many instances where different studies produced conflicting results. In general, we assume that this is a product of the underlying uniqueness of each study

---

[77] *Ibid.*
[78] *Ibid.*
[79] John F. Ryan, Richard Healy, and Jason Sullivan, "Oh, Won't You Stay? Predictors of Faculty Intent to Leave a Public Research University," *Higher Education* 63, no. 4 (April 2012): 421, http://dx.doi.org/10.1007/s10734-011-9448-5.
[80] *Ibid.*

design, variables examined, and the data sets themselves. It is quite possible and even expected that a variable seen to be significant in univariate analysis would not have the same effect in a multivariate analysis based on the effects of other data points on the output variable. Ultimately, we can be confident that based on the theoretical foundation provided in past research, there is merit in studying the predictive relationships between associated demographic and occupational variables with the likelihood of termination for the population being studied.

Also of interest from the literature review is the need to look at mediating and moderating variables. For example, the Moynihan and Landuyt study indicated that those with increased age and length of service have a decreased likelihood of turnover, which they explained as being reflective of a desire for life stability. As such, a measure of life stability could be seen as a mediator for age and tenure (and vice versa). In the case of this study, where possible, demographic variables that mediate the relationship between unobserved variables such as life stability, organizational commitment, and workplace satisfaction will be used, while noting that the intent of the present study is less about the explanatory power of the included variables and more about the performance of the model as a whole in accurately predicting terminations.

The effect of interactions amongst variables was not seen extensively in the literature review; however, their existence is implied in several studies. Using the example in the previous paragraph, we also saw that the higher salary associated with higher levels in government actually increased turnover likelihood, which was explained as likely being the result of pay disparity when compared against opportunities in the private sector. Thus, it is quite possible to see higher aged individuals with longer service

have increased turnover likelihood contrary to previous results under the life stability hypothesis. This implies that there is a potential interaction between these variables where the effect of age and tenure may be moderated by salary or level. As a result, the present study will look at interactions where applicable.

METHODOLOGY

Overview

Two datasets were provided from HRMS, which constituted a sampling of active female members of the CAF for calendar years from 2004 to 2016 (labelled *active;* 9,424 records) and those that voluntarily terminated during the same period (labelled *terminated;* 4, 532 records). Variable names used were the default column names in HRMS. At a total pre-cleaning sample size of 13, 956 records, this sample was deemed to be appropriate for both training and testing purposes; however, the sample was later trimmed to 13, 524 records due to issues in some of the variables, which will be discussed in subsequent sections. Based on the outcome of the literature review, author domain knowledge, and available database fields, attributes were selected that would potentially be direct predictors of terminations or be used to calculate other variables that would function as predictors. While final variable selection is discussed in the next section, in general, requested attributes were categorized as follows:

- Individual Demographic Factors: birth date, marital status, language ability, education, geolocation, and number of dependent children.
- Occupational Factors: join date, rank, rank seniority date, occupation, occupation entrance date, assigned CAF environment, prior service, contract type, termination date (as applicable), qualifications, and posting date.

The R programming language was used for all data processing and analysis. The following sections outline the steps taken for data cleaning, handling missing data, model training, testing, and analysis. Where applicable, background theory is provided on the R packages and ML algorithms used for the study.

Data Cleaning

An initial check was performed to ensure that the data in the two data sets was consistent such that a master data set could be created. This check revealed some minor inconsistencies that were easily rectified; however, it also revealed two major issues. First, the data set containing terminated personnel did not contain qualification levels and posting dates. Thus, level of qualification and the time spent in a specific job or location were precluded as variables of study. Second, it had been desired to consider language-related variables such as FOL and SLP. While FOL was found to be tracked consistently, the results of Second Official Language (SOL) testing were inconsistent and unworkable in the data set of terminated personnel; columns were not consistently named to reflect the results of reading, writing, and oral testing and new sets of columns were created each time a member was tested as opposed to recording the most recent results. As a result, SOL testing results were excluded from study. The active and terminated data sets were then combined into a single data set and checked for variable type inconsistencies, duplicate records, and forbidden numerical values (i.e. < 0) with only minor issues that were easily rectified. Treatment of each variable type is discussed in the following sub-sections.

Categorical Variables

An exhaustive review of categorical variables was conducted. R refers to categorical variables as *factor* variables with each category designated as a *factor level*, so nomenclature will be used interchangeably but will preference use of R's nomenclature for consistency. In general, each factor variable was examined for inconsistencies and the distributions of the frequency of occurrence in each level were

observed. Where needed, levels were recoded to new values to rectify inconsistencies or to reduce the amount of levels. Regarding the latter, some factor variables contained many levels, which were foreseen to potentially affect the performance of model building while also adding complexity that could significantly impact computation time. Where applicable, factor levels were rationalized using business logic to reduce the number of levels into groupings that made sense from a domain knowledge perspective.

Marital Status. Marital Status was found to contain some unneeded levels that duplicated the general meaning of a specific category. For example, there were three levels corresponding to married: Married, Married Regular Forces, and Married Reserve Force. As these granular delineations provided no additional meaning, they were generalized by recoding the granular definitions to the parent category. It was also noted that there were 174 values corresponding to an *unknown* category. There may have been a reason that these were coded in this manner such as a refusal to provide a marital status when the data was collected. As a result, these values were considered to be Missing Not at Random (MNAR), and were treated as a separate category.

The distribution of Marital Status after values were recoded is found at figure 3.1. Married, Single, and Common-Law levels were observed to be the most common, accounting for ~92% of the data set.

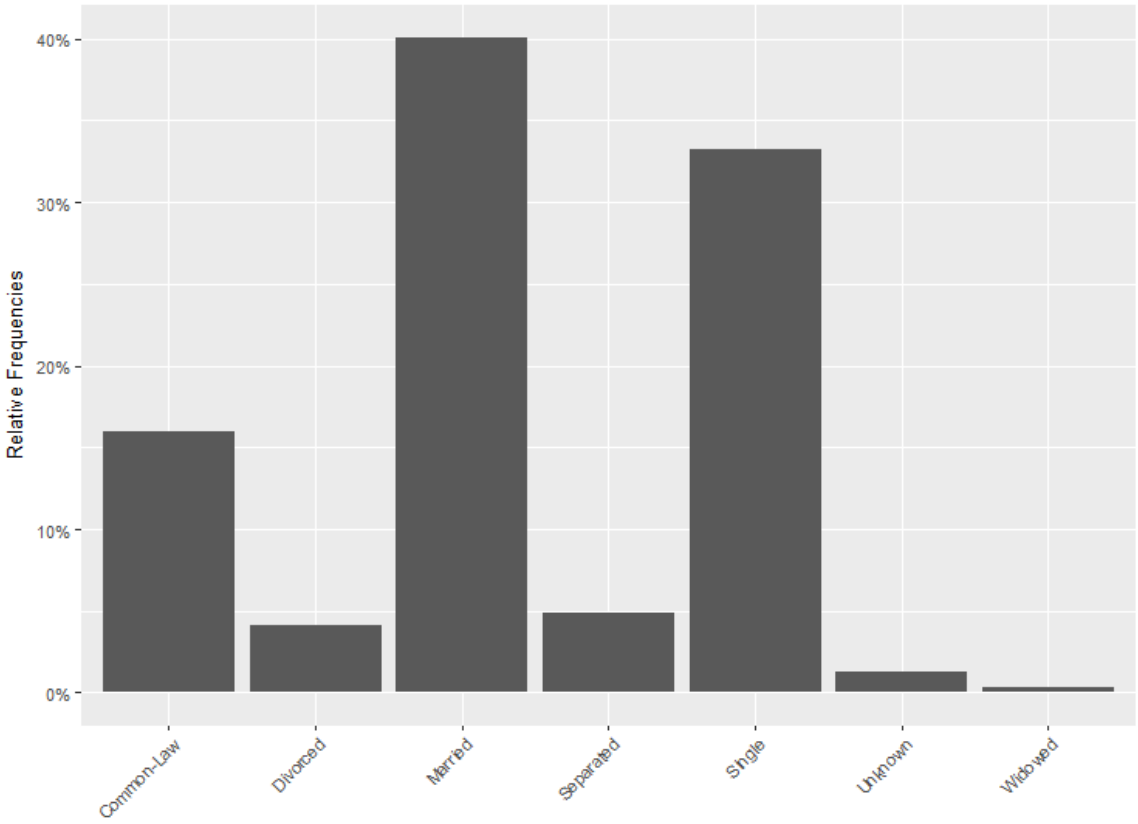

Figure 3.1 – Distribution of Marital Status Variable

Employment Status. Employment Status contained an additional Leave of Absence category that corresponded to active personnel. As the intent of the study was a binomial classification of active and terminated personnel, this was recoded to the active category. The distribution of this variable corresponded to the number of records in each component data set.

Environment. There were no inconsistencies in this variable, which represents the member's environment or Dress Elemental Uniform (DEU) affiliation. It had been considered to recode this variable to reflect the Canadian Forces Occupational Authority (CFOA) responsible for managing a member's specific occupation since joint or *purple* trades are managed by Asst CMP; however, this recoding was seen to cause issues.

Also, there are some joint trades managed by Asst CMP that tend to work in both joint and specific environmental postings such as Land or Air Logistics. As a result, this variable was not changed. From the distribution (Figure 3.2) most records came from the Land and Air environments.

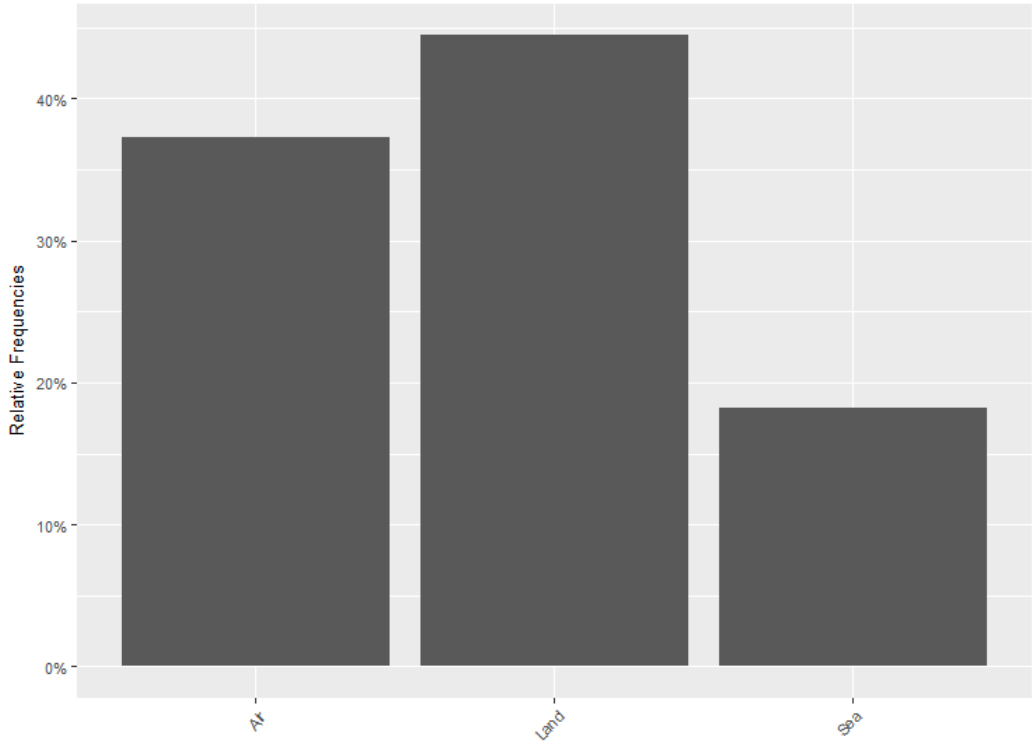

Figure 3.2 – Distribution of Environment Variable

Rank. Levels of this variable separated Navy and Army/Airforce ranks that are actually equivalent. These values could have been recoded to standardized NATO ranks; however, this schema would still result in many levels. As a result, they were recoded into standard Rank Groups used by the CAF.[81] Given the infrequent occurrence of General/Flag Officers, these were included in the Senior Officer category. From the distribution (Figure 3.3), most records corresponded to Junior NCM ranks, which made

[81]Government of Canada, "Ranks and appointment," last modified 19 March 2018, https://www.canada.ca/en/department-national-defence/services/military-history/history-heritage/insignia-flags/ranks/rank-appointment-insignia.html.

up over 50% of the data set. Note that the delineation between officers and NCMs was not included since it can be completely derived from rank group.

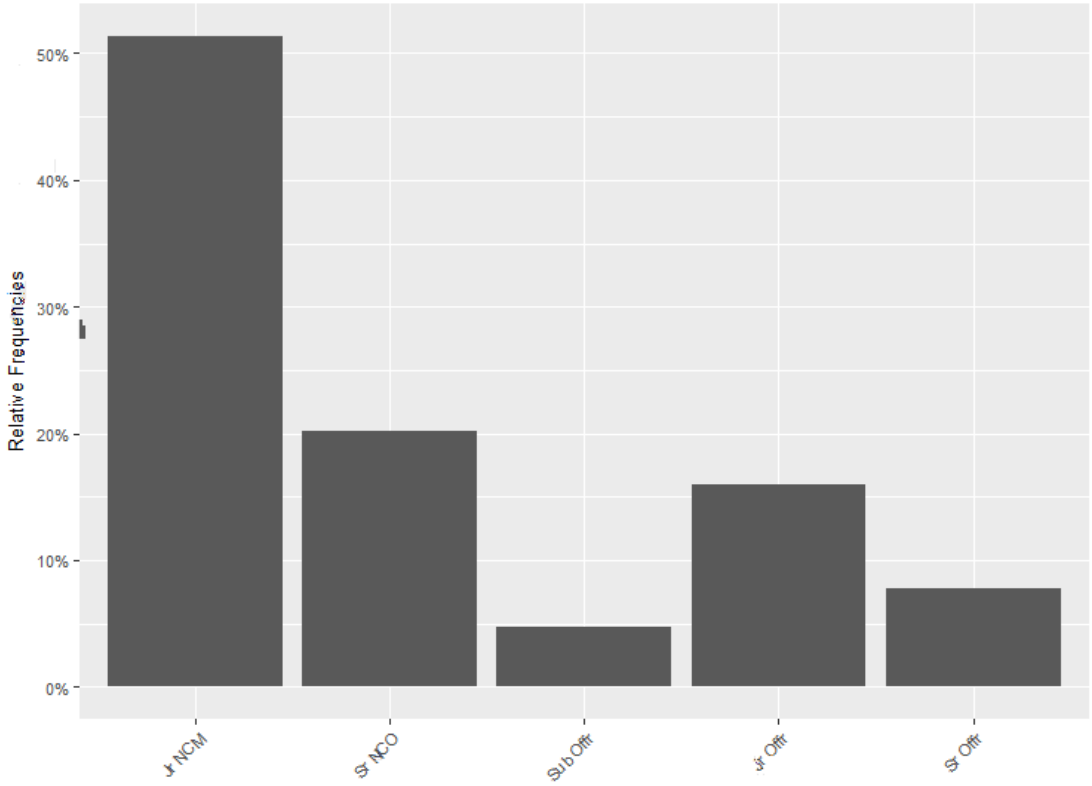

Figure 3.3 – Distribution of Rank Group Variable

Military Occupational Structure Identification Code (MOSID). Given the amount of levels of MOSID, these were recoded to reflect the career field groupings in the CAF Annual Report on Personnel 15/16 (see Appendix 1).[82] There were some issues pertaining to the existence of levels corresponding to MOSID that no longer existed, were restructured, or reflected new trades, so these issues were rectified, leading to the distribution at Figure 3.4. The Operations Support (OS) level heavily dominates the

[82]Department of Defence, *Annual Report on Regular Force Personnel 2015/2016* (Ottawa: Director General Military Personnel Research and Analysis, October 2016), B44-46.

other categories, noting that it encompasses logistics and clerical trades where females are heavily represented.

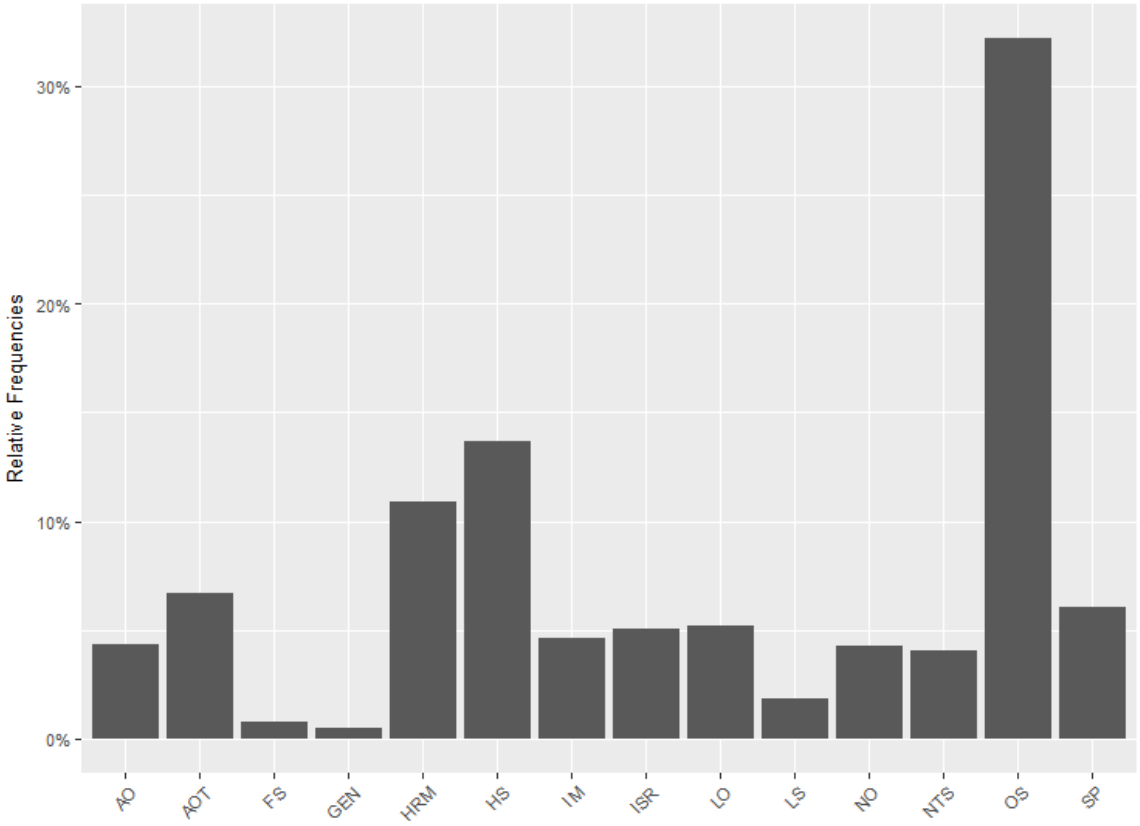

Figure 3.4 – Distribution of Occupation Group Variable

Release Item. This variable was not used in analysis as the intent was to study all voluntary (CAF category 4) releases; however, it was examined to see if it highlighted any issues in the data. In this regard, there was a serious discrepancy noted with the presence of releases for Active personnel, indicating the existence of prior service. The issue encountered was that HRMS only records the most recent termination date and does not differentiate different hire dates when there is a break in service. In many cases, the hire date was very far in the past and while HRMS records former service, it was not clear upon inspection exactly how to break up these dates in order to properly reflect

YOS. As a result, data in these records was considered MNAR and since it could heavily impact results, they were removed from the data set. Dropping these records diminished the sample to 13,528 records.

Highest Education Level. There were 11, 989 records that were labelled as "Not Indicated". As a result, this variable was not included in the model but kept in the data set for possible future use.

Geographic Location. The data in this variable was consistent; however, it was comprised of many levels including many that had a single occurrence. To rectify this, the variable was recoded according to the following business rules:

- Values were not recoded if the location was near a major Canadian Forces Base (CFB) including the recruit training garrison at St-Jean-sur-Richelieu.
- Major metropolitan areas of Toronto, ON and Montreal, QC were also not recoded in order to reflect the significant external economic opportunity and the sizeable garrisons and reserve units in those locations.
- North West and Yukon Territories were grouped together.
- Certain cities that have Reg F units proximate or belonging to a major base were coded as part of that base (i.e. Quebec, QC was coded as Valcartier, QC).
- Members outside of major CFBs but within Canada were recoded to "Other, Canada".
- Members posted outside of Canada were recoded to "OUTCAN".

The intent of using these business rules was to provide more granularity than that provided by simply looking at province, based on the assumption that are likely intra-

province differences. The recoded distribution (Figure 3.5) shows that the National Capital Region (NCR) and St-Jean-sur-Richelieu were the most dominant levels, which is expected given the amount of personnel posted to Ottawa and the high frequency of releases that occur during recruit training. Several CFBs showed infrequent occurrences in the sample, likely corresponding to smaller bases or employment in MOSID that are uncommon amongst women while other large bases such as Petawawa, ON and Halifax, NS had expectedly large concentrations of females.

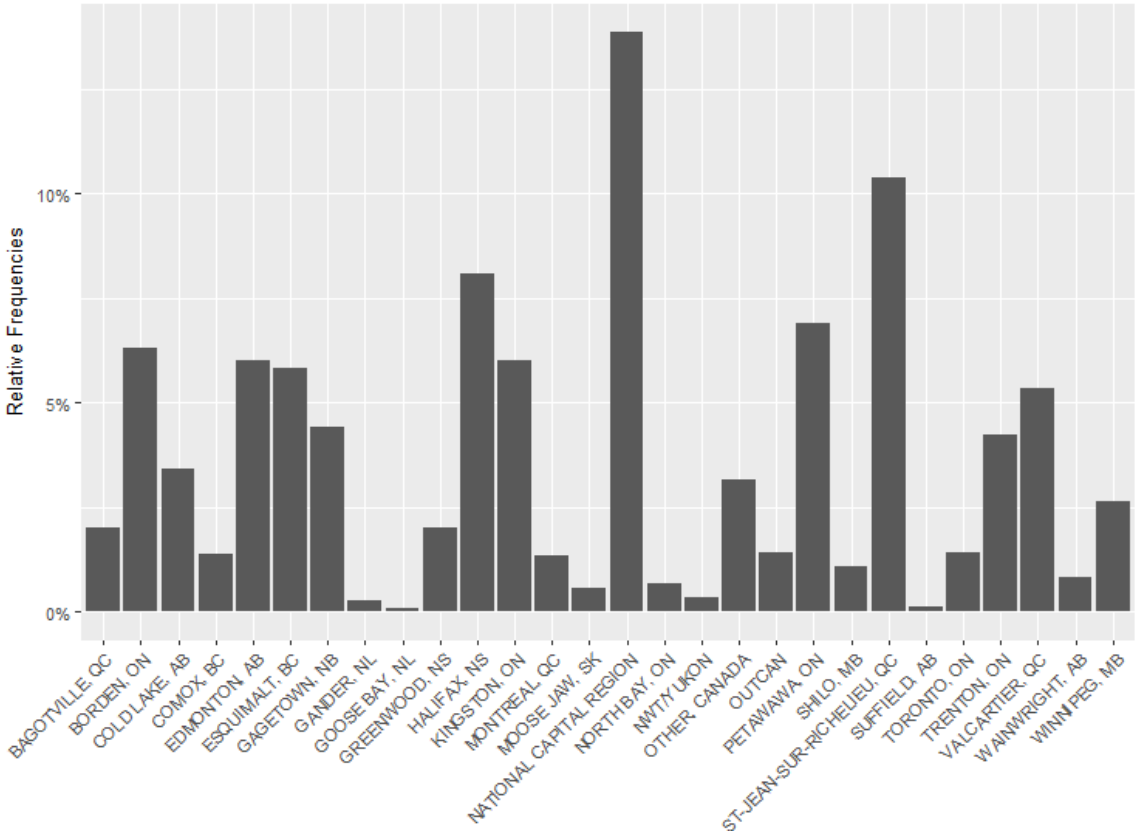

Figure 3.5 – Distribution of Geographic Location Variable

First Official Language. There were no issues with this variable. The distribution (Figure 3.6) shows that ~72% of the sample corresponded to native English speakers.

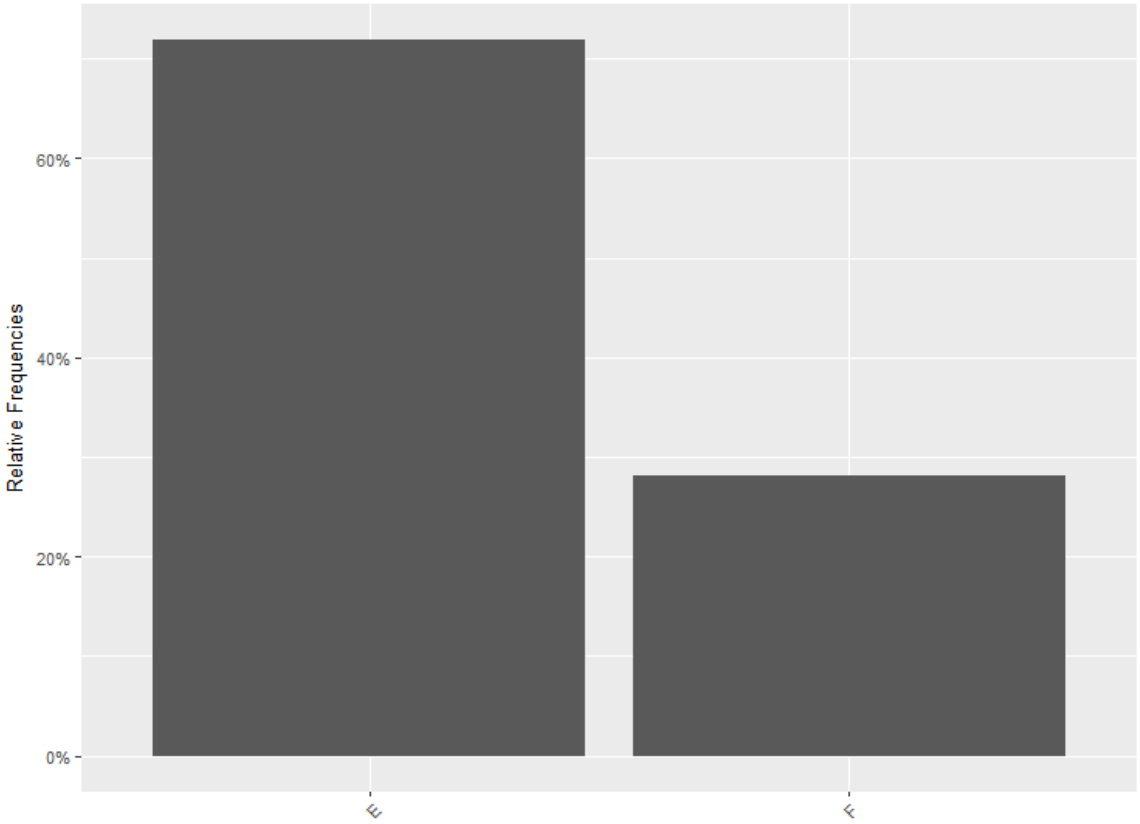

Figure 3.6 – Distribution of First Official Language Variable

Contract Type. There were no major issues with this variable; however, there were a large amount of records that were listed with old contract types that no longer exist due to TOS changes enacted in 2005. Thus, while these are unlikely to be errors as they either represent a grandfathered contract or the contract under which the member was serving at time of release, this variable will change over time as the entire population continues to move to the new structure. The distribution (Figure 3.7) shows a large amount of personnel under the Variable Initial Engagement contract, which is the typical contract for new NCMs in the CAF. This would be expected given the amount of junior

ranked personnel in the sample and the amount of releases that occur during recruit training. The distribution also shows many levels where there are few occurrences.

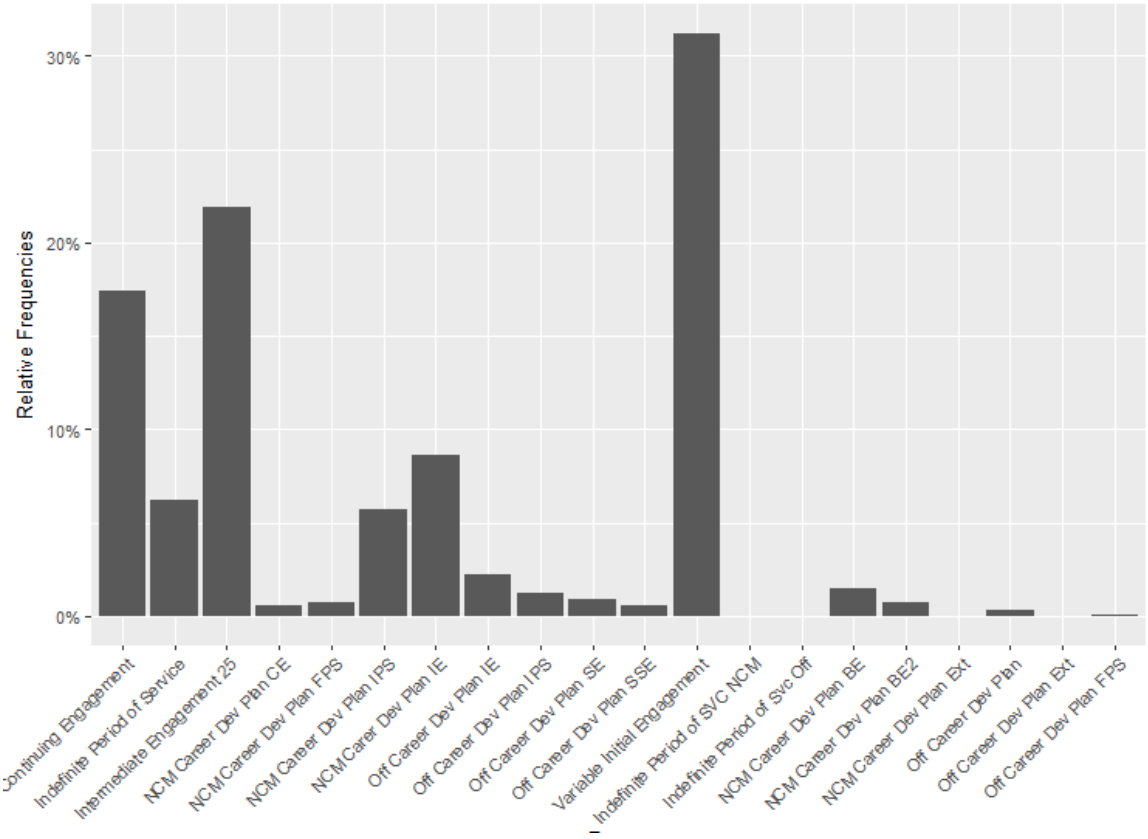

Figure 3.7 – Distribution of Contract Type Variable

Date and Numeric Variables

Date variables were used to derive numeric variables that would work better for modelling: Age, YOS, TIR, and TIO. Before calculating these numeric variables, the date variables were sanity checked and examined for possible errors. Some issues in the date variables possibly corresponding to edge cases as well as some data collection errors were observed. In some cases, these could be fixed while in others, records were deleted, which diminished the final data set to 13, 524 total records. Of note is that rank seniority

date was used to calculate TIR instead of rank effective date as there were issues in this variable for terminated personnel.

The distributions of the resulting calculated variables were examined using violin and Quantile-Quantile (QQ) plots (figures 3.8 through 3.12). The violin plot shows the kernel density estimate (shaded in yellow) for each variable, along with its interquartile range (thick black box) and median (white dot). For each variable, we see a skewed distribution with the greatest amount of observations in lower values, particularly for TIR. The YOS, TIR, and TIO variables all show a similar distribution.

The QQ plots allow for comparison of the distribution of each variable against the Normal Distribution in order to check for normality as it may influence design decisions and which R packages are used. For normal distributions, the points in the QQ plot will fall along the red reference line. It is apparent that YOS, TIR, and TIO did not follow the Normal Distribution while age shows some similarity. To confirm the latter variable's distribution, a Shapiro-Wilk test was performed, which confirmed that age did not follow the Normal Distribution ($p < 2.2 \times 10^{-16}$).[83]

[83]S. Shapiro and M. Wilk, "An Analysis of Variance Test for Normality," *Biometrika* 52, no. 3-4 (1 December 1965): 591–611, https://doi.org/10.1093/biomet/52.3-4.591.

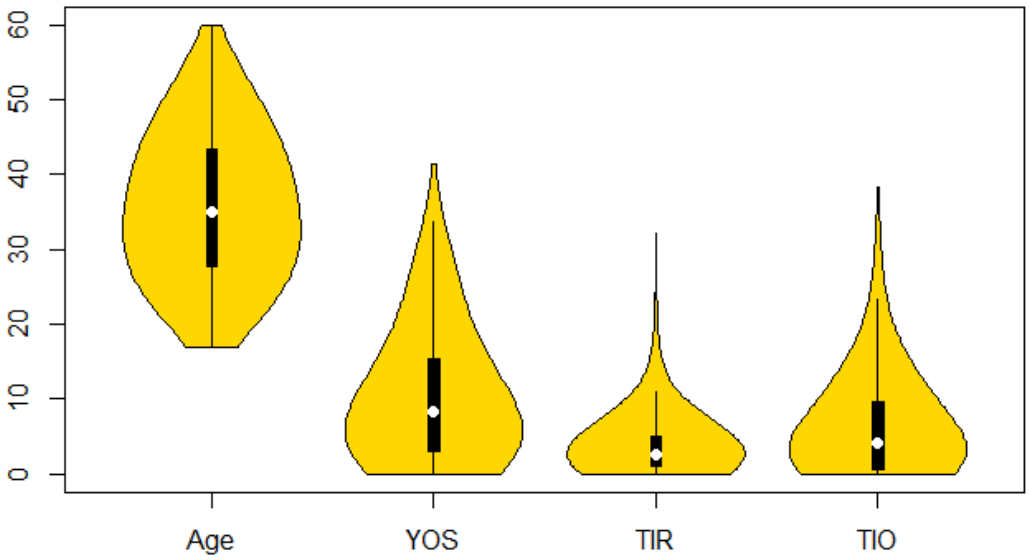

Figure 3.8 – Violin Plot of Date-Calculated Numeric Variables

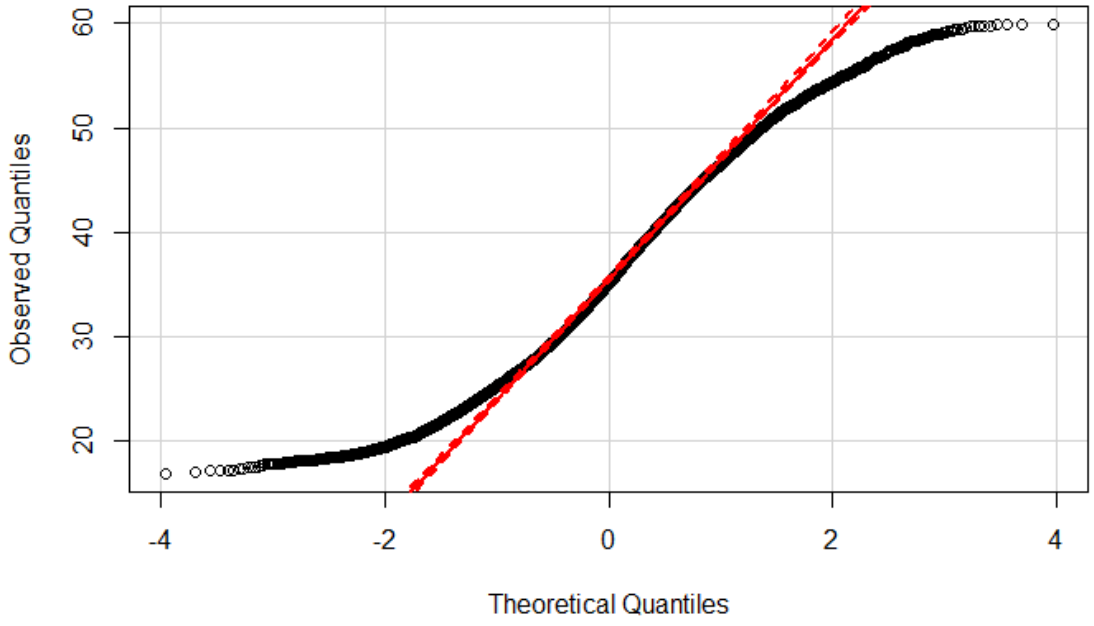

Figure 3.9 – QQ Plot for Age

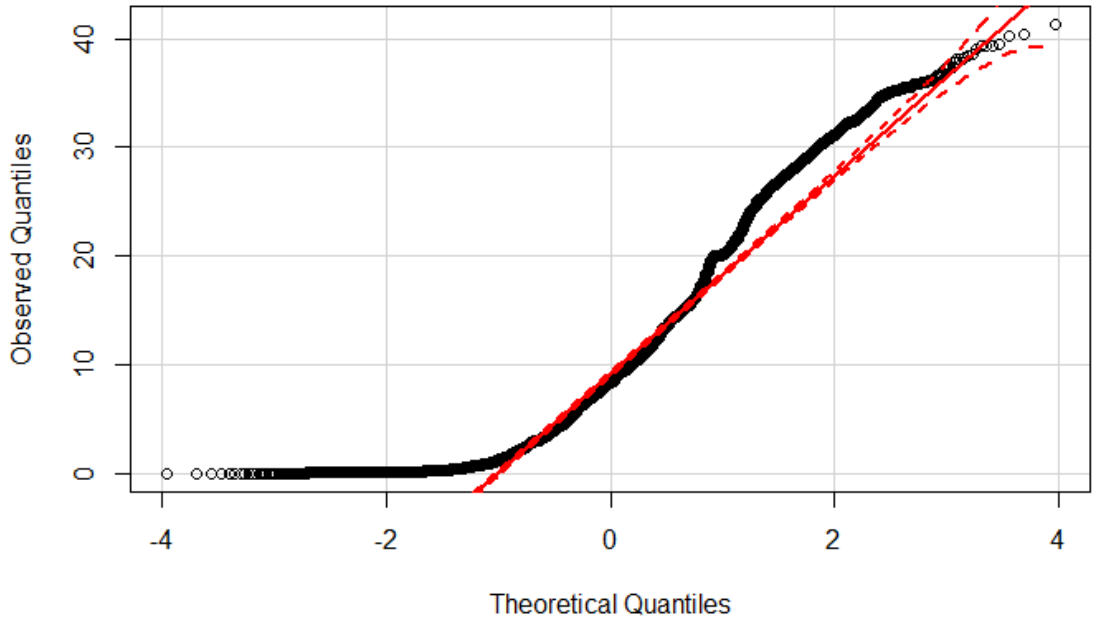

Figure 3.10 – QQ Plot for YOS

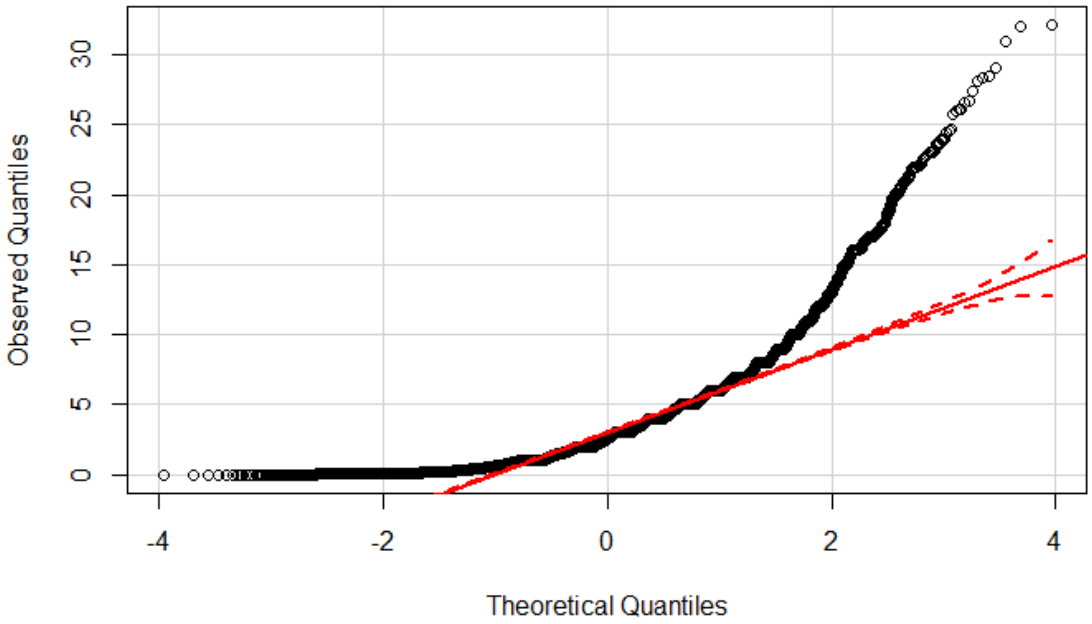

Figure 3.11 – QQ Plot for TIR

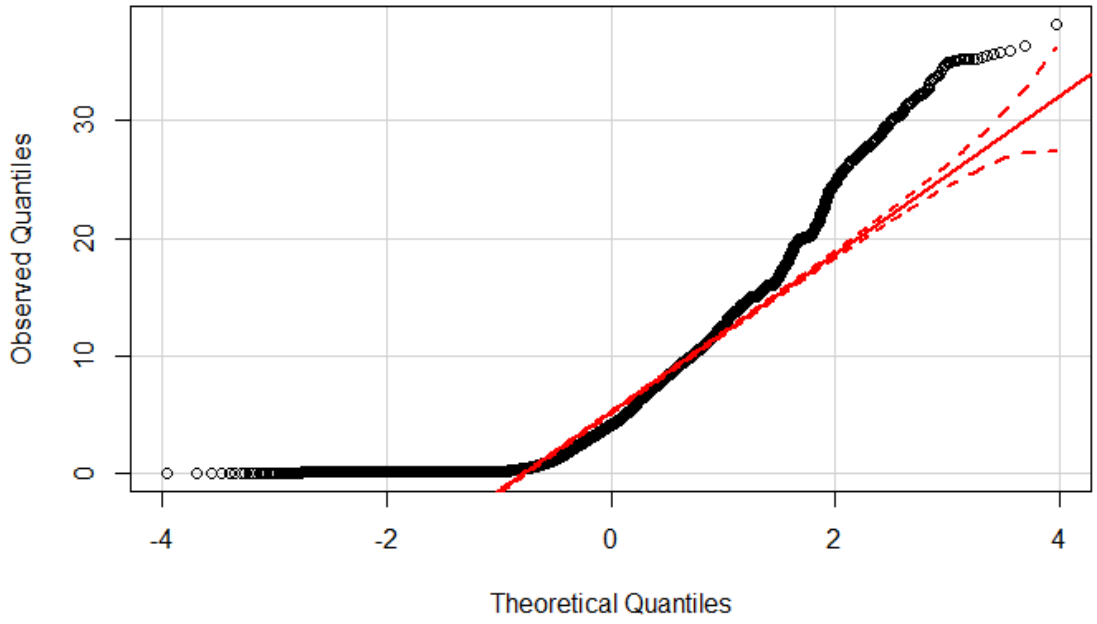

Figure 3.12 – QQ Plot for TIO

The number of children (Children Count) was the last numeric variable included in the data set as extracted from HRMS. The observed histogram (Figure 3.13) shows that most records in the sample did not have children.

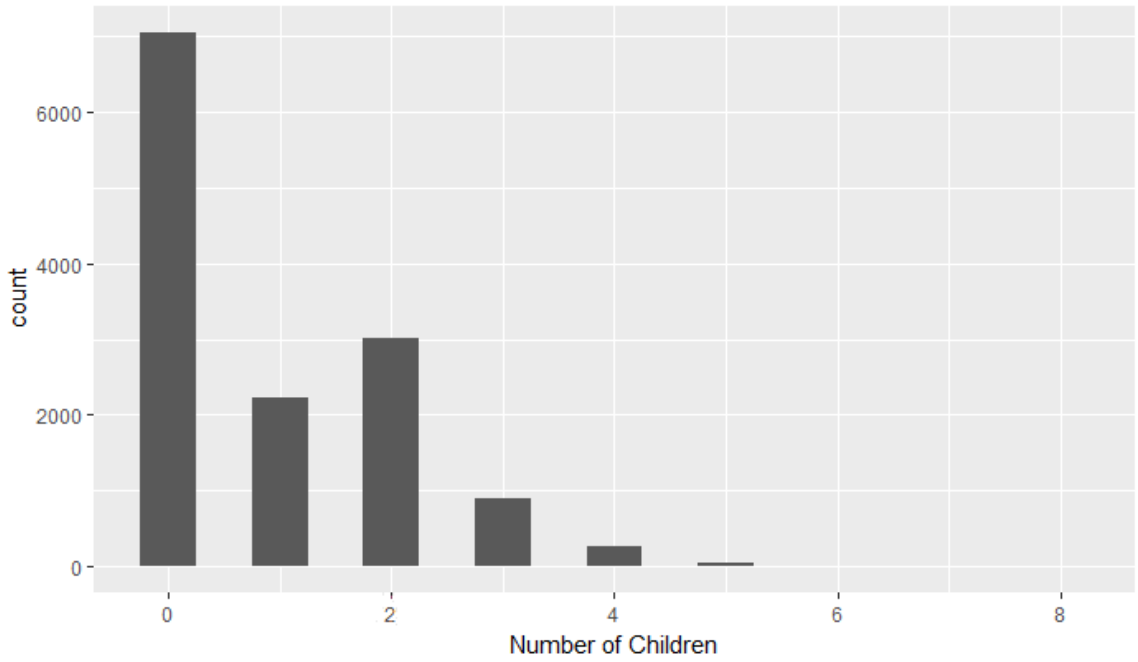

Figure 3.13 – Histogram for the Number of Children for each Member

In addition to analysis of the variables, a histogram was created for the number of voluntary releases for each year for comparison against the attrition population as discussed in the Literature Review. From this histogram (Figure 3.14), it is observed that there were two years (2007 and 2008) that seemingly had a higher number of releases. Referring back to Figure 2.5, this is roughly in line with what was observed for all release types amongst the total population.

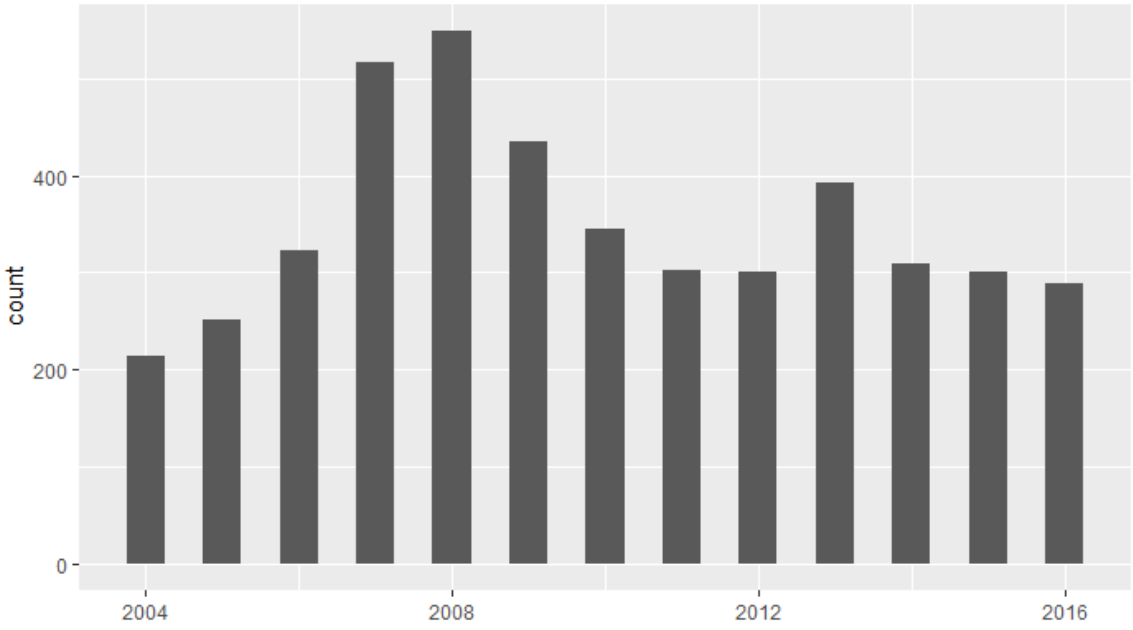

Figure 3.14 – Histogram for the Number of Releases by Year

Handling Missing Data and Imputation

The overall data set was split into training and testing samples before conducting missing data imputation separately on each as opposed to imputing across the entire data set. This was done so as not to leak information between training and testing samples. While the training and testing samples all came from the same overall population, theoretically, the testing data should represent as-of-yet unobserved data such that its own imputation model based on its own component data should be used for missing value substitution. The data sets were randomly split such that the training set contained 70% of records while the testing set contained the other 30%. The proportion of active to terminated personnel (or class balance) was maintained over both data sets.

The overall data set was largely complete with less than 2% of records containing missing information. Using the *VIM* package, a visualization of the missing data was produced (Figure 3.15) showing that the variable with the most missing values was TIR

at ~1.26% followed by five other variables with only a few missing values in each. Figures 3.16 and 3.17 present the missing data for the training and testing data sets respectively. We note that while the training set was missing some data in all of the six missing variables in the overall data set, the test set only had missing data in four of these variables, which was due to the sparsity of the missing data.

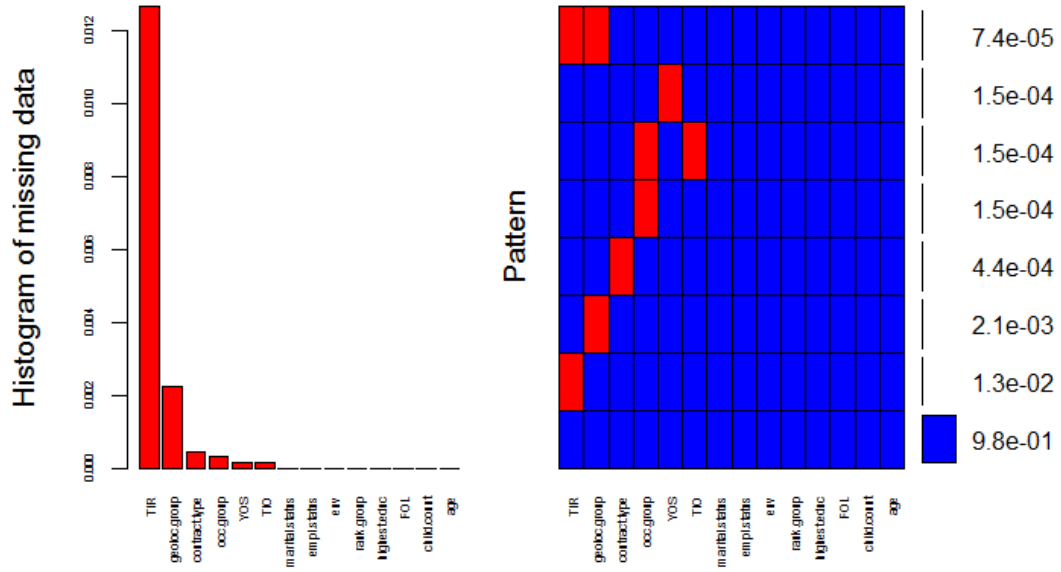

Figure 3.15 – Visualization of Missing Data – Overall Data Set

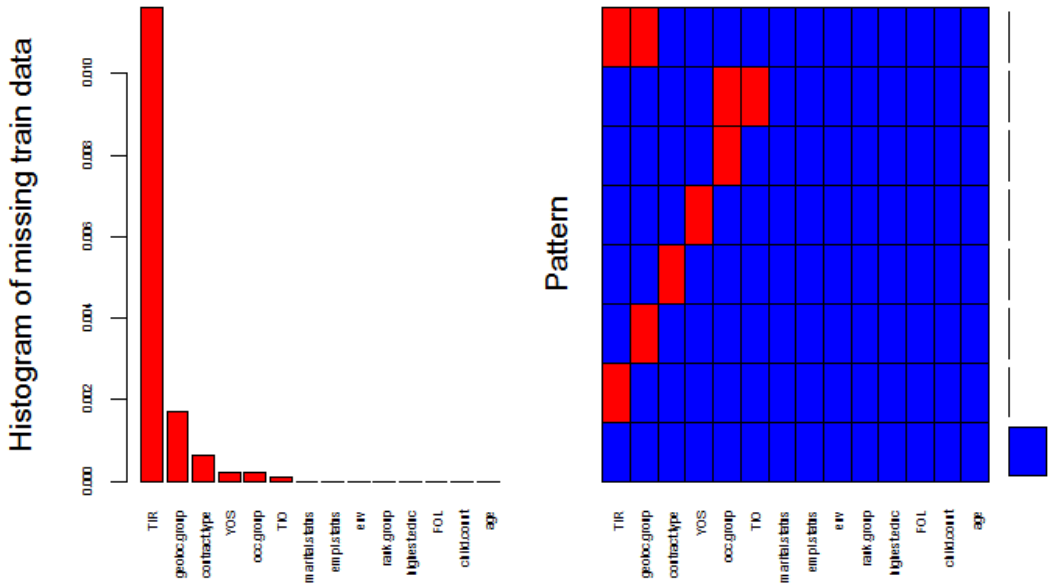

Figure 3.16 – Visualization of Missing Data – Training Set

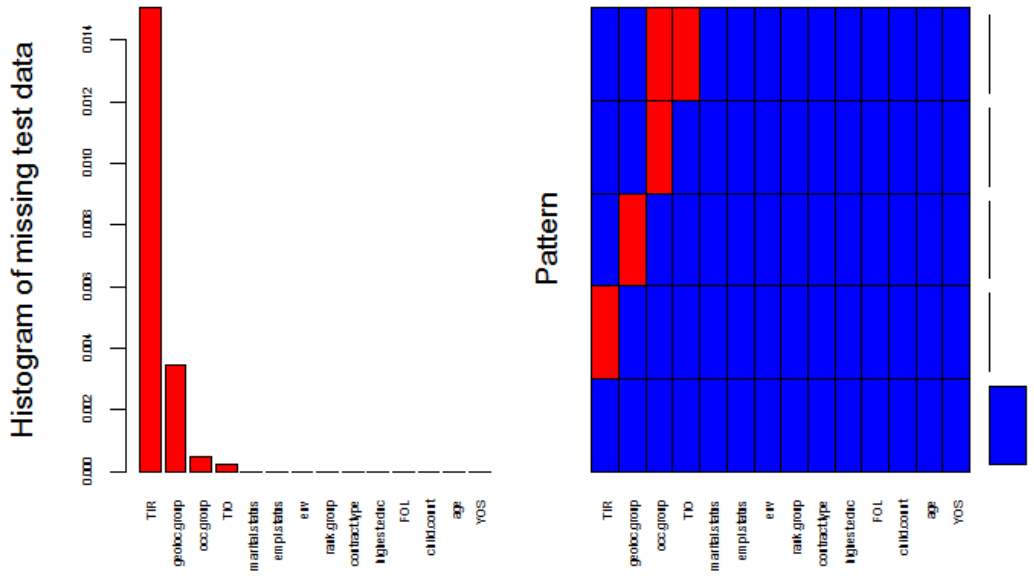

Figure 3.17 – Visualization of Missing Data – Testing Set

Based on analysis, it was assumed that data was Missing at Random (MAR) or Missing Completely at Random (MCAR) and was thus, appropriate to undergo imputation. One key aspect to note is that the intent of imputation is not to produce the most accurate prediction for the theoretical true values of the missing data, but rather, to produce statistically valid data that will not impact model building. That is to say, we expect there to be variance in imputed values since we observe variance in actual values; however, the range in these values should be plausible such that "they could have been obtained had they not been missing."[84]

The key considerations for selecting an imputation method and package included:

- The data consisted of both factor and numeric variables.
- The numeric variables in the data did not follow a Multivariate Normal Distribution.
- The overall size of the data set and the number of factor levels implied the need for an efficient algorithm.

Based on these considerations, Multiple Imputation by Chained Equations (MICE) was used as implemented by the *mice* package in R written by van Buuren and Groothuis-Oudshoorn. The *mice* package uses Fully Conditional Specification (FCS), whereby an imputation model is built on a "variable-by-variable basis by a set of conditional densities, one for each incomplete variable."[85] This involves creating an initial imputation from the non-missing data and then iteratively drawing successive imputations from conditional densities using chained equations, which are tailored to

[84]Stef van Buuren and Karin Groothuis-Oudshoorn, "mice: Multivariate Imputation by Chained Equations in R," *Journal of statistical software* 45, no. 3 (2010): 12, http://www.jstatsoft.org/v45/i03/.
[85]*Ibid.,* 2.

each variable and variable type.[86] This is a suitable method to use when, as in this case, no suitable multivariate distribution (i.e. Multivariate Normal) can be used.[87]

The general method for conducting imputation in ML follows Figure 3.18 whereby an incomplete data set is used to create *m* imputed data sets which are then used to produce *m* fitted models and pooled to create a single aggregated model that accounts for differences within and between the imputed data sets. In the present study, pooling or model averaging were not actually conducted due to difficulties in their application to models produced using gradient descent / regularized regression. The *pool()* function requires that the fitted model object have an associated method for calculating the variance-covariance matrix. By default, the *mice* package uses the *vcov()* function to do this; however, there is no *vcov()* method associated with either of the *glmnet* or *xgboost* packages that were used in training. Further investigation is required, but it is speculated that this is due to the impact of biased estimates arising from regularized regression.

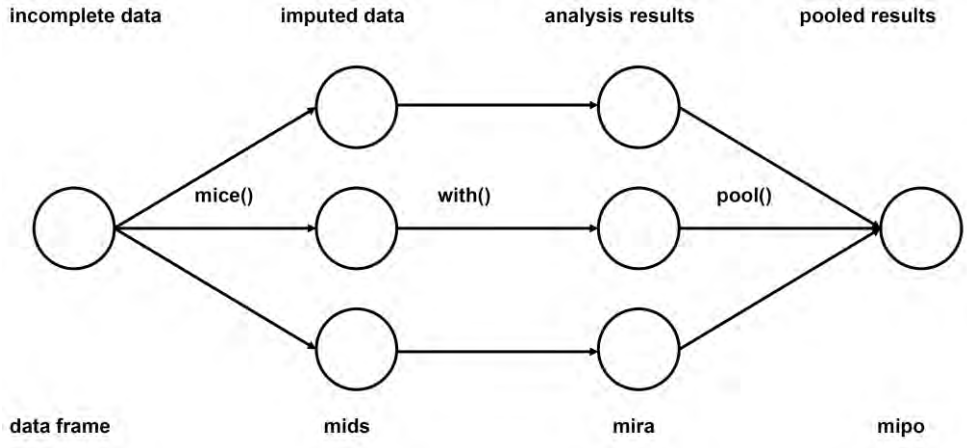

Figure 3.18 – Process for Using Multiply Imputed Data in Machine Learning

[86] *Ibid.*
[87] *Ibid.*

Source: van Buuren and Groothuis-Oudshoorn, *MICE: Multivariate Imputation by Chained Equations,* 5.

In a vignette for the *penalized* package, which performs similar fitting to *glmnet*, Goeman, Meijer, and Chaturvedi, indicate that their package does not provide standard errors with the following explanation:

> The reason for this is that standard errors are not very meaningful for strongly biased estimates such as arise from penalized estimation methods. Penalized estimation is a procedure that reduces the variance of estimators by introducing substantial bias. The bias of each estimator is therefore a major component of its mean squared error, whereas its variance may contribute only a small part.
>
> Unfortunately, in most applications of penalized regression it is impossible to obtain a sufficiently precise estimate of the bias. Any bootstrap-based calculations can only give an assessment of the variance of the estimates. Reliable estimates of the bias are only available if reliable unbiased estimates are available, which is typically not the case in situations in which penalized estimates are used.[88]

As a result, it is assumed that this issue would arise in creating a custom variance-covariance method to work with *glmnet* and *xgboost* and why the models produced by these packages do not include estimates of coefficient variance, thus precluding the use of the *pool()* function. Due to this limitation, models were constructed for each *m* imputed data set, compared visually, and one was randomly selected for further analysis (discussed further in the Assessing Model Performance sub-section later in this chapter).

The modelling techniques used for imputing each variable type are user-set parameters when calling *mice()*. In this case, predictive mean matching (PMM) was used for numeric variables whereas Bayesian Polytomonous Regression (polyreg) was used for categorical variables. PMM is a widely used method pioneered by Little where missing

[88]J. Goeman, R. Meijer, and N. Chaturvedi, "L1 and L2 Penalized Regression Models," last modified 27 May 2016, 18-19, https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

values are imputed using a nearest-neighbour approach with distance determined by the "expected values of the missing variables conditional on the observed covariates."[89] Polytomous regression, on the other hand, is a generalization of logistic regression that can be applied to output variables of more than two categories.[90]

For this study, six imputed data sets were created with ten iterations performed for each missing variable in each data set. Default predictors generated by *mice* were used to predict the missing data (i.e. using all other variables) due to a lack of *a priori* assumptions for covariance amongst predictors and the missing variable. In terms of performance of the imputations, several checks were performed to ensure that the substituted data were valid. A convergence check is shown at Figures 3.19 and 3.20, which graphically depict the mean and standard deviation for imputed values for each data set and iteration; "good" convergence should show that these values mix together over the course of all of the iterations. Visual inspection indicates sufficient mixing and thus, indicates this convergence. Similar results were observed for the testing data set noting the fewer number of variables that were imputed.

---

[89] Gerko Vink, Laurence E. Frank, Jeroen Pannekoek, and Stef van Buuren, "Predictive Mean Matching Imputation of Semicontinuous Variables," Statistica Neerlandica 68, no. 1 (2014): 62. A copy of Little's original work was unavailable, so a secondary source was used. Original citation: Little, R.J.A., "Missing data adjustments in large surveys (with discussion)," *Journal of Business Economics and Statistics,* no. 6 (1988): 287-301.

[90] Jaap Brand, *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets* (Enschede: Print Partners Ispkamp, 1999), 57.

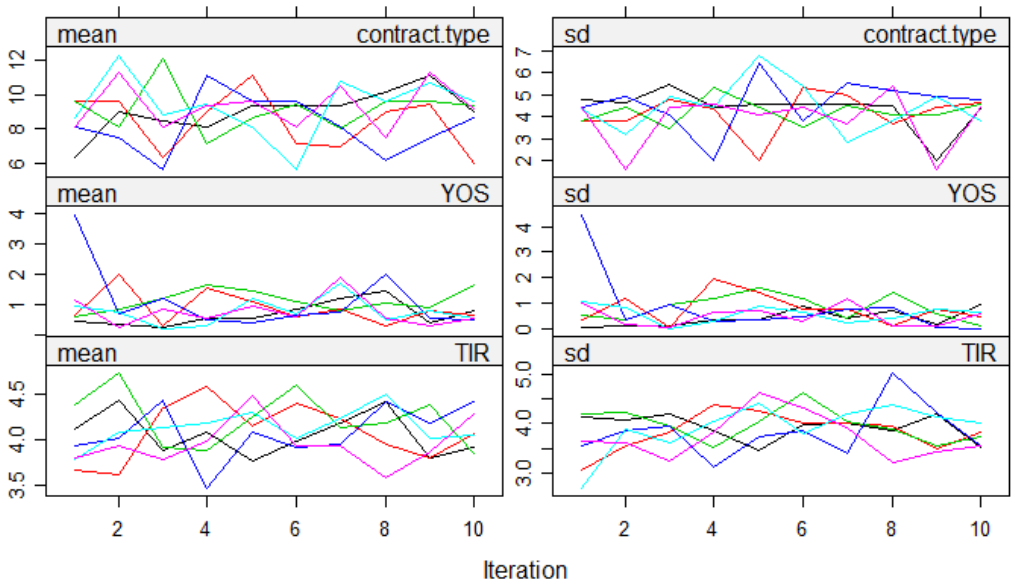

Figure 3.19 – Convergence Check for Contract, YOS, and TIR – Training Data

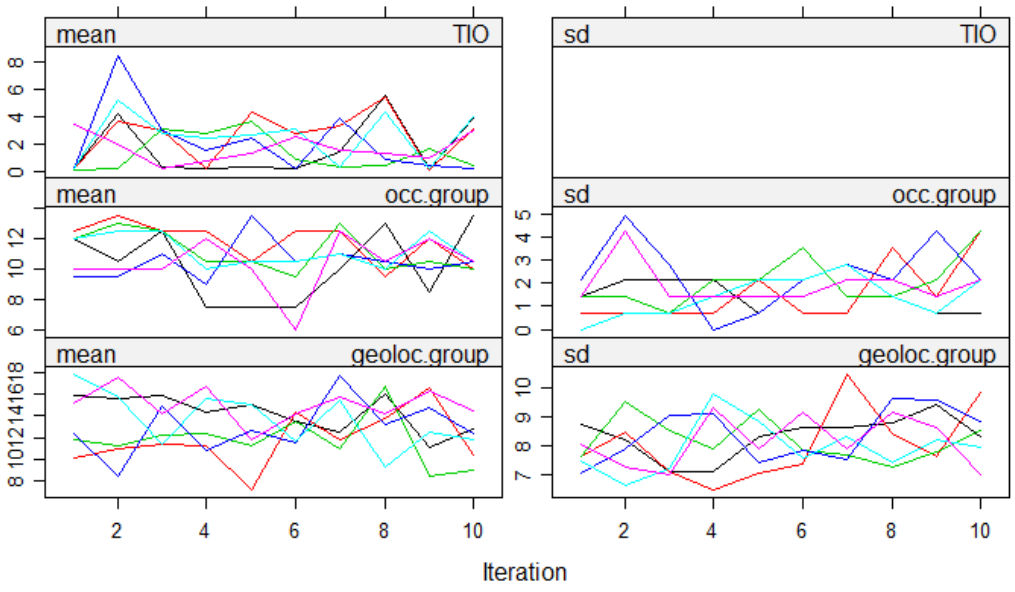

Figure 3.20 – Convergence Check for TIO, Occupation Group, and Location – Training Data[91]

[91]Note there is no standard deviation for TIO as there was only one value to impute for the training data set.

To further analyze the validity of numerical imputations, scatter plots of actual and imputed values were produced for the three numerical variables (Figure 3.21) as well as a density plot for TIR as the variable containing the most missing data (Figure 3.22). Both plots show that imputations produced plausible values when considered against the non-missing data. A separate plot was also produced for imputations of factor variables (Figure 3.23), which again shows plausible imputation. Similar results were seen for the testing data set.

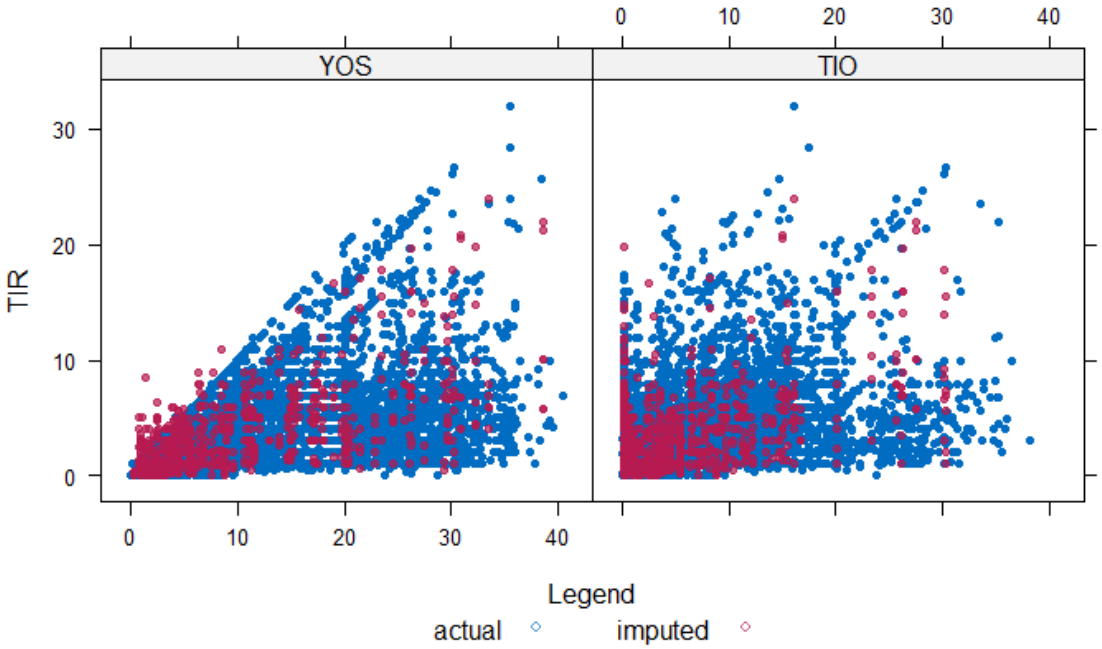

Figure 3.21 – Scatter Plots of Imputed Numeric Variables – Training Data

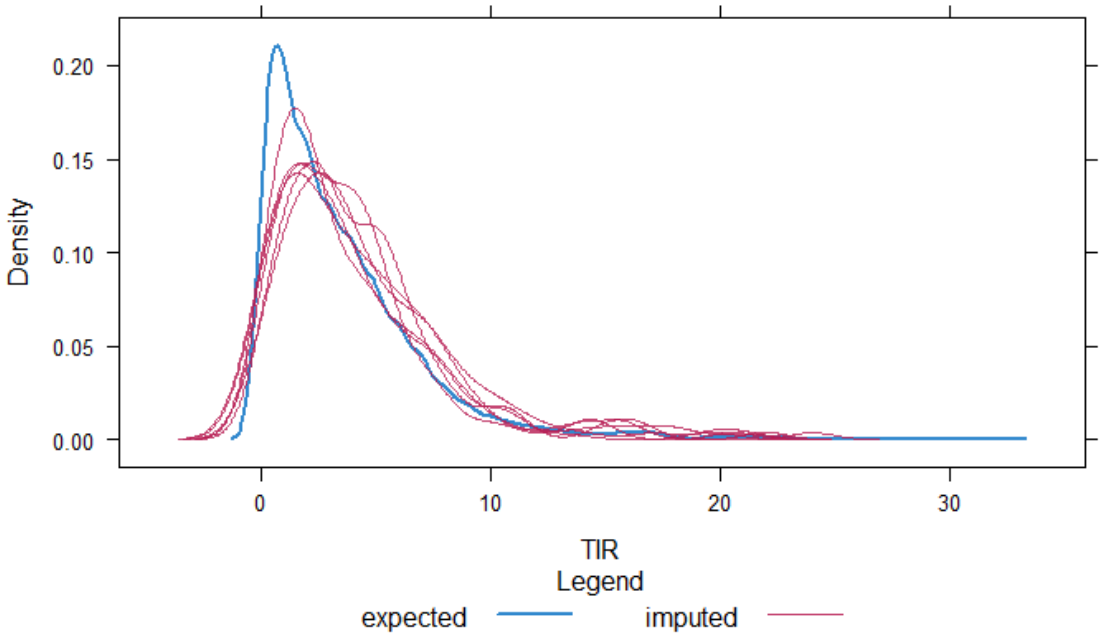

Figure 3.22 – Density Plot of Imputed TIR – Training Data

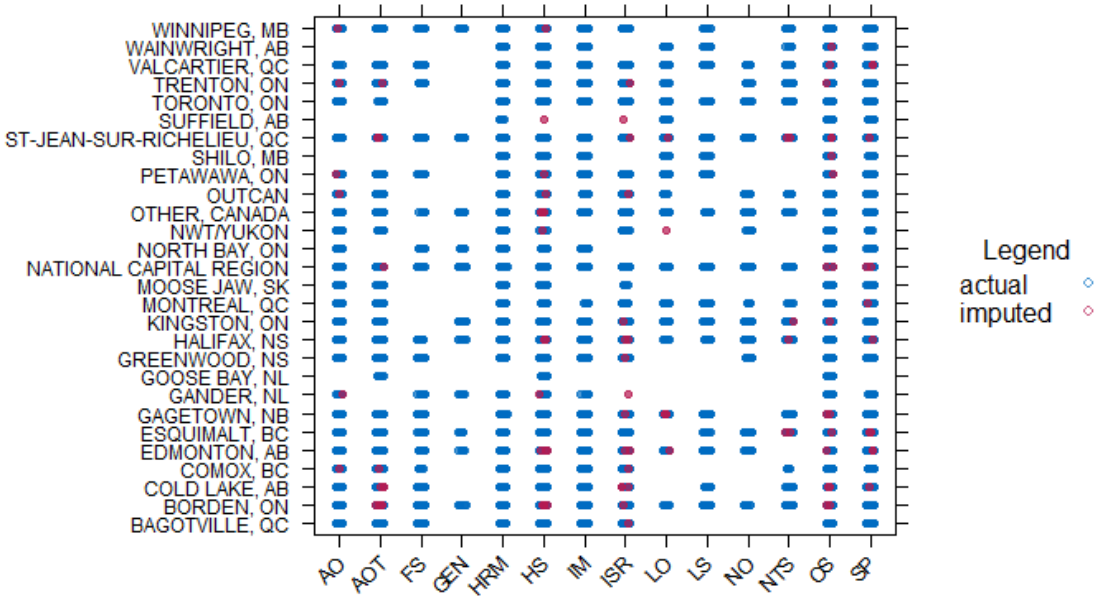

Figure 3.23 – Imputed Values for Location and Occupation Groups – Training Data

Logistic Regression Model

Background and Theory

A terse overview of the theory behind logistic regression as well as the methods used for model tuning / fitting is presented in this section. Logistic regression is similar in concept to linear regression, which is used to estimate the linear relationship between a dependent variable and corresponding independent variables; however, logit modelling is focussed on categorical output variables (i.e. binary outcomes), is based on different underlying assumptions, and uses a different method to estimate parameters (Maximum Likelihood Estimation (MLE)). In this regard, it is considered to be a generalized linear modelling method. Because it based on a sigmoidal logistic function, a value between zero and one is produced making it suitable for probability estimation.

A linear combination of *m* explanatory variables in a regression model with variables x_i and coefficients β_i can be given by:

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m = \beta_0 + \sum_{i=1}^{m} \beta_i x_i \ ^{92}$$

We can take $p(x)$ as the probability of an event occurring while the probability of it not occurring is $1 - p(x)$. The log-odds of these two quantities are expressed as a function of the linear combination of the explanatory variables, which forms the logistic regression equation:

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i \ ^{93}$$

[92]Wikipedia, "Logistic Regression," last accessed 7 April 2018, https://en.wikipedia.org/wiki/Logistic_regression.

[93]*Ibid.*

Where the probability of occurrence is given as the function:

$$p(x) = \frac{1}{1+ e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+ \beta_m x_m)}}$$ [94]

While logistic regression makes use of the MLE to estimate model coefficients, we consider it along with regularized regression for feature selection, which involves regression cost minimization in the trade-off between bias and variance. Roughly stated, a high-biased model is one that is too simple to approximate the relationship and thus, provides a poor fit to the data, whereas high-variance has the opposite issue; the model provides a good approximation of the relationship but is over-fitted to the data, which means the model will not generalize well to future observations.[95] There is a trade-off between the two such that a high-biased model will have low-variance and *vice versa*; thus, part of the intent of model building is to find the optimal balance between bias and variance.

Regularized regression provides an effective method to fit a model that optimizes this trade-off. A penalized logit model will contain all predictors; however, coefficient estimates will undergo *regularization* such that the model *shrinks* them towards zero.[96] This is performed via a fitting penalty that is added to the model's loss function. Since the intent is to minimize the loss function when fitting parameters, the magnitude and amount of coefficients will increase the penalty. There are several different regularization methods; however, two of the most commonly used methods in logistic regression are *ridge* and *least absolute shrinkage and selection operator (lasso)*.

---

[94] *Ibid.*

[95] Gareth James *et al*, *An Introduction to Statistical Learning with Applications in R* (New York: Springer Publishing, 2013), 34-35.

[96] *Ibid.,* 214.

Roughly speaking, the ridge method works by keeping all coefficients but shrinking them towards zero while lasso will force some coefficients all the way to zero based on the value of the tuning parameter.[97]

Despite having some advantages over ridge regression, lasso still has some limitations. As a result, Zou and Hastie developed a more generalized method known as *elastic net regularization* that linearly combines the ridge and lasso penalties.[98] This introduces another parameter α that specifies the type of regularization: ridge for $\alpha = 0$; lasso for $\alpha = 1$; and, elastic net as a linear combination of the two for $0 < \alpha < 1$. Due to its inherent advantages, elastic net regularization was used to train the logistic regression model with α as a tuned hyper-parameter.

Implementation in R

The *glmnet* package in R was used to fit an elastic net logistic regression model to the data. Since the intent was to tune the α and λ hyper-parameters using an automated method, *glmnet* was augmented with the *caret* package as a wrapper to conduct the tuning. This was performed using repeated *k-fold* cross-validation, where *k* represents the number of resamples taken. The *k-fold* cross-validation method resamples the training data into *k* samples, of which *k-1* are used for training the model while the last sample is used for validation based on some performance metric. This is repeated until all *k* resamples are used in validation and then repeated a specified number of times. In this case, the resamples were split into 10 folds and cross-validation was repeated 3 times. 10

[97] *Ibid.*, 215, 219.

[98] Hui Zou and Trevor Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67, no. 2 (2005): 301-320, http://www.jstor.org/stable/3647580.

different values selected by caret were tested for each tuning parameter resulting in 100 pairs of tuning parameters to be validated.

The best tune / final model was based on the fit and tuning parameters that maximized Cohen's Kappa statistic, which measures the accuracy of the classifier taking into account the probability of correctly predicting the output class by chance.[99] Given the slight imbalance of the sample (~66% active and ~33% terminated), model weights were added to penalize inaccurate predictions in the minority class. Two separate sets of models were built: one including interactions and one that included only main effects.

Boosted Tree Model

Background and Theory

Tree-based models are a type of non-linear model that involve "stratifying or segmenting the predictor space into a number of simple regions . . . [which] can be summarized in a [recursive binary decision] tree."[100] Roughly speaking, leaves are created from the root by successive splits in the predictor variables based on minimizing a loss criterion. Prediction of the output variable is made by assigning the most commonly occurring class in that leaf.

[99]Wikipedia, "Cohen's Kappa," last accessed 11 April 2018, https://en.wikipedia.org/wiki/Cohen%27s_kappa.

[100]Gareth James *et al*, *An Introduction to Statistical Learning with Applications in R* (New York: Springer Publishing, 2013), 303.

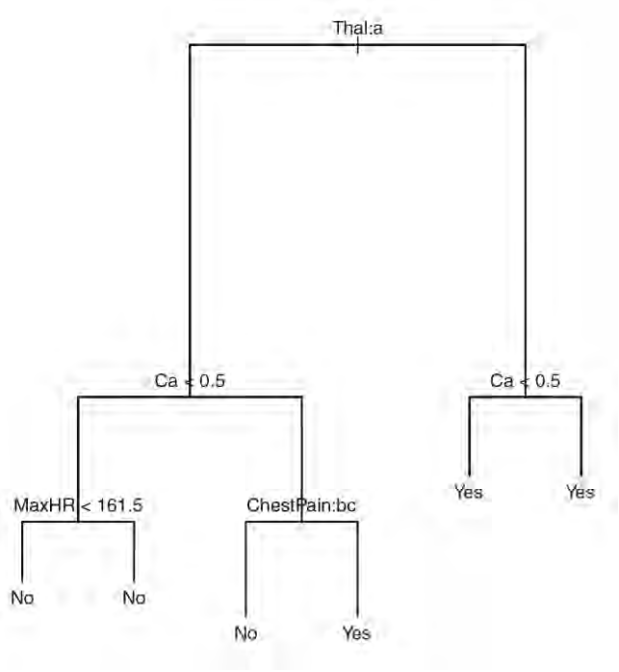

Figure 3.24 – Example Decision Tree with Binary Output

Source: James *et al*, *An Introduction to Statistical Learning with Applications in R,* 313.

Basic tree models, while very visually interpretable, are not typically used as their performance is limited by overfitting in the branches. This is solved by using an ensemble method such as bagging, random forests, or boosting, which involve "producing multiple trees, which are then combined to yield a single consensus prediction," albeit at the expense of interpretability.[101]

The ensemble method selected for this study was boosting, which is similar to bootstrap aggregating (bagging). Bagging "involves creating multiple copies of the original training data set using a bootstrap [resample], fitting a separate decision tree to each copy, and then combining all of the trees in order to create a single predictive model."[102] Boosting uses a similar method; however, trees are grown sequentially using

[101] *Ibid.*
[102] *Ibid.,* 321.

information from the residuals of previous trees to slowly improve the fitted model as a linear combination of all of the trees.[103] Boosting involves a number of tuning parameters, which are discussed in the next section.

Implementation in R

The *xgboost* package in R was used to fit an extreme gradient boosted regression tree to the data wrapped by the *train()* function in *caret* with the same parameters for repeated cross-validation ($k = 10$; repeated 3 times) and performance metric (Cohen's Kappa). The *xgboost()* function uses the following tuning parameters, which are used to regularize the tree being fit in order to avoid overfitting:[104]

- *eta* scales the contribution of each tree, which controls the learning rate and prevents overfitting. It takes a value between 0 and 1.
- *gamma* is the minimum loss reduction required to further split a leaf node.
- *max.depth* provides the depth of the tree.
- *min_child_weight* is a weighting to be used such that if a split leaf node has a sum of instance weight less than min_child_weight, the tree will not split any further.
- *Subsample* is the percentage ratio of the training subsample that *xgboost* uses to train.
- *collsample_bytree* is the percentage ratio of columns used by *xgboost* in each tree.
- *nrounds* is the maximum number of iterations.

---

[103] *Ibid.*

[104] Tianqi Chen *et al*, "xgboost: Extreme Gradient Boosting. R package version 0.6.4.1.," last accessed 20 April 2018, https://CRAN.R-project.org/package=xgboost. The list of tuning parameters was taken from this reference. Info contained in each bullet was taken from this reference.

Given the number of parameters, less granular tuning was applied such that *caret* performed a cross-validated fit with combinations of three values for each parameter as selected by the algorithm. Model weights are not used for boosted trees.

Assessing Model Performance

Given that there were six multiply imputed data sets, six sets of repeated cross-validation models were produced for each ML technique: logit with main-effects only; logit with interactions; and boosted tree. For each technique, the *resamples()* function in *caret* was used to examine the Kappa from each final model produced by repeated cross-validation – 30 resamples for each training data set. Boxplots were produced and analyzed. A random data set of the six was then selected so that the optimally tuned model for each ML technique was based on the same training data set.

Performance and fit for the optimally tuned models were assessed using confusion matrix analysis and the receiver operating characteristic (ROC) with predictions performed on a randomly selected imputed test data set. The metrics used to assess model performance are found below with an example confusion matrix at Table 3.1.

Table 3.1 – Example Confusion Matrix

	Reference	
Prediction	*Positive Class (Terminated)*	*Negative Class (Active)*
Positive Class (Terminated)	True Positives (TP)	False Positives (FP)
Negative Class (Active)	False Negatives (FN)	True Negatives (TN)

False Positive Rate (FPR). FPR is the proportion of True Negatives (TN) misclassified to be in the positive class – i.e. active personnel predicted to be terminated. It is given by the formula:

$$FPR = \frac{FP}{FP+TN}$$ [105]

False Negative Rate (FNR). FNR is the proportion of True Positives (TP) misclassified to be to be in the negative class – i.e. terminated personnel predicted to be active. It is given by the formula:

$$FNR = \frac{FN}{FN+TP}$$ [106]

Sensitivity. Sensitivity, also known as the True Positive Rate (TPR) or recall, is the proportion of TP that were correctly classified. It can be thought of as the probability that the classifier will correctly predict a terminated individual given that they are actually terminated. It is given by the formula:

$$Sensitivity = \frac{TP}{TP+FN} = 1 - FNR$$ [107]

[105] Wikipedia, "Evaluation of Binary Classifiers," last accessed 21 April 2018, https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers.
[106] *Ibid.*
[107] *Ibid.*

Specificity. Specificity, also known as the True Negative Rate (TNR), is the proportion of TN that were correctly classified. It can be thought of as the probability that the classifier will correctly predict an active individual given that they are actually active. It is given by the formula:

$$Specificity = \frac{TN}{FP+TN} = 1 - FPR$$ [108]

Precision or Positive Predictive Value (PPV). PPV, also known as precision, is the proportion of TP out of all positive classifications and can be viewed as the probability that an individual classified as terminated was actually a terminated individual. It is given by the formula:

$$PPV = \frac{TN}{TP+\mathrm{FP}}$$ [109]

Negative Predictive Value (NPV). NPV is the proportion of TN out of all negative classifications and can be viewed as the probability that an individual classified as active was actually an active individual. It is given by the formula:

$$NPV = \frac{TN}{TP+FP}$$ [110]

Accuracy (raw). Accuracy measures the overall agreement between the classifier and the actual values. The issue with accuracy is that it can be boosted by accurate predictions in one class in an unbalanced data set, thus, requiring other measures such as Cohen's Kappa and balanced accuracy to better appreciate overall classifier performance.

---

[108] *Ibid.*
[109] *Ibid.*
[110] *Ibid.*

Cohen's Kappa. Kappa was previously discussed as it is the metric that was maximized in repeated cross-validation tuning; however, we also compute it on the test set for performance analysis. It is calculated based on the observed accuracy of the classifier (p_o) and the hypothetical probability of an accurate prediction by chance (p_e or the expected accuracy given the confusion matrix), given by the formula:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$ [111]

Roughly stated, a classifier that performs no better than random would have a Kappa of 0, while a perfect classifier would have a value of 1. As such, while there are no universal rules for high performing Kappa, it can be considered as the relative performance of the classifier as measured by the interval between random and perfect classification.

Balanced Accuracy. Balanced accuracy takes into account the impact of the classifier's ability to make accurate predictions in each class. As such, it is the arithmetic mean of the specificity and sensitivity of the classifier meaning that underperformance in one of those metrics will bring down the balanced accuracy.

ROC Curve. The ROC Curve is a graphical plot that illustrates the performance of the classifier over all of the values of the decision threshold. It plots the sensitivity (TPR) as a function of the FPR. It is normally compared against a curve produced by a random classifier, depicted by the diagonal on the plot.

[111] Wikipedia, "Cohen's Kappa," last accessed 11 April 2018, https://en.wikipedia.org/wiki/Cohen%27s_kappa.

Area Under ROC (AUROC). AUROC provides a single scalar statistic that summarizes classifier performance. It is the area between the ROC curve and the diagonal that would be expected of a random classifier. Values closer to 1 are better.

Analysis was also performed by assessing the impact of changes to the decision threshold value on the confusion matrix metrics using the optimization tools in the *rocr* and *OptimalCutpoints* packages. The intent behind this portion of the analysis was to illustrate that the classifier can be optimized for some performance metric such as Kappa or account for some business cost associated with misclassification in either the positive or negative class. For example, due to costs associated with terminations, the organization may prefer to ensure that they do not miss on classifying a terminated individual as such (i.e. minimize false negatives). Alternatively, if there is a cost associated with intervention or incentives offered to individuals at risk of attrition, the organization would not want to incur this cost for individuals that are actually not likely to terminate and thus, would not want to inadvertently misclassify active personnel as likely terminations (i.e. minimize false positives). The organization could also seek a decision threshold that balances these qualities, but sufficed to say, it constitutes an important business decision in the application of a classification model of this nature. With this in mind, three confusion matrix analyses were performed based on three adjusted decision thresholds:

- Minimizing overall cost by minimizing both FPR and FNR with each metric weighted equally. This minimizes the total cost of misclassification.

- Minimizing FPR and FNR but applying a weighted cost ratio of 1:2 in order to preference minimizing FN – i.e. preferring greater sensitivity to further shrink the cost of terminations at the expense of increased costs of FP.
- Maximizing the Kappa value using the *opt.cuts()* function as opposed to minimizing the cost of specific cost metrics.

Lastly, a cursory review of variable importance was conducted. As there were many variables and factor levels considered, a full analysis was not performed. Instead, the most important variables as determined by the absolute value of the coefficient in logit regression and the model gain in the boosted tree model were compared using the *varImp()* function in *caret* along with associated graphical plots.

RESULTS AND DISCUSSION

Assessing Results of Repeated Cross-Validation

Kappa values for each repeated cross validation on each imputed training set are depicted as box plots in Figures 4.1 through 4.3. From each box plot, we observe variation between data sets due to the differences in the imputed values but note that they are all fairly close.

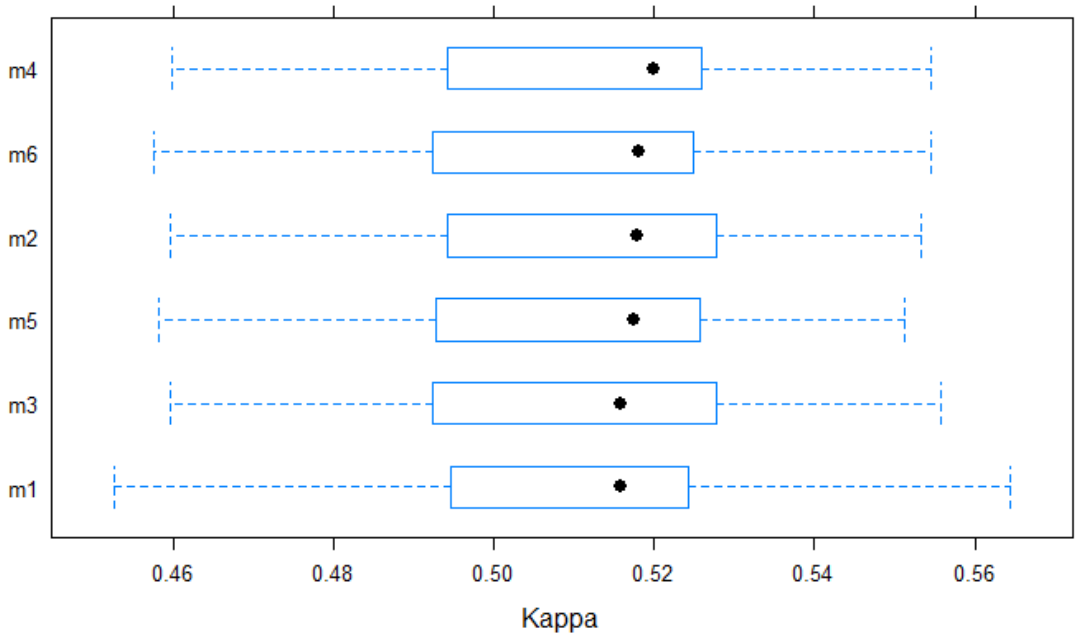

Figure 4.1 – Boxplot of Cohen's Kappa for each Imputed Training Set – Main Effects Logit Model

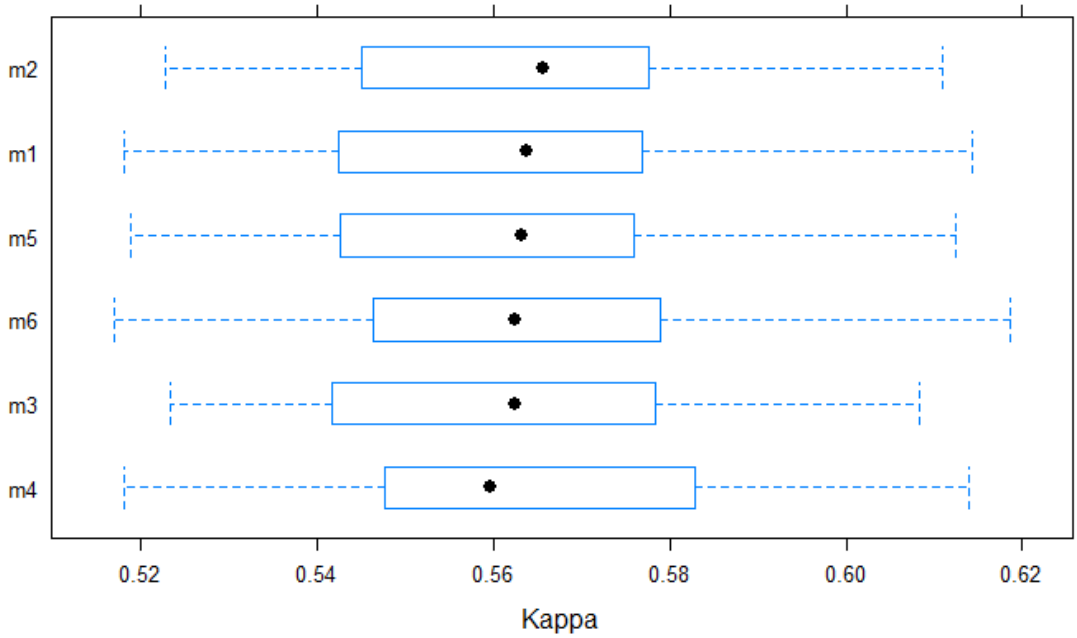

Figure 4.2 – Boxplot of Cohen's Kappa for each Imputed Training Set – Logit Model with Interactions

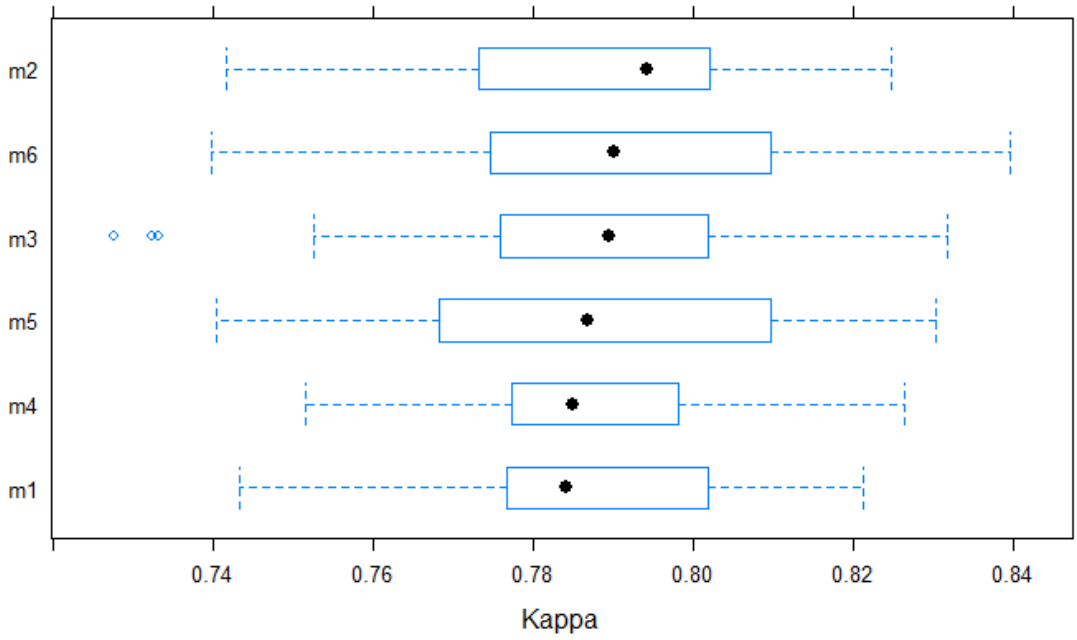

Figure 4.3 – Boxplot of Cohen's Kappa for each Imputed Training Set – Boosted Tree Model

Summary Comparison of Model Fits on Training Data

Before analyzing each model's performance on test data, the results of hyper-parameter tuning are presented in order to look at the characteristics of the best tune for each model. Table 4.1 provides this summary comparison using the randomly selected, second imputed training set.

Table 4.1 – Summary of Training Results for Each Model

	Logit – Main Effects Only	**Logit – Pairwise Interactions**	**XGBoosted Tree**
Hyper-parameters for Best Tune			
alpha (α)	0.7	1	-
lambda (λ)	0.000726	0.00167	-
eta	-	-	0.4
max_depth	-	-	3
gamma (γ)	-	-	0
colsample_bytree	-	-	0.8
min_child_weight	-	-	1
subsample	-	-	1
nrounds	-	-	150
Results for Best Tune (Training Set #2)			
Accuracy (raw)	77.8%	80.0%	90.84%
Standard Deviation Accuracy	1.2%	1.1%	1.0%
Kappa	51.2%	56.3%	78.8%
Standard Deviation Kappa	2.5%	2.4%	2.3%
Number of Non-zero Coefficients	77	733	-
Number of Zeroized Coefficients	1	1649	-

From Table 4.1 we see that the logit model with interactions has been tuned using lasso as the regularization method while the main effects model is between pure elastic net and lasso. As a result, a significant amount of coefficients in the interactions model have been zeroized. From an expected performance perspective, there appears to be a slight difference between the two logit models but both underperform against the boosted tree model by a sizeable margin. Analyzing this further, Figure 4.4 presents a resampled box plot of raw accuracy and Kappa for each model.

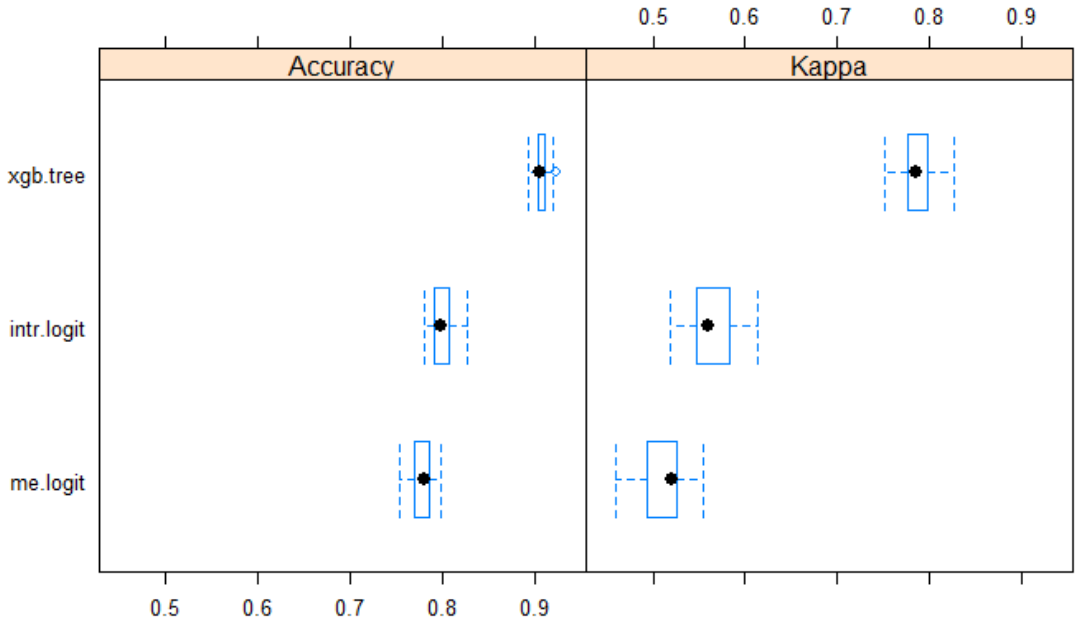

Figure 4.4 – Boxplot of Cohen's Kappa and Accuracy for Each Model Type

The differences between models are visibly noticeable from the box plot; however, the actual statistical significance of these differences was checked with a paired t-test using the *diff()* function in *caret.* Results are presented at Table 4.2 where the upper diagonal presents the estimated difference in Kappa / Accuracy between the models being compared while the lower diagonal provides the p-value under the null hypothesis. The *diff()*function, by default, applies a Bonferroni correction to the p-value to control for issues pertaining to multiple comparisons.[112] We see that the differences between all of the models are statistically significant with the boosted tree model significantly outperforming the other two.

[112]Wolfram Mathworld, "Bonferroni Correction," last accessed 21 April 2018, http://mathworld.wolfram.com/BonferroniCorrection.html.

Table 4.2 – Paired t-test of Differences between Model Types

Test Parameters			
p-value adjustment: bonferroni			
Upper diagonal: estimates of the difference between models			
Lower diagonal: p-value for H_0: difference = 0			
Accuracy (raw)			
	logit – main effects	logit - interactions	xgboost tree
logit – main effects	-	-0.02169	-0.12896
logit - interactions	1.521e-10	-	-0.10728
xgboost tree	< 2.2e-16	< 2.2e-16	-
Kappa			
	logit – main effects	logit - interactions	xgboost tree
logit – main effects	-	-0.05068	-0.27341
logit - interactions	1.473e-11	-	-0.22273
xgboost tree	< 2.2e-16	< 2.2e-16	-

Confusion Matrix Performance

The previous sub-section looked at estimated performance on the training set; however, such analysis must also be applied to the testing set in order to check for overfitting and actual performance on as-of-yet unobserved data. A comparison of the confusion matrices generated from predictions on the test data is found at Table 4.3, which also includes a summary comparison of associated performance metrics. The performance of each model on the test data reinforces what was anticipated from the training data; the boosted tree model outperforms the other two models in every metric.

Table 4.3 – Confusion Matrix Comparison - Default Threshold (*p* = .5)

	Logit – Main Effects Only		**Logit – Pairwise Interactions**		**XGBoosted Tree**	
	Reference					
Prediction	*Term.*	*Active*	*Term.*	*Active*	*Term.*	*Active*
Terminated	992	523	1038	500	1120	126
Active	366	2175	320	2198	238	2572
FPR	19.4%		18.5%		4.7%	
FNR	27.0%		23.6%		17.5%	
Sensitivity (TPR)	73.0%		76.4%		82.5%	
Specificity (TNR)	80.6%		81.5%		95.3%	
PPV	65.5%		67.5%		89.9%	
NPV	85.6%		87.3%		91.5%	
Accuracy (raw)	78.1%		79.8%		91.0%	
Accuracy 95% CI	76.8%	76.8%	78.5%	81.0%	90.1%	91.9%
Expected Accuracy	54.2%		54.0%		56.4%	
Kappa	.522		.561		.794	
Balanced Accuracy	76.8%		79.0%		88.9%	

Receiver Operating Characteristic

A plot of the ROC curves for each model is found at Figure 4.5. The plot shows that the logit model with interactions generally outperforms the main effects model; however, for low values of the FPR, there is some intersection between the two curves, indicating that the interactions model is not completely dominant. That said, the boosted tree model dominates both of the other curves for all values of FPR and a large difference is observed in its AUROC compared to the other two. Thus, the boosted tree shows a noticeably better ability to distinguish between terminated and active classes than the other two models.

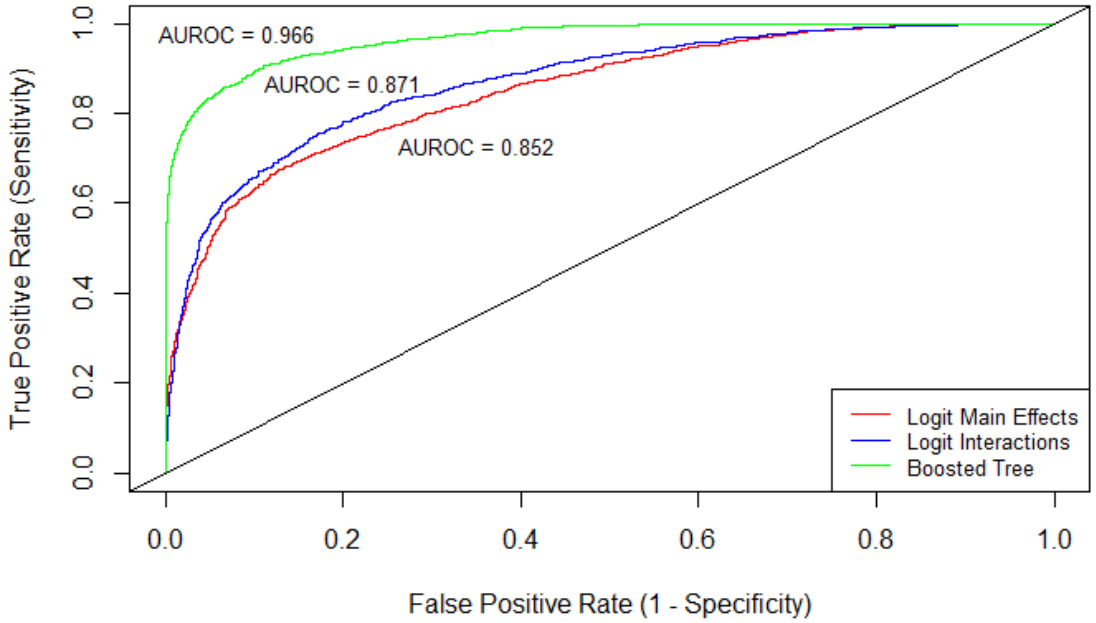

Figure 4.5 – Receiver Operating Characteristic Curves

Adjusting Decision Thresholds

Tables 4.4 through 4.6 present summary comparisons for the three different decision thresholds that were tested with each model: minimizing misclassification cost; using a misclassification cost that was weighted towards minimizing FN over FP; and maximizing Kappa. In general, the boosted tree continued to outperform the other models in all three scenarios but we can see how changing the decision threshold can affect the associated performance metrics, which can be aligned to organizational goals such as minimizing termination costs or producing the best turnover predictions.

Table 4.4 – Confusion Matrix Comparison – Minimized Misclassification Cost Threshold (Even weights; $W_{FP} = W_{FN}$)

	Logit – Main Effects Only		**Logit – Pairwise Interactions**		**XGBoosted Tree**	
Threshold (p =)	.678		.690		.480	
	Reference					
Prediction	*Term.*	*Active*	*Term.*	*Active*	*Term.*	*Active*
Terminated	797	186	846	204	1145	131
Active	561	2512	512	2494	213	2567
FPR	6.9%		7.6%		4.9%	
FNR	41.3%		37.7%		15.7%	
Sensitivity (TPR)	58.7%		62.3%		84.3%	
Specificity (TNR)	93.1%		92.4%		95.1%	
PPV	81.1%		80.6%		89.7%	
NPV	81.7%		83.0%		92.3%	
Accuracy (raw)	81.6%		82.4%		91.5%	
Accuracy 95% CI	80.4%	82.7%	81.1%	83.5%	90.6%	92.4%
Expected Accuracy	58.6%		58.1%		56.0%	
Kappa	.556		.580		.807	
Balanced Accuracy	75.9%		77.4%		89.7%	

Use of this cost minimization makes sense if the cost of misclassifying terminated individuals as active (FN) is equal to the cost of misclassifying active personnel as terminated (FP). From this table, it is observed that accuracy for the logit models has been boosted compared to the default threshold by drastically reducing FP but at the expense of increased FN. As a result, overall accuracy is improved but these models significantly underperform in terms of sensitivity and have a resulting lower balanced accuracy, noting that Kappa is slightly improved. The boosted tree model, on the other hand, shows comparable numbers to the default threshold. Any decreases in performance metrics for the boosted tree are rather small and it has an overall improved balanced accuracy and Kappa.

Table 4.5 – Confusion Matrix Comparison - Weighted Cost Threshold ($W_{FP} = 1$, $W_{FN} = 2$)

	Logit – Main Effects Only		**Logit – Pairwise Interactions**		**XGBoosted Tree**	
Threshold (p =)	.567		.520		.346	
	Reference					
Prediction	*Term.*	*Active*	*Term.*	*Active*	*Term.*	*Active*
Terminated	926	361	1017	449	1218	244
Active	432	2337	341	2249	140	2454
FPR	13.4%		16.6%		9.0%	
FNR	31.8%		25.1%		10.3%	
Sensitivity (TPR)	68.2%		74.9%		89.7%	
Specificity (TNR)	86.6%		83.4%		91.0%	
PPV	72.0%		69.4%		83.3%	
NPV	84.4%		86.8%		94.6%	
Accuracy (raw)	80.5%		80.5%		90.5%	
Accuracy 95% CI	79.2%	81.7%	79.3%	81.7%	89.6%	91.4%
Expected Accuracy	56.2%		54.5%		54.5%	
Kappa	.555		.571		.791	
Balanced Accuracy	77.4%		79.1%		90.3%	

Using this cost minimization makes sense if the cost of misclassifying terminated individuals as active (FN) is the most important consideration. From this table, it is observed that poor sensitivity performance associated with the logit models for the minimized cost threshold is mitigated to a degree. Decreased sensitivity compared to the default threshold is still observed so these models see their performance improvements based on better specificity, noting that performance between the main effects and interactions models were closer than for the other thresholds. The boosted tree model shows similar performance to the previous thresholds, although we observe a sizeable increase in FPR, which is offset by a sizeable decrease in FNR.

Table 4.6 – Confusion Matrix Comparison – Maximized Kappa Threshold

	Logit – Main Effects Only		**Logit – Pairwise Interactions**		**XGBoosted Tree**	
Threshold (p =)	.585		.623		.446	
	Reference					
Prediction	*Term.*	*Active*	*Term.*	*Active*	*Term.*	*Active*
Terminated	902	320	909	279	1170	157
Active	456	2378	449	2419	188	2541
FPR	11.9%		10.3%		5.8%	
FNR	33.6%		33.1%		13.8%	
Sensitivity (TPR)	66.4%		66.9%		86.2%	
Specificity (TNR)	88.1%		89.7%		94.2%	
PPV	73.8%		76.5%		88.2%	
NPV	83.9%		84.3%		93.1%	
Accuracy (raw)	80.9%		82.1%		91.5%	
Accuracy 95% CI	79.6%	82.1%	80.8%	83.2%	90.6%	92.3%
Expected Accuracy	56.5%		57.0%		55.7%	
Kappa	.560		.584		.808	
Balanced Accuracy	77.3%		78.3%		90.2%	

Instead of minimizing misclassification cost, this schema seeks to maximize the Kappa metric which could be seen as maximizing the benefit provided by the selected classifier over a random classifier. As expected, an increase in Kappa is observed for all three models. Similar balanced accuracy is seen when compared to the weighted cost threshold cases, which is indicative of a good balance between specificity, sensitivity, and associated costs.

Variable Importance

Graphical plots of variable importance for each model are presented as Figures 4.6 through 4.8. Factor variable names are, by default, presented as the number that R uses to decode the factor level as that is what it stores in memory; the actual factor level names can be decoded using the matrices at Appendix 2. For models produced using *glmnet,* the plots rank the absolute value of the coefficients for the final tuned model

whereas *xgboost* uses the model gain or summed importance over boosting iterations that is introduced by adding the variable.

Interpretation of these variables also requires some skepticism or acknowledgement of limitations. These values present the *global* interpretability of the coefficients in producing the overall model and not the *local* interpretability for each observation (i.e. whether or not a variable supports or contradicts the overall model for a specific observation). All factor variables are compared to a reference level and thus, must be considered in comparison, not as absolutes. Also, since regularized regression was used, bias has been introduced into the variables, thus shrinking their coefficient estimates or potentially impacting the boosted tree gain, meaning that the full importance of the variable is not necessarily obtained.

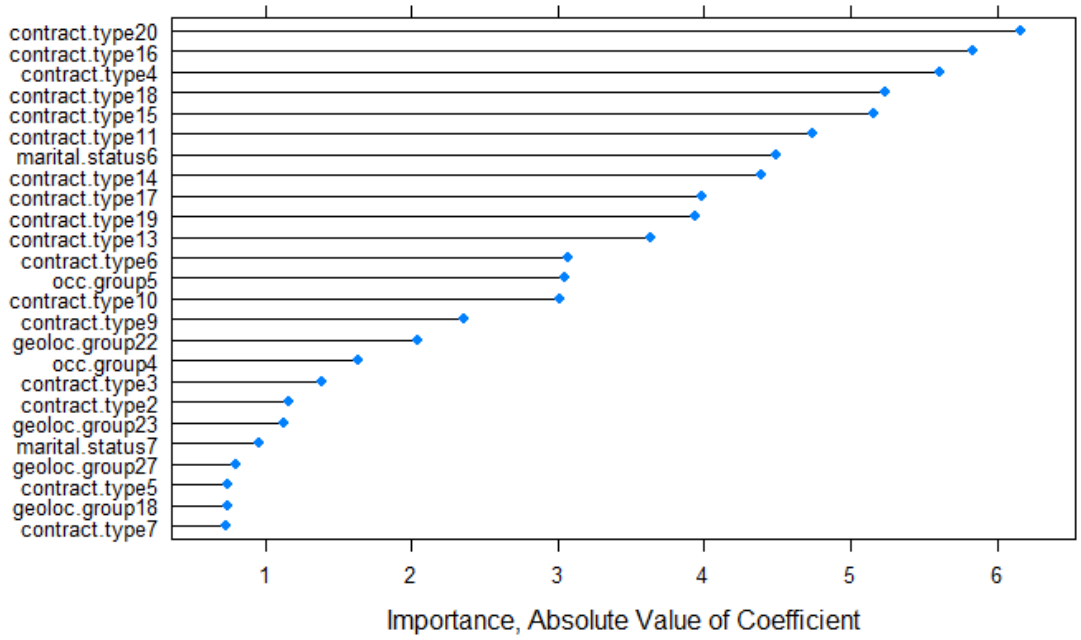

Figure 4.6 – Variable Importance for Main Effects Logit Model

The variable importance for the main effects model appears to change fairly sharply for each of the first 20 variables and then experiences a more gradual decline after that. The most dominant variable appears to be the contract type which should be expected as certain types of contracts should be more likely to appear in certain stages of a member's career when they are at higher attrition risk. As discussed in Chapter 3, this variable contains legacy values from pre-2005 TOS policy changes and thus, the important levels are likely to change over time as the legacy values are phased out. The other interesting variables for this model included occupation and location, potentially reflective of life stage, while unexpectedly, the unknown marital status proved to be important.

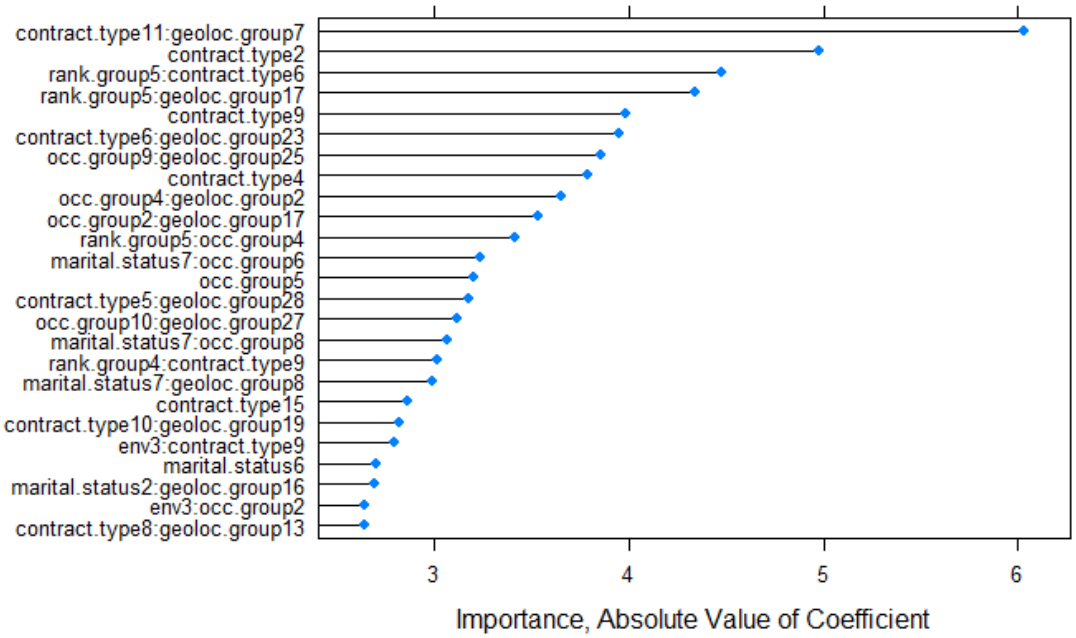

Figure 4.7 – Variable Importance for Logit Model with Interactions

The top four variables experience sharp changes in importance, which begins to smooth in lower tiers. Again, we see the same types of variables as those observed in the main effects model; however, the interactions between variables appear to dominate. That being said, model tuning did not control for improper interactions. For example, the third most important variable was an interaction between senior officers and a contract type only applicable to NCMs. It was anticipated that penalized regression would have taken care of such cases and given that there were no actual observations in the data set corresponding to this interaction, it bears further investigation. Ultimately, this indicates a potential for overfitting and poor applicability of this type of model to the given problem. Future work may involve investigating only specific, allowable interactions and including only those in the regularized regression as opposed to including all pairwise interactions and relying on regularization to bias those that should not exist.

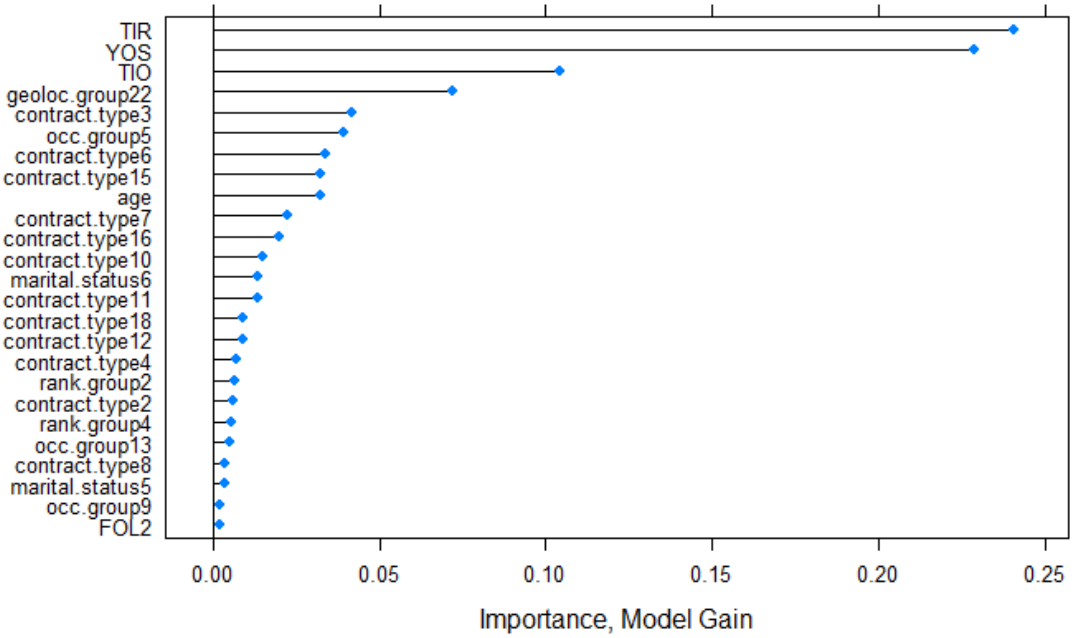

Figure 4.8 – Variable Importance for Boosted Tree Model

This plot shows that the top four variables accounted for significant gain when included in the model. Interestingly, the numeric variables accounted for the three most important variables, with another numeric also in the top 10, indicating that this type of model is perhaps better at accounting for the decision boundary produced by this variable type. Otherwise, we observe some similarities between the important categorical variables for this model versus the other two. For example, when compared to the main effects logit, we observe the following commonalities:

- "St-Jean-sur-Richelieau" (geoloc.group22): likely due to the number of releases processed during recruit training.
- "Human Resources Management" (occ.group5): potentially representing higher attrition in this group.
- "Unknown" marital status and a number of common contract types.

Commentary on Results

From the results, the boosted tree model clearly outperformed the logit models across all performance metrics. This may arise because the boosted tree is more capable of capturing a non-linear decision boundary between the predicted classes. While the logit model does not rely on a linear relationship between dependent and independent variables, it does assume linearity of the independent variables and the log-odds. Looking at the shape of the terminated distribution in the predictors can be revealing in this regard. For example, Figure 4.9 shows a violin plot for YOS separated into the active and terminated classes. Examining the terminated YOS distribution, we can speculate that the log-odds of termination may be higher for the two peak YOS versus other values, thus representing maxima at early and late-career exit points. As such, the relationship between this variable and the log-odds may be non-linear.

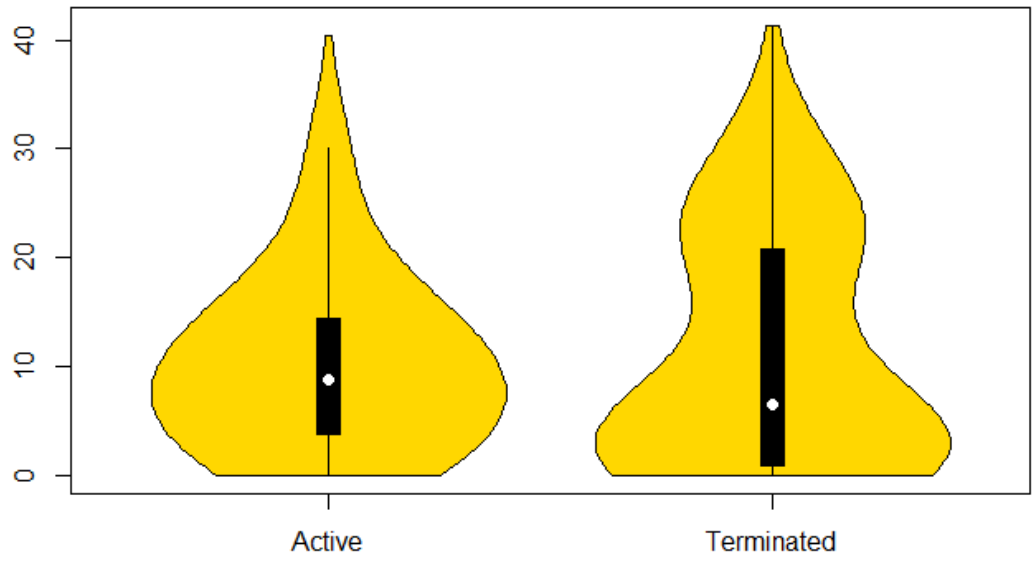

Figure 4.9 – Violin Plots for YOS separated by Output Class

A number of different techniques could be applied to concretely test for and address the violation of the log-odds linearity assumption. It may be possible to code the continuous variables as factors; however, the cut-off for segmenting a continuous variable into a categorical one is somewhat arbitrary. Alternatively, variable transformations could be applied to the continuous variables or a generalized additive model including polynomial terms or spline functions could be created for the logit model to smooth the predictors. Such approaches are outside of the scope of the current study so this could be considered as part of future work; however, given the performance of the boosted tree model and potential of other applicable non-linear models, this may be unnecessary.

Collinearity was not seen to impact the results but is also discussed here. Estimates of the Generalized Variance Inflation Factors (GVIF) for the predictors are presented at Table 4.7, calculated using the *vif()* function in the *car* package using a non-regularized, general linear model. GVIF, as proposed by Fox and Monette, are reduced to the familiar VIF quantity for one degree of freedom (Df), but require an adjustment when multiple Df are present – i.e. for factor variables, since they are associated with more than one coefficient.[113] This adjustment, shown in column 4 of the table, allows for direct comparison of VIFs for variables with different Df by applying the "5/10 rules" to the values in column 5.[114] From this column, YOS is observed to be the only variable that may cause collinearity issues.

---

[113] John Fox and Georges Monette, "Generalized Collinearity Diagnostics," *Journal of the American Statistical Association* 87, no. 417 (March 1992): 180, https://search-proquest-com.cfc.idm.oclc.org/docview/274859555?accountid=9867.

[114] Gareth James *et al*, *An Introduction to Statistical Learning with Applications in R* (New York: Springer Publishing, 2013), 101. A VIF that exceeds 5 or 10 indicates possible collinearity issues.

Table 4.7 – Table of GVIF and Df-adjusted GVIF

	GVIF	Df	$GVIF^{(1/(2*Df))}$	$GVIF^{(1/(Df))}$
Marital Status	1.684134	6	1.044395	1.091
Environment	3.307639	2	1.348589	1.819
Rank Group	7.804728	4	1.292840	1.671
Contract Type	11.986949	19	1.067547	1.140
FOL	1.276861	1	1.129983	1.277
Child Count	1.406675	1	1.186033	1.407
Age	3.213244	1	1.792552	3.214
YOS	8.582898	1	2.929658	8.583
TIR	1.933389	1	1.390464	1.933
TIO	2.333025	1	1.527424	2.333
Occupation Group	6.942906	13	1.077375	1.161
Geolocation Group	7.005347	27	1.036707	1.074

While this analysis shows that YOS may have a collinearity issue, in actual fact, we can be more forgiving of the higher GVIF for this variable due to the regression techniques applied. We would not want to exclude YOS as a variable of study given its *a priori* expectation as being very important to forecasting attrition; however, this is not required since elastic net and boosted tree regression mitigate the effects of collinearity. For the elastic net, this is done through the introduction of bias in the predictors, whereby collinear predictors are either averaged (ridge) or one of them removed (lasso), noting that the pure lasso solution (i.e. logit interactions model) may remove a key predictor from the model.[115] The boosted tree model addresses multi-collinearity in a similar manner to lasso regularization, whereby the tree can choose one of the correlated predictors for a split and discount the other; however, the other correlated predictor may be used in a later tree.[116] Ultimately, as the focus of the study is on accurate prediction

[115] Hui Zou and Trevor Hastie, "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67, no. 2 (2005): 313, http://www.jstor.org/stable/3647580.

[116] Roy Rosemarin, "Is Multicollinearity a Problem with Gradient Boosted Trees?," last modified January 2013, https://www.quora.com/Is-multicollinearity-a-problem-with-gradient-boosted-trees. This

and not focussed on obtaining the best coefficient estimates, collinearity does not significantly impact the output.

Lastly, computation resource usage, while not a specific focus of this work, was considered. Rough measures were taken for the computation time for training over the six imputed data sets for each ML model type. The main effects logit model and boosted tree showed comparable performance, taking roughly 1.5 hours to complete with given resources. Comparatively, the interactions logit model was extremely computationally intensive, taking almost 48 hours to complete under the same conditions. Given that such models are intended to provide tools to policy analysts, computation time is an important consideration as it will impact the timeliness of producing reports. When taken with predictive performance, it adds another attractive feature to the use of a gradient boosted tree model for this problem.

Overall, all models exhibited more difficulty in assigning terminated personnel to the positive class than assigning active personnel to the negative class, consistently skewing towards specificity over sensitivity. Performance in terms of balancing FP and FN was tweaked by adjusting decision thresholds; however, we also observed the expected trade-off between Type I and II errors, while still seeing a general preference towards higher FNR. These results may stem from the slight class imbalance but it is also possible that terminations are simply harder to predict. This could be because of heterogeneity in the Terminated population and the different types of individuals that release, although we would also expect some heterogeneity in the Active class. As a

---

reference is an answer to a question on quora.com; however, it is supported by reference material: Trevor Hastie, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. (New York: Springer Publishing, n.d.), 608-610.

result of this difference, it is up to the organization to conduct cost-benefit analysis in order to determine the optimal decision threshold based on organizational goals, having presented three different options. Heterogeneity in the output classes could also be investigated in the future through the use of clustering analysis.

Applicability to the CAF

While this study provided some indication of the importance of certain variables in modelling voluntary releases amongst females, its utility lies in prediction. In this regard, there are several applications for the model in the CAF. One such important application is attrition forecasting to predict future releases and inform plans for recruiting, intake, and training. Previous CAF research has been focussed on developing methodologies for forecasting bulk attrition rates and has achieved particularly good accuracy. That being said, applying the same methodologies to forecasting attrition amongst sub-groups such as MOSID, rank, or other groupings can run into issues due to small sample sizes.[117] The models built in the current study should be able to circumvent this limitation because they are trained on the entire data set and predictions are applied to each individual, which can be grouped accordingly. In addition to forecasting attrition for future planning, models can also be used to identify individuals at high risk of releasing, which will allow the organization to intervene in an attempt to prevent the release. Table 4.8 demonstrates an example of this, where the number of predicted releases, including those that constitute members of a high risk group ($p \geq .9$), are grouped by Environment, Rank Group, and Occupation Group. A subset was taken to

[117]Nancy Otis and Michelle Straver, *Review of Attrition and Retention Research for the Canadian Forces,* DRDC CORA 2008-030 (Ottawa: Defence Research and Development Canada Centre for Operational Research and Analysis, October 2008): 35-36, http://cradpdf.drdc.gc.ca/PDFS/unc78/p530400.pdf.

display all occupation groups but show only junior NCMs and junior officers in the Army.

Table 4.8 – Example Report of Predicted Terminations by Groups of Interest

Env	Rank Group	Occ Group	Predicted	Actual	FP	FN	High Risk
Land	Jr NCM	FS	0	1	0	1	0
Land	Jr NCM	HRM	0	0	0	0	0
Land	Jr NCM	HS	27	37	1	11	12
Land	Jr NCM	IM	18	21	0	3	10
Land	Jr NCM	ISR	4	7	1	4	1
Land	Jr NCM	LO	38	41	6	9	15
Land	Jr NCM	LS	8	9	0	1	5
Land	Jr NCM	OS	73	73	13	13	40
Land	Jr NCM	SP	3	5	1	3	1
Land	Jr Offr	HRM	0	0	0	0	0
Land	Jr Offr	HS	5	9	0	4	3
Land	Jr Offr	IM	1	1	0	0	0
Land	Jr Offr	ISR	0	0	0	0	0
Land	Jr Offr	LO	3	3	0	0	3
Land	Jr Offr	LS	2	2	0	0	2
Land	Jr Offr	OS	8	9	0	1	4
Land	Jr Offr	SP	7	8	2	3	2

The table also includes the number of actual releases, FP, and FN for the purposes of gauging accuracy. The report performs fairly well in terms of accuracy and could thus function as a baseline planning tool for Career Managers, Trade Advisors, and other interested parties; however, we do not that the smaller sub-groups are very fragile to misclassification error since small fluctuations amount to significant overall error when applying predictions to small groups. FP and FN impact the stability of predictions as expected but overall this query demonstrates the utility of predictive modelling to aid in planning and turnover intervention. This is just one application and queries made on predictions are malleable to the desires of the organization.

Other applications may include identifying and predicting trends over time based on re-training the model on new information or predicting the impact of new policies on actual turnover, noting again that the models built for this study were not designed with causal inference in mind. While these modelling techniques have general applicability, their use in studying voluntary attrition amongst females is important. Given the need to increase female representation in the CAF, such tools will aid in identifying high-risk groups to allow leadership and human resources personnel to intervene and potentially prevent a release given that they have adequate policy tools at their disposal. Such tools can also be used on particular subsets of the data in order to identify trends. For example, if the CAF wants to address first-year attrition amongst females, it could conduct analysis on records from this cohort in order to build a predictive model that can identify individuals at high risk of early career turnover. Based on these predictions, the CAF could develop mechanisms that seek to either lower an individual's risk or mitigate their loss in order to improve the flow of females through the basic recruit training system.

FUTURE WORK AND CONCLUSIONS

Limitations and Future Work

While this study demonstrated the ability to use ML in order to predict turnover amongst the CAF female population, several limitations were identified that should be addressed in future work. First, while multiple imputation produced plausible values across data sets, we observed differences between data sets based on the different imputed values. In this study, there was not a lot of missing data, but the differences between imputed data sets would be more pronounced for a data set with more missing data. As a result, while the CAF should endeavour to ensure that data sets are as complete as possible, new techniques will be required to address this situation. While such techniques require wider work and discussion in the data science community, one such possible method involves using lasso or lasso averaging estimation (LAE) as implemented by the recently released *mami* package to "combine shrinkage estimators with different tuning parameters."[118] This package was not used in the present study due to time constraints and the fact that it does not support tree-based models.

Another limitation that was observed was the violation of the log-odds non-linearity assumptions in the logit model. As discussed, there are several techniques that can be used to address this situation; however, the increased performance of the tree-based model indicates that it may be preferable to further investigate other non-linear, non-parametric, and/or other methods as they may be superior in capturing the decision boundary for this type of problem. To that end, future work should include comparison

[118]Michael Schomaker and Christian Heumann, "Model Averaging and Model Selection after Multiple Imputation using the R-package MAMI," last modified 12 December 2017, 9, http://mami.r-forge.r-project.org/.

of the boosted tree model in this study against models produced by random forests, clustering analysis, neural nets, etc. In doing so, the best model for this particular problem can be selected. That being said, the true utility of such work may lie in constructing an ensemble classifier that can capitalize on the unique strengths of its component models in predicting outcomes. For example, stacked generalization, first proposed by Wolpert and implemented through R packages such as *caretEnsemble,* can be used to create a meta-model that combines component models into an ensemble classifier that may outperform any individual model.[119]

There is also ongoing research in the CAF that can influence or augment future work in this area. One limitation in the present study was that it contained only non-longitudinal data. It did not involve collecting data points over multiple time periods, so the effects of changes in variables over time or the ability to bound predictions to certain time horizons was not possible. According to the CMP intranet site, they are currently engaging in a research study called Project Horizon, whose focus is to "examine the drivers of early commitment and retention using a panel design, i.e. repeated measurements of the same group of individuals over time."[120] This data set would allow for the use of time series methods to forecast attrition, such as the ability to predict the likelihood of termination in the next year. This may be difficult to use globally since the data in HRMS is not collected in this manner, but if mechanisms can be implemented to

[119]David H. Wolpert, "Stacked Generalization," *Neural Networks* 5, no. 2 (1992): 241-259.

[120]Department of National Defence, *Personnel Portfolio (Status as of 31 March 2017)* (Ottawa: Director General Military Personnel Research and Analysis, 31 March 2017), DWAN: http://cmp-cpm.mil.ca/assets/CMP_Intranet/docs/en/support/dgmpra/personnel-portfolio-fy-2016-17-q4-all-activities.pdf.

record snapshots of HRMS data over time, it could become another tool used by analysts in forecasting the individual likelihood of attrition.

Conclusions

Voluntary Attrition is extremely costly for organizations since it implies losing an asset that directly contributes to business outputs, which must be replaced through an expensive process of recruiting, training, acclimation, and experience. Given the CAF's goals of growing its female workforce to 25%, an additional cost is incurred as preventable attrition detracts from this goal. Employee turnover is a natural process that all organizations must live with; however, data science provides tools that allow human resource organizations to either intervene to prevent attrition or at the very least, plan accordingly for how attrition will impact its workforce and future plans.

This study sought to answer questions on whether or not human resources data and ML techniques could be used to predict voluntary releases amongst CAF females, how well such techniques could perform, and which techniques worked best for this classification problem. Three models were trained and then independently tested: a logit model that included only main effects; a logit model that included both main effects and all pairwise interactions between variables; and an extreme gradient boosted tree model. These models were built using regularization methods and repeated cross-validation in order to avoid overfitting the data in the final, tuned model. In the logit models, hyper-parameters were tuned to select the optimal elastic net regularization while for the boosted tree, regularization was achieved by tuning the tree boosting hyper-parameters.

While all three were able to perform the task, the boosted tree model significantly outperformed the other two in all classifier performance metrics. Under the default decision threshold, the boosted tree was able to predict terminations with 91.0% accuracy, 88.9% balanced accuracy, and a Kappa score of .794, which all indicate strong agreement between the boosted tree and the actual values when accounting for Type I/II errors and the probability of chance agreements. As a result, a boosted tree model provides an appropriate ML base for use in a tool designed to assist human resources analysts with attrition planning.

This study also showed that the decision threshold could be adjusted in order to minimize certain misclassification costs, maximize certain metrics, or otherwise align the output of the model to the metrics that are most important for the organization based on cost-benefit analysis. In this manner, trade-offs between metrics were displayed since, in general, the classifiers had more difficulty in predicting terminations, demonstrating lower sensitivity and suffering from higher false negatives; however, adjusting the decision threshold tended to improve performance metrics such as balanced accuracy and Cohen's Kappa.

The applicability of these models in routine CAF activities was also discussed. Given their predictive power and the ability to reuse the code used to build them, these models could be continually improved using new data and provide an analyst with the ready means to forecast attrition in the female (or wider CAF) population. Because predictions are built on individual records, bulk forecasts can be built but can also be subdivided by occupation, rank, environment, or other appropriate grouping, which will allow for more granular forecasting. It is also possible that human resources

interventions or incentives could be offered to high-risk individuals or groups in order to attempt to prevent a voluntary release from occurring.

Ultimately, such applications, while limited in this study, point to the wider applicability of ML to the CAF. Continued academic study of attrition and its causes is important, but it is possible to operationalize such academic study into approaches that can be used to readily answer a myriad of questions that are important in routine and operational decision making. This should include further use of analytics and ML in human resources analysis but can be equally applied to decisions on force readiness, capability development, intelligence, public affairs, and other areas that generate significant data for inference and predictive purposes. In doing so, the CAF can take true steps towards harnessing the power of data in defence of the nation.

APPENDIX 1 – OCCUPATION GROUP CODING

Table A1.1 – Occupation Group by MOSID

Occ Group	MOSID	MOS Description
AO	00019	Airborne Electronic Sensor Operator
AO	00021	Flight Engineer
AO	00101	Search And Rescue Technician
AO	00182	Air Combat Systems Officer
AO	00183	Pilot
AO	00184	Aerospace Control
AO	00337	Aerospace Control Operator
AOT	00135	Aviation Systems Technician
AOT	00136	Avionic Systems Technician
AOT	00138	Aircraft Structures Technician
AOT	00185	Aerospace Engineering
AOT	00189	Construction Engineering
AOT	00261	Air Weapons System Technician
AOT	00343	Non-Destructive Testing Technician
AOT	00363	Air Maintenance Supervisor
FS	00149	Firefighter
FS	00301	Refrigeration and Mechanical Technician
FS	00302	Electrical Distribution Technician
FS	00303	Electrical Generation Systems Technician
FS	00304	Plumbing and Heating Technician
FS	00305	Water, Fuels and Environment Technician
FS	00306	Construction Technician
FS	00307	Construction Engineer Superintendent
HRM	00208	Personnel Selection
HRM	00211	Training Development
HS	00152	Medical Laboratory Technologist
HS	00153	Medical Radiation Technologist
HS	00155	Biomedical Electronics Technologist
HS	00192	Health Care Administration
HS	00193	Health Service Operations
HS	00194	Pharmacy
HS	00195	Nursing Officer Specialist
HS	00334	Medical Technician
HS	00335	Dental Technician
IM	00109	Aerospace Telecommunications And Information System
IM	00340	Communications and Electronics Engineering (Air)
IM	00341	Signals

Occ Group	MOSID	MOS Description
IM	00362	Army Communication & Information Systems Specialist
ISR	00099	Intelligence Operator
ISR	00100	Meteorological Technician
ISR	00120	Communicator Research
ISR	00137	Imagery Technician
ISR	00203	Public Affairs Officer
ISR	00213	Intelligence
ISR	00238	Geomatics Technician
LO	00005	Crewman
LO	00010	Infantryman
LO	00178	Armour
LO	00179	Artillery
LO	00180	Infantry
LO	00181	Engineer
LO	00339	Combat Engineer
LO	00368	Artilleryman
LS	00129	Vehicle Technician
LS	00130	Weapons Technician (Land)
LS	00134	Material Technician
LS	00187	Electrical And Mechanical Engineering
LS	00327	Electronic-Optronic Technician (Land)
NO	00114	Naval Combat Information Operator
NO	00115	Naval Electronic Sensor Operator
NO	00207	Maritime Surface And Sub-Surface
NO	00299	Naval Communicator
NO	00324	Sonar Operator
NTS	00105	Boatswain
NTS	00124	Hull Technician
NTS	00125	Electrical Technician
NTS	00165	Steward
NTS	00342	Clearance Diver
NTS	00344	Naval Combat Systems Engineering
NTS	00345	Marine Systems Engineering
NTS	00346	Naval Engineering
NTS	00366	Weapons Engineering Technician
NTS	00367	Marine Engineer
OS	00164	Cook
OS	00167	Postal Clerk
OS	00168	Supply Technician
OS	00169	Ammunition Technician

Occ Group	MOSID	MOS Description
OS	00170	Traffic Technician
OS	00171	Mobile Support Equipment Operator
OS	00298	Resource Management Support Clerk
OS	00322	Court Reporter
OS	00328	Logistics
SP	00161	Military Police
SP	00166	Musician
SP	00190	Physiotherapy Officer
SP	00191	Dental Officer
SP	00196	Medical
SP	00197	Bioscience
SP	00198	Social Work
SP	00204	Legal
SP	00210	Music
SP	00214	Military Police Officer
SP	00349	Chaplain
SP	00357	Chemical, Biological, Radioactive, Nuclear Operator

Source: Department of National Defence, "Annual Report on Regular Force Personnel 2015/2016", B44-B46.

Table A1.2 – Abbreviations of Occupation Groups

AO	Air Operations
AOT	Air Operations Technical Support
FS	Facility Support
GEN	Generalist
HRM	Human Resources Management
HS	Health Services
IM	Information Management
ISR	Intelligence, Surveillance, and Reconnaissance
LO	Land Operations
LS	Land Support
NO	Naval Operations
NTS	Naval Technical Support
OS	Operations Support
SP	Specialist

Source: Department of National Defence, “Annual Report on Regular Force Personnel 2015/2016”, D2.

APPENDIX 2 – FACTOR LEVELS VERSUS NAMES USED IN REGRESSION

When returning coefficient names, R returns the variable name and the number of the factor level that it stores. The tables below outline how the regression model variables were coded for the factor variables.

Table A2.1 – Marital Status Levels vs Regression Model Names

Name of Level	Name used in Regression Model
Common-Law	Reference Level
Divorced	marital.status2
Married	marital.status3
Separated	marital.status4
Single	marital.status5
Unknown	marital.status6
Widowed	marital.status7

Table A2.2 – Environment Levels vs Regression Model Names

Name of Level	Name used in Regression Model
Air	Reference Level
Land	env2
Sea	env3

Table A2.3 – Rank Group Levels vs Regression Model Names

Name of Level	Name used in Regression Model
Jr NCM	Reference Level
Sr NCO	rank.group2
Sub Offr	rank.group3
Jr Offr	rank.group4
Sr Offr	rank.group5

Table A2.4 – Contract Type Levels vs Regression Model Names

Name of Level	Name used in Regression Model
Continuing Engagement	Reference Level
Indefinite Period of Service	contract.type2
Intermediate Engagement 25	contract.type3
NCM Career Dev Plan CE	contract.type4
NCM Career Dev Plan FPS	contract.type5
NCM Career Dev Plan IPS	contract.type6
NCM Carer Dev Plan IE	contract.type7
Off Career Dev Plan IE	contract.type8
Off Career Dev Plan IPS	contract.type9

Off Career Dev Plan SE	contract.type10
Off Career Dev Plan SSE	contract.type11
Variable Initial Engagement	contract.type12
Indefinite Period of SVC NCM	contract.type13
Indefinite Period of Svc Off	contract.type14
NCM Career Dev Plan BE	contract.type15
NCM Career Dev Plan BE2	contract.type16
NCM Career Dev Plan Ext	contract.type17
Off Career Dev Plan	contract.type18
Off Career Dev Plan Ext	contract.type19
Off Career Dev Plan FPS	contract.type20

Table A2.5 – First Official Language Levels vs Regression Model Names

Name of Level	Name used in Regression Model
English (E)	Reference Level
French (F)	FOL2

Table A2.6 – Occupation Group Levels vs Regression Model Names

Name of Level	Name used in Regression Model
AO	Reference Level
AOT	occ.group2
FS	occ.group3
GEN	occ.group4
HRM	occ.group5
HS	occ.group6
IM	occ.group7
ISR	occ.group8
LO	occ.group9
LS	occ.group10
NO	occ.group11
NTS	occ.group12
OS	occ.group13
SP	occ.group14

Table A2.7 – Geolocation Group Levels vs Regression Model Names

Name of Level	Name used in Regression Model
BAGOTVILLE, QC	Reference Level
BORDEN, ON	geoloc.group2
COLD LAKE, AB	geoloc.group3
COMOX, BC	geoloc.group4
EDMONTON, AB	geoloc.group5
ESQUIMALT, BC	geoloc.group6

GAGETOWN, NB	geoloc.group7
GANDER, NL	geoloc.group8
GOOSE BAY, NL	geoloc.group9
GREENWOOD, NS	geoloc.group10
HALIFAX, NS	geoloc.group11
KINGSTON, ON	geoloc.group12
MONTREAL, QC	geoloc.group13
MOOSE JAW, SK	geoloc.group14
NATIONAL CAPITAL REGION	geoloc.group15
NORTH BAY, ON	geoloc.group16
NWT/YUKON	geoloc.group17
OTHER, CANADA	geoloc.group18
OUTCAN	geoloc.group19
PETAWAWA, ON	geoloc.group20
SHILO, MB	geoloc.group21
ST-JEAN-SUR-RICHELIEU, QC	geoloc.group22
SUFFIELD, AB	geoloc.group23
TORONTO, ON	geoloc.group24
TRENTON, ON	geoloc.group25
VALCARTIER, QC	geoloc.group26
WAINWRIGHT, AB	geoloc.group27
WINNIPEG, MB	geoloc.group28

BIBLIOGRAPHY

Adler, Daniel, Duncan Murdoch et al. “rgl: 3D Visualization Using OpenGL.” R package version 0.99.16. Last accessed 20 April 2018. https://CRAN.R-project.org/package=rgl.

Adler, Daniel. “vioplot: Violin Plot.” R package version 0.2. Last accessed 20 April 2018. https://cran.r-project.org/package=vioplot.

Bates, Douglas and Martin Maechler. “Matrix: Sparse and Dense Matrix Classes and Methods.” R package version 1.2-12. Last accessed 20 April 2018. https://CRAN.R-project.org/package=Matrix.

Bliss, William G. "Calculating the Costs of Employee Turnover." *Fairfield County Business Journal* 40, no. 32 (2001): 12.

Brand, Jaap. *Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets.* Enschede: Print Partners Ispkamp, 1999.

Buuren, S. van, and Karin Groothuis-Oudshoorn. "mice: Multivariate Imputation by Chained Equations in R." *Journal of statistical software* 45, no. 3 (2010): 1-68. http://www.jstatsoft.org/v45/i03/.

Calitoiu, Dragos. *Logistic Regression Model for Medical Attrition.* DRDC-RDDC-2016-RXXX [DRAFT]. Ottawa: Director General Military Personnel Research and Analysis, November 2016.

Calitoiu, Dragos. *Logistic Regression Model for Voluntary Attrition.* DRDC-RDDC-2016-RXXX [DRAFT]. Ottawa: Director General Military Personnel Research and Analysis, November 2016.

Chen, Tianqi, Tong He, Michael Benesty, Vadim Khotilovich and Yuan Tang. “xgboost: Extreme Gradient Boosting. R package version 0.6.4.1.” Last accessed 20 April 2018. https://CRAN.R-project.org/package=xgboost.

Cox, D. R. "The Regression Analysis of Binary Sequences." *Journal of the Royal Statistical Society. Series B (Methodological)* 20, no. 2 (1958): 215-42. http://www.jstor.org/stable/2983890.

Davis, Karen D. “Understanding Women's Exit from the Canadian Forces: Implications for Integration?” In *Wives and Warriors: Women and the Military in the United States and Canada,* edited by Laurie Weinstein and Christie C. White, 179-188. Westport, CT: Bergin & Garvey, 1997.

Department of National Defence. *Annual Report on Regular Force Attrition 2014/2015.* Ottawa: Director General Military Personnel Research and Analysis, June 2016.

Department of National Defence. *Annual Report on Regular Force Personnel 2015/2016.* Ottawa: Director General Military Personnel Research and Analysis, October 2017.

Department of National Defence, *Audit of Force Reduction Program.* Ottawa: Director General Audits, January 1997. http://www.forces.gc.ca/assets/FORCES_Internet/docs/en/about-reports-pubs-audit-eval/705529.pdf.

Department of National Defence. *Canadian Forces Employment Equity Report 2011-2012.* Ottawa: DND Canada, October 2012. http://publications.gc.ca/collections/collection_2014/mdn-dnd/D3-31-2012-eng.pdf.

Department of National Defence. *Personnel Portfolio (Status as of 31 March 2017).* Ottawa: Director General Military Personnel Research and Analysis, 31 March 2017. DWAN: http://cmp-cpm.mil.ca/assets/CMP_Intranet/docs/en/support/dgmpra/personnel-portfolio-fy-2016-17-q4-all-activities.pdf.

Dowle, Matt and Arun Srinivasan. "data.table: Extension of `data.frame`." R package version 1.10.4-3. Last accessed 20 April 2018. https://CRAN.R-project.org/package=data.table.

Fang, Manchun and Paul Bender. *Canadian Forces Attrition Forecasts: What One Should Know.* DRDC CORA 2008-060. Ottawa: Defence Research and Development Canada Centre for Operational Research and Analysis, December 2008. http://www.dtic.mil/dtic/tr/fulltext/u2/1001567.pdf.

Fang, Manchun and Stephen Okazawa. *Forecasting Attrition Volume: A Methodological Development.* DGMPRA TM 2009-025. Ottawa: Defence Research and Development Canada and Director General Military Personnel Research & Analysis, December 2009. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc126/p532444_A1b.pdf.

Fischetti, Tony. "assertr: Assertive Programming for R Analysis Pipelines." R package version 2.5. Last accessed 20 April 2018. https://CRAN.R-project.org/package=assertr

Fischetti, Tony, Eric Mayor, and Rui Miguel Forte. R: Predictive Analysis. 1st ed. Birmingham: Packt Publishing, 2017.

Fox, John and Georges Monette. "Generalized Collinearity Diagnostics." *Journal of the American Statistical Association* 87, no. 417 (March 1992): 178-183. https://search-proquest-com.cfc.idm.oclc.org/docview/274859555?accountid=9867.

Fox, John and Sanford Weisberg. *An {R} Companion to Applied Regression.* 2nd ed. Thousand Oaks: Sage, 2011. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of Statistical Software* 33, no.1 (2010): 1-22. http://www.jstatsoft.org/v33/i01/.

Goeman, J., R. Meijer, and N. Chaturvedi. "L1 and L2 Penalized Regression Models." last modified 27 May 2016. https://cran.r-project.org/web/packages/penalized/vignettes/penalized.pdf.

Government of Canada. "Ranks and appointment." Last modified 19 March 2018. https://www.canada.ca/en/department-national-defence/services/military-history/history-heritage/insignia-flags/ranks/rank-appointment-insignia.html.

Grant, Trista and Canada. Director Military Employment Policy. Attrition and Retention in the Canadian Forces: A Demographic Study of the 10 to 22 YOS Cohort, Consolidated Survey Results, and some Suggestions for Retention Strategies. Ottawa: Director Military Employment Policy, 2002.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer Publishing, n.d.

Hayes, Tracy Machelle. "Demographic Characteristics Predicting Employee Turnover Intentions." Order No. 3728489, Walden University, 2015. https://search-proquest-com.cfc.idm.oclc.org/docview/1734042628?accountid=9867.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. New York: Springer Publishing, 2013.

Jonge, Edwin de and Mark van der Loo. "editrules: Parsing, Applying, and Manipulating Data Cleaning Rules." R package version 2.9.0. Last accessed 20 April 2018. https://CRAN.R-project.org/package=editrules

Kowarik, Alexander and Matthias Templ. "Imputation with the R Package VIM." *Journal of Statistical Software* 74, no. 7 (20 April 2018): 1-16. doi:10.18637/jss.v074.i07.

Knowles, Jared E. "eeptools: Convenience Functions for Education Data." R package version 1.1.1. Last accessed 20 April 2018. https://CRAN.R-project.org/package=eeptools.

Kuhn, Max with contributions from Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, Zachary Mayer, Brenton Kenkel, the R Core Team, Michael Benesty, Reynald Lescarbeau, Andrew Ziem, Luca Scrucca, Yuan Tang, Can Candan and Tyler Hunt. "caret: Classification and Regression Training." R package version 6.0-79. Last accessed 20 April 2018. https://CRAN.R-project.org/package=caret.

Lee, Jennifer E. C., Donald R. McCreary, and Martin Villeneuve. "Prospective Multifactorial Analysis of Canadian Forces Basic Training Attrition." *Military Medicine* 176, no. 7 (July 2011): 777-84. https://search-proquest-com.cfc.idm.oclc.org/docview/876042799?accountid=9867.

Little, R.J.A. "Missing data adjustments in large surveys (with discussion)." *Journal of Business Economics and Statistics,* no. 6 (1988): 287-301.

Lopez-Raton, Monica, Maria Xose Rodriguez-Alvarez, Carmen Cadarso Suarez, and Francisco Gude Sampedro. "OptimalCutpoints: An R Package for Selecting Optimal Cutpoints in Diagnostic Tests." *Journal of Statistical Software* 61, no. 8 (2014): 1-36. http://www.jstatsoft.org/v61/i08/.

Michaud, Kathy and Canada. Director General Military Personnel Research and Analysis. The Canadian Forces Exit Survey 2005-2008: Descriptive Analysis. Ottawa: Director General Military Personnel Research & Analysis, 2011.

Microsoft Corporation and Steve Weston. "doParallel: Foreach Parallel Adaptor for the 'parallel' Package." R package version 1.0.11. Last accessed 20 April 2018. https://CRAN.R-project.org/package=doParallel.

Milton-Bache, Stefan and Hadley Wickham. "magrittr: A Forward-Pipe Operator for R." R package version 1.5. Last accessed 20 April 2018. https://CRAN.R-project.org/package=magrittr

Moynihan, Donald P. and Noel Landuyt. "Explaining Turnover Intention in State Government: Examining the Roles of Gender, Life Cycle, and Loyalty." *Review of Public Personnel Administration* 28, no. 2 (2008): 120-143. https://doi.org/10.1177/0734371X08315771.

Office of the Auditor General of Canada. "Canadian Armed Forces Recruitment and Retention – National Defence." Last accessed 10 March 2018. http://www.oag-bvg.gc.ca/internet/English/parl_oag_201611_05_e_41834.html#.

Okazawa, Stephen. *Measuring Attrition Rates and Forecasting Attrition Volume.* DRDC-CORA-TM-2007-02. Ottawa: Centre for Operational Research and Analysis, January 2007. http://cradpdf.drdc.gc.ca/PDFS/unc66/p527519.pdf.

Otis, Nancy and Michelle Straver. *Review of Attrition and Retention Research for the Canadian Forces.* DRDC CORA 2008-030. Ottawa: Defence Research and Development Canada Centre for Operational Research and Analysis, October 2008. http://cradpdf.drdc.gc.ca/PDFS/unc78/p530400.pdf

Penney, Christopher. *Modelling Voluntary Attrition of Civilian Personnel in the Department of National Defence.* DRDC-CORA-CR-2013-063. Ottawa: Centre for Operational Research and Analysis, April 2013.

Penney, Christopher E. *A Survival Analysis of ADM (Materiel) Workforce Attrition*. DRDC-RDDC-2016-R162. Ottawa: Defence Research and Development Canada – Centre for Operational Research, October 2016. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc247/p804591_A1b.pdf.

Pitts, David, John Marvel, and Sergio Fernandez. "So Hard to Say Goodbye? Turnover Intention among U.S. Federal Employees." Public Administration Review 71, no. 5 (2011): 751-760. http://onlinelibrary.wiley.com.cfc.idm.oclc.org/doi/10.1111/j.1540-6210.2011.02414.x/abstract.

Pollack, Lance M., Cherrie B. Boyer, Kelli Betsinger, and Mary-Ann Shafer. "Predictors of One-Year Attrition in Female Marine Corps Recruits." *Military Medicine* 174, no. 4 (April 2009): 382-391. https://search.proquest.com/docview/217051260?accountid=9867.

Privy Council Office (PCO) of Canada. *Increasing recruitment of women into the Canadian Armed Forces.* Ottawa: PCO Innovation Hub, June 2017. https://www.canada.ca/content/dam/pco-bcp/documents/pdfs/CAF_ENG.pdf.

R Core Team. "R: A Language and Environment for Statistical Computing." Last accessed 5 May 2018. https://www.R-project.org/.

Rosemarin, Roy. "Is Multicollinearity a Problem with Gradient Boosted Trees?." Last modified January 2013. https://www.quora.com/Is-multicollinearity-a-problem-with-gradient-boosted-trees.

Ryan, John F., Richard Healy, and Jason Sullivan. "Oh, Won't You Stay? Predictors of Faculty Intent to Leave a Public Research University." *Higher Education* 63, no. 4 (April 2012): 421-437. doi:http://dx.doi.org/10.1007/s10734-011-9448-5. https://search.proquest.com/docview/927705340?accountid=9867.

Schomaker, Michael and Christian Heumann. "Model Averaging and Model Selection after Multiple Imputation using the R-package MAMI." Last modified 12 December 2017. http://mami.r-forge.r-project.org/.

Shapiro, S. and M. Wilk. "An Analysis of Variance Test for Normality." *Biometrika*, Volume 52, Issue 3-4 (1 December 1965): 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Sing T., Sander O., Beerenwinkel N., and Lengauer T. "ROCR: Visualizing Classifier Performance in R." *Bioinformatics* 21, no. 20 (2005): 78-81. http://rocr.bioinf.mpi-sb.mpg.de.

Smith, David G. and Judith E. Rosenstein. "Gender and the Military Profession: Early Career Influences, Attitudes, and Intentions." *Armed Forces & Society* 43, no. 2 (2016): 260-279. https://doi.org/10.1177/0095327X15626722.

Sokri, Abderrhamane. *A Socio-economic Analysis of Military Attrition: The Case of Non-Commissioned Members of the Canadian Armed Forces*. DRDC CORA TM 2013-088. Ottawa: Defence Research and Development Canada – Centre for Operational Research, June 2013. http://cradpdf.drdc-rddc.gc.ca/PDFS/unc263/p537714_A1b.pdf.

Stewart, James B., and Juanita M. Firestone. 1992. "Looking for a Few Good Men: Predicting Patterns of Retention, Promotion, and Accession of Minority and Women Officers." *American Journal Of Economics And Sociology* 51, no. 4: 435-458. http://search.ebscohost.com/login.aspx?direct=true&db=eoh&AN=0273184&site=ehost-live.

Vink, Gerko, Laurence E. Frank, Jeroen Pannekoek, and Stef van Buuren. "Predictive Mean Matching Imputation of Semicontinuous Variables." Statistica Neerlandica 68, no. 1 (2014): 61-90.

Wickham, Hadley. *ggplot2: Elegant Graphics for Data Analysis.* New York: Springer-Verlag, 2009.

Wickham, Hadley, Romain Francois, Lionel Henry, and Kirill Müller. "dplyr: A Grammar of Data Manipulation." R package version 0.7.4. Last accessed 20 April 2018. https://CRAN.R-project.org/package=dplyr

Wickham, Hadley. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40, no. 1 (2011): 1-29. http://www.jstatsoft.org/v40/i01/.

Wikipedia. "Cohen's Kappa." Last accessed 11 April 2018. https://en.wikipedia.org/wiki/Cohen%27s_kappa.

Wikipedia. "Evaluation of Binary Classifiers." Last accessed 21 April 2018. https://en.wikipedia.org/wiki/Evaluation_of_binary_classifiers.

Wikipedia. "Logistic Regression." Last accessed 7 April 2018. https://en.wikipedia.org/wiki/Logistic_regression.

Wikipedia, “Regularization (mathematics),” last accessed 11 April 2018, https://en.wikipedia.org/wiki/Regularization_(mathematics).

Wolfe, Jessica, Kiban Turner, Marie Caulfield, Tamara L. Newton, and et al. "Gender and Trauma as Predictors of Military Attrition: A Study of Marine Corps Recruits." *Military Medicine* 170, no. 12 (December 2005): 1037-43. https://search.proquest.com/docview/217065022?accountid=9867.

Wolfram Mathworld. “Bonferroni Correction.” Last accessed 21 April 2018. http://mathworld.wolfram.com/BonferroniCorrection.html.

David H. Wolpert, “Stacked Generalization,” *Neural Networks* 5, no. 2 (1992): 241-259.

Zou, Hui, and Trevor Hastie. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 67, no. 2 (2005): 301-20. http://www.jstor.org/stable/3647580.