Katie McCrum, William Reynolds, and Brenden Siekman
5 December 2022

# OR568: Applied Predictive Analytics Project
# Predicting Month to Month Attrition of U.S. Army Captains

**Background:**

Ensuring the U.S. Army is sufficiently staffed across all ranks is an ongoing effort for our all-volunteer force. Not only must the Army recruit adequate numbers, it must retain Soldiers at every level in order to have developed and experienced leaders. Future attrition, or the departure of a service member, directly influences how the Army makes decisions with regards to personnel. The Office of the Deputy Chief of Staff for Personnel, Department of the Army, also known as the "HQDA G-1," is responsible for manpower management and is deeply invested in a model that can predict attrition. In order to help inform future decisions regarding recruiting, promoting, and retaining people, the G-1 needs a tool that will allow them to accurately estimate approximately how many personnel will leave service in any given month, 12 to 24 months in the future. The G-1 is especially interested in service members in the rank of Captain, as they encompass the largest percentage of Officers and also have the highest attrition rate.

**Objective:**

Our objective is to determine which factors are the best predictors of future attrition, how current economic factors may affect an individual's decision to continue service, and to develop a model that can better predict future monthly attrition rates than the G-1's current model.

**Methods:**

Our team broke the problem down into two types of models: individual and summary models. The individual models examine the factors predicting individual loss, while the summary models group the data by month to predict the number of losses in a particular month. The latter type interests the client more. While G-1 is not particularly interested in predicting the loss of a particular Officer, the results of the individual models helped indicate which are the more significant factors when predicting loss, which we can then incorporate into a group model through an aggregation of selected factors for a particular month.

**Data:**

The data used in this project comes from the HQDA G-1, the Bureau for Labor Statistics (BLS), and the Organization for Economic Cooperation and Development (OECD). The original dataset from the Army G-1 includes 54 predictors (cleaned and transformed into 37 predictors), 3,392,489 observations over 75,785 unique IDs, and consists primarily of nominal and qualitative data at the individual service member level.[1] Predictors can be grouped into several categories: career data, personal data, familial data, education data, and loss data. Career data encompasses information regarding an individual's military career to include basic branch, number of deployments, dates of rank and accession, and additional skills. Personal and familial data includes general demographic data, marital status and number of dependents, special considerations, and spouse's branch of service when applicable. Education data identifies if the individual is currently enrolled in a degree program as well as their most advanced level of education. Loss data provides dates for initiated separation paperwork and projected loss date where applicable. Dropped variables include predictors that were redundant (i.e. both INV_CF1AOC1 and INV_CRMGOF provide information on the individual's branch therefore only one is needed) or convoluted (INV_NBRMOA provides the number of months spent overseas but does not indicate whether those months are accompanied assignments or deployments). Redundant variables, prior to being dropped, were used to assist in filling in missing data in their partner predictors if applicable. Further processing and mutations of data are expounded on within each model section. The dataset provided includes monthly numbers for each unique ID still remaining in the Captain population (i.e. not attrited out by natural or induced causes such as promotion or separation) from November 2008 to October 2022.

[1] Office of the Deputy Chief of Staff for Personnel, Department of the Army. *Army Captain Attrition Dataset.* (October 2022).

The datasets from the BLS originate from the monthly Job Openings and Labor Turnover Survey (JOLTS)[2] and the Labor Force Statistics from the Current Population Survey (CPS)[3] at the national level. They include job vacancies and unemployment numbers (non-farming, seasonally adjusted, age 16 and up) which we used to calculate the vacancy to unemployment ratio (V/U Ratio). We used the monthly consumer confidence index (CCI) from the OECD which provides an indication of the populations' confidence in their future economic situation six months from the survey date.[4] The data set includes monthly numbers from January 2000 to September 2022, however only data that coincides with our G-1 dataset (2008-2022) is used in this analysis. The full range of variable names, types, descriptions, and sample subsets are outlined in Appendix A (Figures A1 and A2).

**Data Cleaning and Preprocessing:**

Our first step was to identify empty values that are actually missing data or if the empty values indicated a "blank" category as a factor type. For the actual missing values, we used a variety of techniques. Because each unique ID had multiple entries, we could isolate IDs with missing values and fill some of the missing data that were present in other rows for the same ID. Some fields were subsets of other fields. For example, INV_CRMGOF is a more general piece of information than INV_CF1AOC1. We can fill in missing INV_CRMGOF data if we have INV_CF1AOC for the same ID. Some data was missing at random for all ID values. If there were no good generalizations we omitted the data, as we had a sufficient number of data points while omitting some incomplete rows. If we were grouping data into fewer categories, we typically grouped missing values with the majority value.

For summary models, the data was aggregated while grouped by date of data collected (DATA_DT) using dplyr[5] library with Rstudio. In summarizing data, we summarized by counting rows (e.g. n_distinct(INV)), counting types (e.g. SOC == ROTC) or non-blank (!is.na(LOSS_DT)). To obtain ratio values for each, we mutated the row by dividing by the inventory (INV). These allow predictors for linear and time series models.

**Exploratory Data Analysis:**

Initial exploratory analysis compared the population ratio versus the loss population of several predictor variables. Analysis of the source of commission revealed that officers enter the military via the United States Military Academy (USMA), the Reserve Officers' Training Corps (ROTC), Officer Candidate School (OCS), and direct commission (in this data they are labeled as "other"). Approximately 70% of the Army commissions from ROTC- the majority from a scholarship program), 17% from USMA, and the remaining 13% from both OCS and direct commission. Initial analysis indicates, especially after 2015, that source of commission is not a blatant indicator of loss (Figure C1, Appendix C). Career emphasis splits individuals into basic branch, functional area, and acquisition corps, with the majority remaining in their basic branch. For this discussion we will include the acquisition corps under the title of functional area as their accession process is the same. In almost all circumstances, officers can not make the switch from their basic branch to a functional area until they have completed their key developmental assignment as a Captain, typically around the six to eight year mark. These officers choose this career change into a more technical area of concentration of their specific interest and in many cases are sent to additional schooling resulting in additional service commitments. Likely due to the fact that an individual in a functional area makes the decision to move to the more focused area of work coupled with the potential for an additional service commitment, we see a

---

[2] Bureau of Labor Statistics. *Job Opening and Labor Turnover (JOLTS) Dataset.* (September 2022). https://data.bls.gov/PDQWeb/jt.

[3] Bureau of Labor Statistics. *Current Population Survey Dataset.* (September 2022). https://data.bls.gov/PDQWeb/ln.

[4] Organization for Economic Cooperation and Development. *Consumer Confidence Index Survey Dataset.* (October 2022). https://data.oecd.org/leadind/consumer-confidence-index-cci.htm.

[5] Wickham H, François R, Henry L, Müller K (2022). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

smaller loss rate observed in this population (Figure C2). In exploring gender effects, women attrit as approximately 20% of the loss population however we observe a comparable loss rate among men and women (between 38-40%) when accounting for percent of total population (Figure C3). Marital status appears to play a bigger role in an individual's decision to remain in the Army- approximately 66% of the Captain population is married and married individuals comprise 56% of the loss population. However, 51% of our non-married (includes divorces, widowed, annulled, single) population attrit as opposed to 33% of the total married population. Additionally, when controlling for those married to another service member versus those who are not, the rate remains almost equal (Figure C4).

Both CCI and V/U ratio were transformed into indicator variables for the purpose of our analysis based on the following cutoffs: a value of 100 or greater for the CCI is considered to be a "positive economic outlook" and a V/U ratio value of one or greater indicates available jobs are greater than the number of unemployed persons (an unemployed person is defined as someone who has searched for a job in the last 30 days). Our hypothesis is that when we observe a positive economic outlook and/or a V/U ratio greater than one, losses will increase as stability in the civilian workforce is more likely. We graphed the CCI and V/U ratio levels from 2008-2020, at first glance a direct relationship is not apparent (Figure C5, Figure C6).

**Current model (G-1):**

G-1 is currently using a weighted-moving-average time series model, which predicts the upcoming year's months loss rates on the previous three years' months. The weights are 0.5, 0.3, and 0.2 for the previous year, two years previous and three years previous, respectively. The mean absolute percent error (MAPE) of the current model is ~8% with the forecasted loss typically slightly above actual loss rate (Figure C7).

**Logistic Regression Model:**

To develop the logistic regression model, we had to look at an appropriate grouping of the data. Each Captain, identified by ID number, had a data entry (row) for each month as a Captain. In the latest date for each ID, if the Captain had a loss date or loss type, then the Captain was a loss. If not, then the Captain was not a loss. Therefore we grouped the dataframe for the individual analysis model by each unique ID's latest data entry. This allowed us to predict loss for each ID.

To develop useful predictors for the individual model, we had to transform several of the columns into new columns with usable values or factors. Many factors had several categories that we shrunk to either binary indicators or a smaller number of grouped factors. As an example of converting to a binary indicator, INV_PHYC indicated one of several categories of physical condition. If an ID was any category below the top category, then the new indicator was TRUE as having a sub-optimal physical category. For an example of combining factors, there were 83 unique assignment considerations (INV_ASCO1, INV_ASCO2, and INV_ASCO3). We ran a logistic regression predicting loss only using assignment considerations and sorted the resulting coefficients for each assignment consideration by coefficient and p-value. We could then group the assignment considerations into "strong positive", "strong negative", "weak positive", "weak negative", and "no assignment consideration," based on the strength of the coefficient while screening assignment considerations with p-values over 0.05. There were two cases with a clear causal negative effect that we grouped into "strong negative" without a significant p-value, however. These considerations were "mandatory separation," and "documented sex related offense." Both of these had a low enough frequency to have an insignificant p-value, but we can infer that officers with these assignment considerations will be losses.

For other data columns, we examined a change between an ID's first entry and an ID's last entry to surmise a new factor, such as a change in marital status or change in branch. One factor in particular, submission of separation paperwork, is a strong predictor of loss, but is not useful to examine at the latest data entry, as all officers are required to submit separation paperwork to exit the Army. For this factor in particular, we looked at the ID's earliest entry and created a binary indicator for if the officer had already submitted separation paperwork. In this way, we could factor in the separation paperwork's effect without overfitting.

We also had to control for predictors that were conflated with time. These included predictors such as age or number of deployments. While correlated, the causal relationship for these predictors is the reverse of what we are examining. Those Captains who are older will have naturally stayed in the system longer and, thus have a lower prevalence of loss. Therefore, we needed to examine age at one moment in time, specifically when the officer was promoted to Captain, by looking at the unique ID's earliest entry. We can then convert this age into a factor that will not be conflated with time at the ID's latest entry (is28: indicator for if ID is 28 or older at promotion to CPT).

Through a combination of the techniques described above, we transformed the 54 original data columns into 33 predictive factors. See Appendix B for the full list of transformations. We kept these 33 factors based on a logistic regression model using only that factor, selecting individually statistically significant factors for the model that combined all selected factors. Once combined, some of these predictors did lose statistical significance, however.

To develop the combined model, we randomly split the data frame into a training and test data, using a 70-30 percent split. Our logistic regression model predicted loss on the test data with an 95% CI accuracy of [0.756, 0.767] and a Kappa of 0.485 at a probability cutoff of 0.5. The confusion matrix and statistics for this model are depicted to the right. The receiver operating characteristic (ROC) curve depicts the sensitivity vs. specificity tradeoff for selected logistic model cutoff values. Our ROC has an area under the curve (AUC) of 0.824. See Appendix D for the ROC plot.

```
Confusion Matrix and Statistics

          0     1
  0 11791  3241
  1  2176  5512

               Accuracy : 0.7616
                 95% CI : (0.756, 0.7671)
    No Information Rate : 0.6147
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.4849

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6297
            Specificity : 0.8442
         Pos Pred Value : 0.7170
         Neg Pred Value : 0.7844
             Prevalence : 0.3853
         Detection Rate : 0.2426
   Detection Prevalence : 0.3384
      Balanced Accuracy : 0.7370

       'Positive' Class : 1
```

These metrics indicate our model performs significantly better than random chance or assuming the base rate at predicting the attrition of a selected Captain. With McFadden's R-Squared value of only 0.2497, we can see that there are several other considerations that influence an individual's decision to stay in the Army beyond our examined predictors.

Examining which factors are most significant in predicting individual loss or retention, we can look at our sorted table of coefficients from the logistic regression below. The coefficients in the table have been transformed from log(odds) to probability of loss. The strongest indicators of retention are highest in the table with the lowest probabilities, while the strongest predictors of loss are at the bottom of the table with the highest probabilities. The three strongest loss predictors of Captains are the three branches that combine into the Logistics branch upon completion of mandatory professional education for Captains, the Captains Career Course. Since the "branch" factor looks at the Captain's branch at their latest data entry, Captains with one of the three branches "OD", "TC", or "QM" indicate Captains that have not completed the career course. These three indicators are not particularly useful, as it is better to look at the "early branches" of logistic Captains since the ones retained will have switched branches to "LG." It is useful in that it informs us that completion of the career course for logisticians indicates a lower probability of loss. Other top five strongest predictors of loss include: Captains who have separation paperwork submitted upon promotion to Captain, Captains commissioning via OCS (both from prior enlisted personnel #2 and from college #5), Captains who have a "strong negative" assignment consideration, and Captains in the Military Police branch. The top five significant factors influencing Captain retention include: Captains with a Stabilized Tour, Captains pursuing education, Captains branched Acquisition Corps, and Captains with either a "strong positive" additional skill identifier or with a "strong positive" assignment consideration.

**Logistic Regression Coefficients Sorted by Ascending Probability of Loss**

| Predictor | coef | p-value |
|---|---|---|
| earlyBranch | 1.00E-04 | 0.8862 |
| StabTourTRUE | 0.0967 | 0.0000 |
| edu_compFALSE | 0.1955 | 0.0000 |
| branchAC | 0.2009 | 0.0000 |
| SPosASITRUE | 0.2548 | 0.0000 |
| SPosASCOTRUE | 0.2805 | 0.0000 |
| branchPO | 0.2899 | 0.0000 |
| branchFunc | 0.2905 | 0.0000 |
| branchSF | 0.3029 | 0.0000 |
| branchCA | 0.3061 | 0.0000 |
| DODSPSTRUE | 0.3107 | 0.0000 |
| earlyBranchLG | 0.3109 | 0.0000 |
| edu_highestDOC | 0.3153 | 0.0000 |
| is28TRUE | 0.3231 | 0.0000 |
| earlyBranchMP | 0.3402 | 0.0040 |
| branchCY | 0.3557 | 0.0044 |
| edu_highestMAST | 0.3686 | 0.0000 |
| WNegASCOTRUE | 0.3725 | 0.0002 |
| (Intercept) | 0.4005 | 0.0027 |
| WPosASITRUE | 0.4021 | 0.0000 |
| earlyBranchAV | 0.4215 | 0.1208 |
| kidsTRUE | 0.4232 | 0.0000 |
| INV_REDCATN | 0.438 | 0.0000 |
| earlyBranchMI | 0.4444 | 0.0441 |
| branchFI | 0.4457 | 0.4652 |
| INV_REDCATH | 0.4526 | 0.0000 |
| married_initialFALSE | 0.4531 | 0.0001 |
| INV_REDCATA | 0.4581 | 0.0001 |

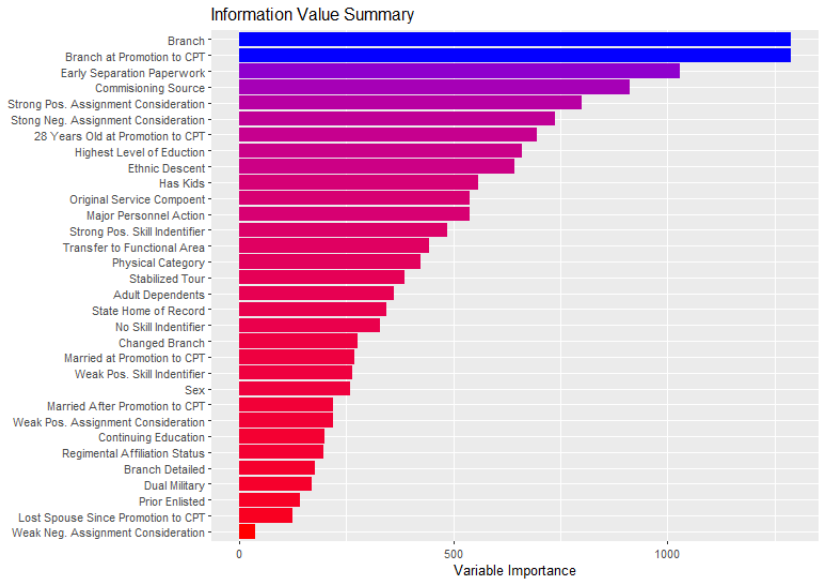
| Predictor | coef | p-value |
|---|---|---|
| INV_REDCATT | 0.4618 | 0.2357 |
| SOCOTHER | 0.4728 | 0.0746 |
| lost_spouseTRUE | 0.4752 | 0.1604 |
| change_branchTRUE | 0.4778 | 0.1463 |
| VTIP_funcTRUE | 0.4785 | 0.5340 |
| INV_SEXF | 0.4824 | 0.0365 |
| NoASITRUE | 0.4886 | 0.0751 |
| priorTRUE | 0.4893 | 0.5493 |
| SOCROTC-NON | 0.4901 | 0.2212 |
| earlyBranchCY | 0.4904 | 0.8787 |
| earlyBranchFA | 0.5027 | 0.9321 |
| got_marriedTRUE | 0.508 | 0.5203 |
| earlyBranchEN | 0.5149 | 0.7049 |
| branchCM | 0.5149 | 0.7816 |
| INV_REDCATX | 0.5151 | 0.4499 |
| earlyBranchAG | 0.52 | 0.6075 |
| earlyBranchSC | 0.528 | 0.3299 |
| branchAD | 0.529 | 0.5657 |
| HOR_STXDIF | 0.532 | 0.0000 |
| branchAR | 0.5328 | 0.3931 |
| earlyBranchSF | 0.5343 | 0.1405 |
| earlyBranchFunc | 0.536 | 0.2951 |
| active_origFALSE | 0.5361 | 0.0000 |
| branchAG | 0.5363 | 0.3824 |
| earlyBranchAC | 0.5501 | 0.6710 |
| RGAASTATRUE | 0.5528 | 0.0000 |
| branchSC | 0.5533 | 0.1270 |
| branchIN | 0.5554 | 0.0870 |

| Predictor | coef | p-value |
|---|---|---|
| branchFA | 0.5636 | 0.0861 |
| earlyBranchAD | 0.5648 | 0.1540 |
| earlyBranchAR | 0.5673 | 0.0293 |
| SOCUSMA | 0.5759 | 0.0000 |
| MPATYP_TAFALSE | 0.5805 | 0.0000 |
| earlyBranchCM | 0.5888 | 0.0597 |
| earlyBranchCA | 0.5906 | 0.0048 |
| branchMI | 0.5907 | 0.0067 |
| adult_depsFALSE | 0.5976 | 0.0000 |
| bDetailTRUE | 0.6013 | 0.0000 |
| earlyBranchFI | 0.6041 | 0.1633 |
| earlyBranchPO | 0.6204 | 0.0049 |
| earlyBranchOD | 0.6214 | 0.0001 |
| WPosASCOTRUE | 0.6282 | 0.0000 |
| branchEN | 0.633 | 0.0020 |
| branchAV | 0.6477 | 0.0051 |
| earlyBranchTC | 0.6574 | 0.0000 |
| edu_highestASSO | 0.6582 | 0.1734 |
| earlyBranchQM | 0.6644 | 0.0000 |
| PHYCsubA | 0.6935 | 0.0000 |
| SOCOCS_COL | 0.698 | 0.0000 |
| branchMP | 0.698 | 0.0004 |
| SNegASCOTRUE | 0.7951 | 0.0000 |
| SOCOCS_INS | 0.7958 | 0.0000 |
| earlySep_ppwTRUE | 0.8072 | 0.0000 |
| branchQM | 0.8589 | 0.0000 |
| branchTC | 0.8857 | 0.0000 |
| branchOD | 0.8955 | 0.0000 |

## Random Forest:

Using the same predictors as in the logistic regression, we also constructed a random forest model. The random forest performed slightly better on the out of sample data with 95% CI for accuracy of [0.7713, 0.7822] and a Kappa of 0.5095. A closer look at the performance reveals a lower sensitivity, but higher specificity. This indicates that the random forest is better at correctly predicting which Captains will be retained (lower false positive rate), but lower performance in correctly predicting which Captains will be losses (higher false negative rate). If the difference between type I and type II errors does not matter, the random forest is the better choice. If minimizing type II errors is more important, then the logistic regression does a better job. In the chart below we can see which predictors are the most important when predicting loss according to the random forest model.

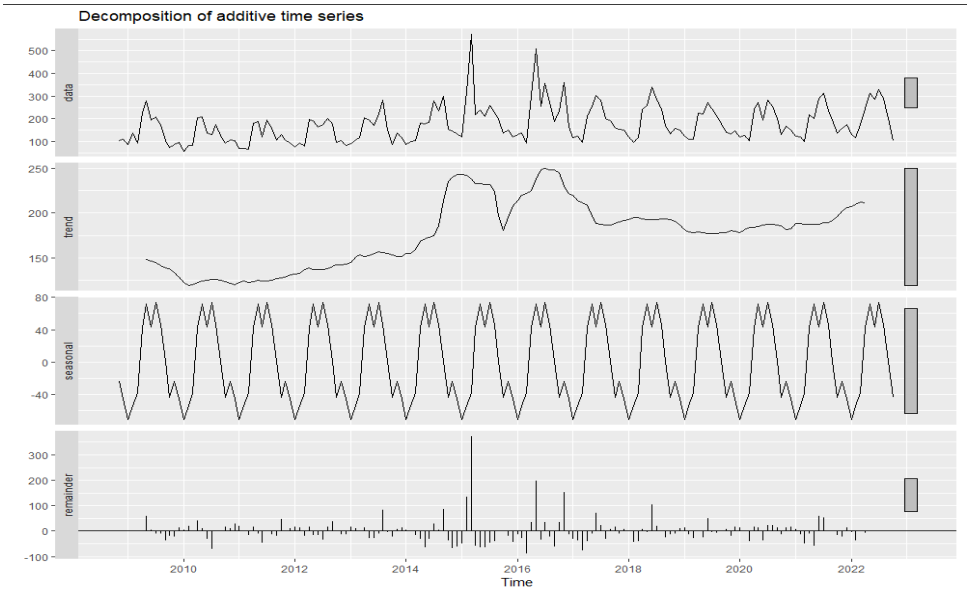
```
Confusion Matrix and Statistics


       0    1
  0 12351  3456
  1  1616  5297

               Accuracy : 0.7768
                 95% CI : (0.7713, 0.7822)
    No Information Rate : 0.6147
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5095

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6052
            Specificity : 0.8843
         Pos Pred Value : 0.7662
         Neg Pred Value : 0.7814
             Prevalence : 0.3853
         Detection Rate : 0.2331
   Detection Prevalence : 0.3043
      Balanced Accuracy : 0.7447

       'Positive' Class : 1
```




## Summary Model Comparison and Selection:

**Time Series Analysis:**

The summarized data is a Time Series due to being collected monthly over a course of time. As such, some assumptions about linear models do not hold true. Time series naturally have a high correlation data due to seasonality and overall trend. To analyze the time series trends and seasonality, we decomposed using the R time-series function (ts) and found a strong seasonality with the months over years with a slight overall trend. This is likely caused by the fact that most losses occur in the summer months as they are tied to the month that the Officer commissions which usually occurs from May to August.




Further decomposition did not reveal any other seasonalities. The decomposition also showed a large spike in 2015 and 2016 which coincided with a change in Army Manpower strength. We controlled for this with a "badYear" indicator variable in the summary models.

**Time Series Models:**

With aggregation and time series, we can look into linear models with Time Series Models. Due to the high correlation of time series, we must utilize other metrics than R2 and sum of squared errors. We chose to use MAPE as the primary statistic of comparison but also looked at Akaike's Information Criterion (AIC), AIC corrected (AICc) and the Bayesian Information Criterion (BIC)[6]. In all cases we wish the model to have low numbers for each statistic.

Additionally, we need to control for Trend and Seasonality. Trend takes as a single modifier or beta value. To account for seasonality, we could factor by month, produce fourier functions[7] or use advanced techniques like ARIMA and TBATS.[8] We first looked at losses as an aggregate but found better performance with treating losses as a rate divided by the inventory (INV).

We train each model with the first 14 years of information and utilize the last year of data as the testing data. Some general trends that we saw

The factoring by month model shows the below formula as a time series linear model (TSLM):

$$loss_{rate}(t) = \beta_0 + \beta_{Trend}(t) + \beta_{Season} Month(t) + Ind.(badYear)$$

The TSLM with Fourier Seasonality followed this formula:

---

[6] Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, "Evaluating the Regression Model," 2nd edition, OTexts: Melbourne, Australia. OTexts. https://otexts.com/fpp2/regression-evaluation.html.

[7] Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, "Some useful predictors," 2nd edition, OTexts: Melbourne, Australia. OTexts, https://otexts.com/fpp2/useful-predictors.html.

[8] Hyndman, Rob J. "Forecasting Weekly Data." *Hyndsight Blog*, 5 Mar. 2014, https://robjhyndman.com/hyndsight/forecasting-weekly-data/.

$$loss_{rate}(t) = \beta_0 + \beta_{Trend}(t) + \sum_{n=1}^{K} \left(a_n cos(\frac{n\pi t}{12}) + b_n sin(\frac{n\pi t}{12}) + Ind.(badYear)\right)$$

The best fitting model had a K=3 which shows 3 Fourier functions to model over each year.

The ARIMA Model found a best fit with (0,1,2) errors. This has a similar formula with the Fourier TSLM but with no intercept coefficient and only one Fourier Season:

$$loss_{rate}(t) = \beta_{Trend}(t) + a_n cos(\frac{n\pi t}{12}) + b_n sin(\frac{n\pi t}{12}) + Ind.(badYear)$$

The TBATS model represents five techniques in a single model[9].

$$y_t = \ell_{t-1} + b_{t-1} + s_{t-1} + \alpha d_t \quad (1)$$

$$b_t = b_{t-1} + \beta d_t \quad (2)$$

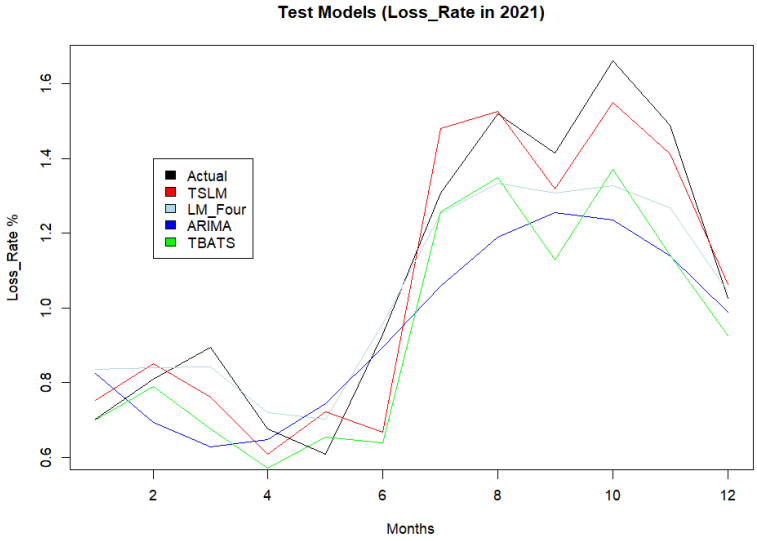
$$s_t = \sum_{j=1}^{8} s_{j,t} \quad (3)$$

$$s_{j,t} = s_{j,t-1} \cos\left(\frac{2\pi jt}{52.18}\right) + s^*_{j,t-1} \sin\left(\frac{2\pi jt}{52.18}\right) + \gamma_1 d_t \quad (4)$$

$$s^*_{j,t} = -s_{j,t-1} \sin\left(\frac{2\pi jt}{52.18}\right) + s^*_{j,t-1} \cos\left(\frac{2\pi jt}{52.18}\right) + \gamma_2 d_t, \quad (5)$$ [10]

The first formula (1) shows loss_rate as y(t) with a lagging variable $l_{t-1}$, a Box-Cox transformation (2), seasonality $s_t$ (3), two fourier transformations with exponential smoothing parameters ($\gamma_n$) (4 & 5), and finally an alpha value to [11]. Using the tbats R function, we found a best fit with Box-Cox transformation with value of 0.101, no overall trend (beta), no damping, gamma values of (-0.00054,0.00996), and a .4014694 moving average coefficient.

**Evaluating models:**




Utilizing the last year of data, we tested each model compared MAPE, AIC, AICC, and BIC. Overall, the model that performed the best was the TSLM with Fourier Seasonality.

---

[9] "Time-Series Forecasting Using TBATS Model." Big Data Analytics. Tenth Planet, November 23, 2020. https://blog.tenthplanet.in/time-series-forecasting-tbats/.

[10] Hyndman, Rob J. "Forecasting Weekly Data."

[11] Hyndman, Rob J. "Forecasting time series with complex seasonal patterns using exponential smoothing." *Hyndsight Blog*, 31 Dec. 2011, https://robjhyndman.com/publications/complex-seasonality/.

| STAT | TSLM | TSLM w/ Fourier | ARIMA | TBATS |
|---|---|---|---|---|
| **MAPE** | 9.982113 | 9.509152 | 16.41882 | 13.94898 |
| **AIC** | 25.29856 | 25.29856 | 52.51 | 272.082 |
| **AICc** | 29.24059 | 29.24059 | 52.92 | 272.3487 |
| **BIC** | 70.94994 | 70.94994 | 67.7 | 1719.417 |

**Adding Predictors:**

The fourier series provided the best model but due to coding issues, we were only able to make the TSLM with predictors. We summarized the training data with 19 predictors (See figure D.2) with month seasonality for a total of 32 variables. This produced a model above with higher statistics and only a few predictors being significant. (ROTC, OCS, and Ethnic High Risk Groups). This gives the impression that not using the predictors is better for the forecast. We also tried using Economic Variables only, turning off the Bad Year indicator variable which produced a higher MAPE. This kept the MAPE around 19% until we lagged the variables by six months. This reduced our sample size but showed lagged CCI to have a negative coefficient (less losses) with significance at a p-value of 4.30e-05 and lagged V/U index with a positive coefficient (more losses) at p-value of .0512. This shows promise in economic variables.

| STAT | TSLM w/ Fourier (Prev. Best) | TSLM w/ ALL Predictors | w/ Significant Predictors | w/ Economic Only |
|---|---|---|---|---|
| **MAPE** | 9.51 | 15.35 | 10.57 | 11.34 |
| **AIC** | 25.29 | 492.70 | 497.1 | 498.6416 |
| **AICc** | 29.24 | 513.8 | 502.7 | 503.671 |
| **BIC** | 70.94 | 594.6 | 551.0 | 549.6 |

**Conclusions and Future Work:**

The best model for predicting the number of losses by month is the time series model with Fourier Series Seasonality. Using this method, we achieve a mean absolute percent error of 9.51% which is higher than G-1's current model of 7.80%. However, there is indication that with proper techniques and further study, some predictors may be significant. This could be experimented further with lagging variables at various intervals. Further, using Fourier Series with this prediction seems promising as well.

Using individual loss prediction alone will not predict when a Captain will be lost. However, could use the individual framework to predict time remaining in the population for Captains. An additional predictor would be months time in grade as a Captain in addition to the other factors used in logistic regression. The new response variable would be the months remaining as a Captain. If the total time for the Captain is less than the time required for promotion, we can infer that the Captain is lost. This model would also allow us to predict in which month an individual Captain will leave the Army.

Currently the noise in the predictions of loss combined with the error of predicting when the loss will occur makes this aggregation of individual prediction less reliable than a time-series forecasting method.

## Appendix A: Data Variables, Types, Descriptions, Subset (Base Datasets)

**Figure A1:** G-1 Data

| # | Variable | Type | Description | Data Format |
|---|---|---|---|---|
| **1** | DATA_DT | date | Date Data Pulled | YYYY-MM-DD e.g 2020-01-01 |
| **2** | INV_ASCO1 | chr | Assignment Consideration One | W7,D5,W7,C7… |
| **3** | INV_ASCO2 | chr | Assignment Consideration Two | BLANK,W7,BLANK,C8… |
| **4** | INV_ASCO3 | chr | Assignment Consideration Three | BLANK,BLANK,BLANK,W7… |
| **5** | INV_ASEPDT | date | Date Of Projected Separation | NA,2016-12-01,NA…. |
| **6** | INV_ASICO2 | chr | Additional Skill Indicator 1 Commissioned | 5W,5P,R9… |
| **7** | INV_BASD | date | Basic Active Service Date | 2009-07-03,2009-06-16,2010-02-23… |
| **8** | INV_BOSD | date | Basic Officer Service Date | NA,NA,2016-07-11,NA… |
| **9** | INV_BREX | date | Branch Detail Expiration Date | 2013-11-01,NA,NA,NA… |
| **10** | INV_CF1AOC1 | chr | Career Field 1 / Area Of Concentration | 35D,11A,38A… |
| **11** | INV_CRMGOF | chr | Control Branch - Career Management Office | MI, IN, CA….. |
| **12** | INV_DOB | date | Date Of Birth | 1987-03-24,1987-08-29,1988-04-22… |
| **13** | INV_DODSPS | chr | Spouse Service Branch | BLANK,A,BLANK…. |
| **14** | INV_DTLBR | chr | Detail Branch | IN, BLANK,BLANK,AR,CM… |
| **15** | INV_ETHGRP | chr | Ethnic Group Code | X,Y,X,X,X,K… |
| **16** | INV_FSA | int | Basic Year Group | 2009,2010,2010…. |
| **17** | INV_IRISK | chr | Individual Risk Assessment Code | BLANK |
| **18** | INV_MARST | chr | Marital Status | M,S,D,M,M,M… |
| **19** | INV_MOP | chr | Manner Of Performance | BLANK |
| **20** | INV_MPATYP | chr | Major Personnel Action Type | TA,TA,AH,TA… |
| **21** | INV_NBRMOA | chr | Number Of Months Overseas | 38,9,6,8,26… |
| **22** | INV_NDEPA | int | Number Of Adult Dependents | 1,2,1,1,1…. |
| **23** | INV_NDEPNC | int | Number Of Dependent Children | 0,2,3,2,0,0… |
| **24** | INV_NOLOT | chr | Number Of Long Overseas Tours | 1,1,0,0,1,BLANK,BLANK…. |
| **25** | INV_NOSOT | chr | Number Of Short Overseas Tours | 1,1,BLANK,1,0,1…. |
| **26** | INV_NPCSCF | chr | Number Of Permanent Changes Of Station | 3,4,4,7,4,6… |
| **27** | INV_OCN_EDUCAT1 | chr | Civilian Education Category | C,C,C,C,P.... |
| **28** | INV_OCX_OSASD | date | Overseas Assignment Start Date | 2012,-10-08,NA,2019-10-24… |
| **29** | INV_ORSTT | chr | Stabilized Tour Type | BLANK |
| **30** | INV_PDOR | date | Permanent Date Of Rank | 2013-11-01,2013-11-01,2014-05-01… |
| **31** | INV_PHYC | chr | Physical Category Status Code | BLANK…A…T…F…W….W....BLANK… |
| **32** | INV_PJAOC | chr | Projected Area Of Concentration | 35D,11A,38A,92A… |
| **33** | INV_PRDVEM | chr | Primary Professional Development Emphasis | 1,3,1,1,2,1,1,1… |
| **34** | INV_ASICO1 | chr | Additional Skill Indicator 1 Commissioned | 5P,5S,2B,K9…. |
| **35** | INV_RACPOP | chr | Race/Population Group Code | C,M,C,X…. |
| **36** | INV_REDCAT | chr | Racial/Ethnic Descent Category | C,H,C,X,C,C,A…. |
| **37** | INV_RELDEN | chr | Religious Denomination | 13,62,40,13,GC,EC…. |

| | | | | |
|---|---|---|---|---|
| **38** | INV_RGAAST | chr | Regimental Affiliation Assignment Status | BLANK,A,B,B,N,B,B…. |
| **39** | INV_RGTAFF | chr | Regimental Affiliation | BLANK,CORPQM,CORPMI,CORPCY… |
| **40** | INV_SAPRDT | date | Date Separation Application Approved | 2019-05-08,NA…. |
| **41** | INV_SCOA | chr | Service Component Of Original Appointment | R,R,R,V,R,R…. |
| **42** | INV_SEX | chr | Sex/Gender Identification | M,F,M,M,M,F… |
| **43** | INV_SRECDT | date | Date Separation Application Received | 2016-12-01,NA,NA… |
| **44** | INV_OCN_CEDG-1 | chr | Civilian Education Degree | BA,BS,MA,BS,BFA… |
| **45** | INV_SREQDT | date | Date Separation Requested | 2016-03-16,NA…. |
| **46** | INV_SSUBDT | date | Date Separation Application Submitted | NA,2019-03-12,NA….. |
| **47** | INV_STBR | int | State Of Birth | 55,55,17,37,21…. |
| **48** | INV_STHRED | chr | Home Of Record - State | 55,17,37,21,26... |
| **49** | LOSS_DT | date | Expected Loss Date | NA,NA,2020-02-01…. |
| **50** | LOSS_TYPE | date | Loss Type | BLANK,NATURAL,BLANK,INDUCED…. |
| **51** | MSV_OFF_QY | int | Months In Service As An Officer | 124,123,117… |
| **52** | SOC | chr | Source Of Commision | ROTC-NON,OCS_COL,ROTC-SCH,USMA… |
| **53** | TIG_QY | int | Time In Grade In Months | 75,75,69,58,64… |

**Figure A2:** BLS/OECD Data

| # | Variable | Type | Description | Subset |
|---|---|---|---|---|
| **1** | Vacancy | num | Job Vacancy Numbers in thousands | 4447,4024,4071… |
| **2** | Unemployment | num | Unemployment Numbers in thousands | 6583,7042,7142… |
| **3** | V/U Ratio | num | Job Vacancy to Unemployment Ratio | 0.6755,0.5714,0.5700… |
| **4** | CCI | num | Monthly Consumer Confidence index | 101.093,100.698,100.5274… |

## Appendix B: Unique Values and Data Cleaning Steps for Predictors

| # | Variable | Data Cleaning Method | UNIQUE VALUES |
|---|---|---|---|
| **1** | DATA_DT | group_by(DATA_DT) → group method<br><br>Individual Method:<br>group_by(ID)>filter(DATA_DT==max(DATA_DT)) | "2008-11-01" "2008-12-01" "2009-01-01" "2009-02-01" …<br>"2022-07-01" "2022-08-01" "2022-09-01" "2022-10-01" |
| **2** | INV_ASCO1 | Group into Strong Positive, Weak Positive, Strong Negative, and Weak Negative, and None based on coefficient from glm predicting loss. Screened by p-value. | " " "A1" "A2" "A3" "A4" "A6" "A8" "A9" "B1" "B2" "B3" "B4" "B5" "B6" "B7" "B8" "B9" "C1" "C2" "C3" "C4" "C7" "C8" "C9" "D1" "D5" "D7" "D8" "D9" "E1" "E5" "F4" "F7" "F9" "G-1" "G2" "I2" "K1" "K2" "K3" "K7" "L1" "L3" "L4" "L5" "LE" "M2" "M3" "M5" "M6" "M8" "O3" "P1" "P4" "R1" "S1" "S8" "S9" "T1" "U5" "V1" "V2" "V5" "V6" "V8" "W2" "W5" "W7" "X1" "X2" |
| **3** | INV_ASCO2 | Grouped with above | " " "A3" "B5" "B8" "B9" "C1" "C3" "C4" "C7" "C8" "D1" "D5" "D8" "D9" "E2" "E5" "F1" "F3" "F4" "F7" "F8" "F9" "G-1" "G2" "H2" "I2" "K2" "K3" "K4" "K6" "L1" "L3" "L4" "L5" "L8" "LE" "M2" "M3" "M5" "M6" "M7" "M8" "O3" "O4" "P1" "P4" "R1" "S1" "S2" "S8" "S9" "T1" "U1" "U5" "V1" "V2" "V5" "V8" "W2" "W4" "W5" "W7" "W9" "X1" "X2" |
| **4** | INV_ASCO3 | Grouped with above | " " "C3" "C4" "C7" "C8" "D5" "D8" "D9" "E2" "E5" "F4" "F5" "F7" "F9" "G-1" "G2" "H2" "I2" "K3" "K4" "K5" "K6" "L1" "L3" "L4" "L8" "LE" "M2" "M3" "M6" "M7" "M8" "O3" "O4" "P1" "P4" "R1" "S8" "S9" "U1" "U5" "V1" "V2" "V5" "V6" "V8" "W2" "W5" "W7" "X1" "X2" |
| **5** | INV_ASICO1 | Group into Strong Positive, Weak Positive, and None based on coefficient from glm predicting loss. Screened by p-value. | " " "00" "1B" "1D" "1E" "1G" "1H" "1J" "1K" "1S" "1X" "1Y" "1Z" "2A" "2B" "2E" "2F" "2G" "3A" "3C" "3D" "3F" "3G" "3H" "3J" "3K" "3M" "3P" "3Q" "3R" "3S" "3W" "3X" "3Y" "3Z" "4B" "4J" "4M" "4P" "4R" "4S" "4T" "4V" "4W" "4X" "4Z" "5A" "5B" "5C" "5D" "5E" "5F" "5H" "5J" "5K" "5L" "5N" "5P" "5Q" "5R" "5S" "5T" "5U" "5V" "5W" "5X" "5Y" "6B" "6C" "6E" "6M" "6P" "6Q" "6W" "6Y" "6Z" "7A" "7B" "7G" "7J" "7Q" "7R" "7Y" "8J" "8R" "9E" "9I" "A2" "A3" "A4" "A6" "B1" "B2" "B3" "B4" "B5" "C1" "C2" "C3" "C6" "C8" "D1" "D2" "D3" "D4" "D5" "D7" "D9" "E4" "E5" "E8" "F3" "F6" "G3" "G5" "G6" "G7" "G8" "G9" "I3" "K4" "K5" "K6" "K9" "L1" "L3" "L5" "L6" "L7" "L8" "M3" "M5" "M6" "M7" "N4" "N7" "N9" "P1" "P4" "Q4" "Q7" "R1" "R2" "R4" "R7" "R8" "R9" "S1" "S4" "S7" "S8" "S9" "T1" "T3" "T4" "T5" "T7" "T8" "U8" "U9" "V3" "V8" "V9" "W1" "W2" "W3" "W4" "W5" "W6" "W7" "W8" "X3" "Y7" |
| **6** | INV_ASICO2 | Grouped with above | " " "00" "1B" "1D" "1E" "1G" "1H" "1J" "1K" "1S" "1X" "1Y" "1Z" "2A" "2B" "2E" "2F" "2G" "2V" "3A" "3B" "3C" "3D" "3E" "3F" "3H" "3J" "3K" "3M" "3Q" "3R" "3S" "3W" "3X" "3Y" "3Z" "4B" "4J" "4M" "4P" "4R" "4T" "4V" "4W" "4X" "4Z" "5A" "5B" "5C" "5D" "5E" "5H" "5J" "5K" "5L" "5M" "5N" "5P" "5Q" "5R" "5S" "5T" "5U" "5V" "5W" "5X" "5Y" "6B" "6C" "6E" "6F" "6G" "6H" "6M" "6N" "6P" "6Q" "6R" "6U" "6V" "6W" "6Y" "6Z" "7A" "7B" "7G" "7J" "7Q" "7R" "7S" "7Y" "8A" "8J" "8K" "8L" "8R" "8X" "9B" "9E" "A3" "A4" "B1" "B2" "B3" "B4" "B5" "C1" "C2" "C3" "C6" "C8" "D2" "D5" "D7" "D9" "E4" "E5" "F3" "F4" "F5" "F6" "G3" "G5" "G6" "G7" "G8" "G9" "K4" "K5" "K6" "K9" "L3" "L5" "L6" "L7" "L8" "M6" "N4" "N7" "N9" "P1" "P4" "Q4" "Q7" "R1" "R2" "R4" "R7" "R8" "R9" "S1" "S4" "S7" "S8" "S9" "T1" "T3" "T4" "T5" "T6" "T7" "U8" "U9" "V3" "V8" "V9" "W1" "W2" "W3" "W4" "W5" "W6" "W7" "W8" "X3" "Y7" |
| **7** | INV_BASD | AVG_TOS = mean(DATA_DT-INV_BASD))<br><br>Dropped from individual predictions | "1964-12-01" "1965-11-23" "1967-05-09" "1967-12-11" …<br>"2022-04-10" "2022-05-04" "2022-06-08" "2022-07-17" |
| **8** | INV_BOSD | PriorService = sum(!is.na(INV_BOSD)) | "1978-07-11" "1981-12-17" "1984-12-22" "1985-05-22" …<br>"2019-08-11" "2019-08-28" "2019-09-15" "2019-11-26" |
| **9** | INV_BREX | Drop | "1996-09-01" "1999-01-01" "1999-06-01" "2001-03-01" …<br>"2022-09-01" "2022-12-01" "2023-03-01" "2023-08-01" |
| **10** | INV_CF1AOC1 | Drop: used more general category of INV_CRMGOF<br><br>Used to help in data cleaning | "001" "002" "003" "004" "005" "006" "007" "009" "00A" "00D" "00E" "011" "013" "015" "016" "017" "018" "021" "022" "025" "027" "029" "030" "031" "101" "102" "103" "104" "105" "106" "107" "108" "109" "110" "111" "112" "113" "114" "115" "116" "117" "118" "119" "11A" "11Z" "121" "122" "123" "124" "126" "127" "128" "129" "12A" "12B" "12D" "130" "131" "13A" "14A" "14B" "14E" "15A" "15B" "15C" "15D" "17A" "17B" "17D" "17X" "18A" "19A" "19B" "19C" "201" "202" "203" "204" "205" "206" "207" "208" "209" "210" "211" "213" "214" "215" "217" "218" "219" "21A" "21B" "220" "222" "224" "225" "228" "24A" |

| | | | |
|---|---|---|---|
| | | | "25A" "25E" "25G" "26A" "26B" "27A" "29A" "301" "302" "303" "304" "305" "306" "307" "308" "309" "30A" "310" "312" "314" "315" "316" "317" "319" "31A" "320" "321" "322" "323" "324" "325" "327" "329" "331" "34A" "35A" "35B" "35C" "35D" "35E" "35F" "35G" "36A" "37A" "37X" "38A" "38S" "38X" "401" "402" "403" "404" "405" "406" "407" "408" "409" "40A" "412" "413" "414" "416" "418" "420" "421" "424" "425" "426" "427" "428" "429" "42B" "42C" "42H" "44A" "45A" "46A" "46X" "48A" "48B" "48C" "48E" "48G" "48I" "48J" "48X" "49A" "49X" "501" "502" "503" "505" "506" "509" "50A" "510" "511" "514" "515" "517" "519" "51A" "51C" "51S" "51T" "51Z" "522" "523" "524" "525" "527" "529" "52A" "52B" "530" "531" "53A" "54A" "56A" "57A" "59A" "601" "602" "603" "604" "605" "606" "607" "608" "609" "610" "611" "612" "613" "614" "615" "616" "617" "618" "620" "622" "624" "626" "627" "628" "630" "64A" "65A" "65B" "65C" "65D" "66B" "66E" "66H" "66S" "67A" "67B" "67C" "67J" "701" "702" "703" "705" "706" "707" "708" "709" "70B" "710" "711" "712" "713" "714" "715" "716" "717" "718" "719" "720" "721" "723" "724" "725" "726" "729" "730" "731" "74A" "74B" "74C" "801" "802" "803" "804" "805" "806" "807" "810" "811" "812" "813" "815" "816" "817" "818" "819" "822" "823" "824" "825" "826" "827" "828" "829" "830" "88A" "88D" "89E" "901" "904" "905" "906""907" "908" "909" "90A" "910" "911" "912" "913" "914" "915" "917" "918" "919" "91A" "91B" "920" "921" "924" "925" "926" "927" "928" "929" "92A" "92B" "92D" "92F" "930" |
| **11** | INV_CRMGOF | Grouped all numbered branches into category: "functional" | " " "24" "26" "29" "30" "34" "36" "37" "38" "40" "42" "43" "46" "48" "49" "50" "51" "52" "53" "57" "58" "59" "90" "91" "99" "AC" "AD" "AG" "AM" "AR" "AV" "CA" "CM" "CY" "EN" "FA" "FI" "IN" "LG" "MI" "MP" "OD" "PO" "QM" "SC" "SF" "TC" |
| **12** | INV_DOB | Used to calculate age. Used age at promotion to CPT as a factor. Binary indicator above or below 28 years old. Determined 28 as best cut off value through trial and error | "1946-04-11" "1947-01-30" "1947-07-29" "1948-01-15" "1948-01-28" … "1997-06-20" "1997-07-04" "1997-07-11" "1997-09-15" "1998-02-09" "1999-09-24" |
| **13** | INV_DODSPS | Converted to binary indicator: Blank=FALSE | "" " " "0" "1" "4" "5" "6" "A" "F" "M" "N" "P" |
| **14** | INV_DTLBR | Converted to binary indicator: Blank=FALSE | "" " " "AD" "AG" "AR" "CM" "EN" "FA" "FI" "IN" "JA" "MI" "OD" "SC" "TC" |
| **15** | INV_ETHGRP | Drop | " " "1" "2" "3" "4" "5" "6" "7" "8" "9" "D" "E" "G" "H" "J" "K" "L" "Q" "S" "V" "W" "X" "Y" "Z" |
| **16** | INV_FSA | Drop | 0 1003 1971 1979 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2022 |
| **17** | INV_IRISK | risk_ind = mean(=="H")<br>Drop due to statistical insignificance | " " "H" "L" "M" |
| **18** | INV_MARST | Grouped: "S", "M" or "other" | " " "A" "D" "I" "L" "M" "S" "W" "Z" |
| **19** | INV_MOP | Dropped due to G-1 discontinued use of predictor | " " "A" "B" "C" "M" "P" |
| **20** | INV_MPATYP | Converted to Binary: "TA" or "other" | " " "AA" "AB" "AC" "AE" "AG" "AH" "NA" "RG" "TA" "TB" |
| **21** | INV_NBRMOA | Drop | " " "00" "01" "02" "03" "04" "05" "06" "07" "08" "09" "10" "11" "12" "13" "14" "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39" "40" "41" "42" "43" "44" "45" "46" "47" "48" "49" "50" "51" "52" "53" "54" "55" "56" "57" "58" "59" "60" "61" "62" "63" "64" "65" "66" "67" "68" "69" "70" "71" "72" "73" "74" "75" "76" "77" "78" "79" "80" "81" "82" "83" "84" "85" "86" "88" "90" "91" "92" "93" "96" "97" "98" |
| **22** | INV_NDEPA | Avg_Dep=mean(INV_NDEPA+INV_NDEPNC)<br><br>Converted to Binary indicator: Adult Dependents; 0=FALSE | 0 1 2 3 4 5 6 7 |
| **23** | INV_NDEPNC | Converted to Binary indicator: kids; 0=FALSE | 0 1 2 3 4 5 6 7 8 9 10 11 12 |
| **24** | INV_NOLOT | AVG_LDEP = mean(INV_NOLOT))<br><br>dropped from individual predictions: conflated with time | "" " " "0" "1" "2" "3" "4" "5" "6" |
| **25** | INV_NOSOT | AVG_SDEP = mean(INV_NOSOT))<br><br>dropped from individual predictions: conflated with time | "" " " "0" "1" "2" "3" "4" "5" "6" "7" "8" "9" |
| **26** | INV_NPCSCF | Drop | " " "0" "1" "2" "3" "4" "5" "6" "7" "8" "9" |
| **27** | INV_OCN_EDUCAT1 | Converted to binary indicator: complete: C=TRUE | " " "C" "P" |
| **28** | INV_OCX_OSASD | Drop | "1988-05-17" "1990-08-12" "1990-09-20" "1991-01-03" … "2022-10-27" "2022-10-28" "2022-10-30" "2022-10-31" |
| **29** | INV_ORSTT | Converted to Binary indicator: blank=FALSE | "" " " "A" "B" "C" "D" "E" "F" "G" "H" "I" "J" "K" "L" |
| **30** | INV_PDOR | Drop | "1980-10-19" "1981-05-13" "1983-07-19" … "2022-07-14" "2022-08-01" "2022-09-01" |
| **31** | INV_PHYC | Converted to Binary Indicator: "A" and all others combined into "subA" | " " "A" "B" "C" "D" "E" "F" "G" "H" "J" "L" "M" "N" "P" "S" "T" "U" "V" "W" "X" "Y" |

| | | | |
|---|---|---|---|
| **32** | INV_PJAOC | Drop | " " " Z" "00A" "00B" "00D" "00E" "00Z" "01A" "01B" "01C" "01D" "01F" "01M" "02A" "02B" "02C" "03A" "05A" "11A" "11Z" "12A" "12B" "12D" "12Z" "13A" "13Z" "14A" "14B" "14E" "14Z" "15A" "15B" "15C" "15D" "15G" "15I" "15K" "15L" "15M" "15N" "15Y" "15Z" "16C" "17A" "17B" "17X" "17Z" "18A" "18Z" "19A" "19B" "19C" "19Z" "21A" "21B" "21D" "21Z" "24A" "24B" "24X" "24Z" "25A" "25G" "25Z" "26A" "26B" "27A" "27B" "27Z" "29A" "29Z" "30A" "30Z" "31A" "31Z" "34A" "34Z" "35A" "35B" "35C" "35D" "35E" "35F" "35G" "36A" "36Z" "37A" "37X" "37Z" "38A" "38C" "38S" "38X" "38Z" "39Z" "40A" "40Z" "41Z" "42B" "42C" "42H" "42Z" "43Z" "44A" "44Z" "45Z" "46A" "46X" "46Z" "47A" "48A" "48B" "48C" "48D" "48E" "48F" "48G" "48H" "48I" "48J" "48P" "48X" "48Z" "49A" "49W" "49X" "49Z" "50A" "50Z" "51A" "51B" "51C" "51D" "51P" "51R" "51S" "51T" "51Z" "52A" "52B" "52Z" "53A" "53B" "53X" "53Z" "54Z" "56A" "56Z" "57A" "57Z" "58A" "59A" "59Z" "62Z" "65A" "65B" "65D" "65Z" "66H" "66N" "66Z" "67A" "67C" "67J" "67Z" "70A" "70B" "70H" "70K" "70Z" "72D" "72Z" "73A" "74A" "74B" "74C" "74Z" "88A" "88B" "88Z" "89E" "89Z" "90A" "91A" "91E" "91Z" "92A" "92Z" "Z " |
| **33** | INV_PRDVEM | Converted to Factor. Blanks grouped with "1" | " " "1" "2" "3" |
| | | | |
| **34** | INV_RACPOP | Drop | " " "A" "C" "E" "M" "N" "R" "S" "X" "Z" |
| **35** | INV_REDCAT | Converted to factor | "A" "C" "H" "N" "T" "X" |
| **36** | INV_RELDEN | Drop | " " "00" "01" "02" "04" "06" "07" "08" "10" "11" "12" "13" "14" "16" "18" "19" "20" "24" "26" "32" "34" "36" "38" "40" "41" "44" "45" "46" "47" "48" "49" "50" "53" "55" "56" "57" "58" "60" "62" "64" "66" "68" "70" "72" "73" "74" "75" "99" "A0" "A7" "AB" "AC" "AD" "AF" "AG" "AK" "AM" "AN" "AO" "AP" "AQ" "AV" "BA" "BB" "BC" "BD" "BE" "BF" "BG" "BH" "BT" "C4" "CA" "CB" "CC" "CD" "CE" "CF" "CG" "CH" "CI" "CJ" "DA" "DB" "DC" "DD" "DE" "DF" "DG" "DJ" "DK" "DL" "DM" "DN" "DP" "DQ" "DR" "DS" "DT" "DU" "DZ" "EA" "EC" "ED" "EE" "EJ" "EL" "EM" "EN" "EP" "EQ" "ER" "FA" "FB" "FC" "FH" "FQ" "FR" "FT" "FY" "G3" "GA" "GB" "GC" "GD" "HU" "JA" "JB" "JC" "JD" "JE" "JF" "JG" "JJ" "JK" "JL" "JM" "JN" "JO" "JP" "JQ" "JR" "JT" "JU" "KF" "KK" "LA" "LB" "LC" "LD" "LE" "LF" "LG" "LH" "LI" "MA" "MB" "MC" "MJ" "NB" "NC" "ND" "NE" "NF" "NG" "NH" "OO" "PP" "VV" "XC" "XN" "XX" "YY" "ZB" "ZY" |
| **37** | INV_RGAAST | Converted to binary factor: RGAASTA: A= TRUE | "" " " "A" "B" "N" |
| **38** | INV_RGTAFF | Drop | "" " " " A " " BCOR" "0001AD" "0001AR" "0001AV" "0001FA" "0001IN" "0001SF" "0002AD" "0002AR" "0002AV" "0002CV" "0002FA" "0002IN" "0003AD" "0003AR" "0003AV" "0003FA" "0003IN" "0004AD" "0004AR" "0004AV" "0004FA" "0004IN" "0005AD" "0005CV" "0005FA" "0005IN" "0006AD" "0006AV" "0006FA" "0006IN" "0007AD" "0007AR" "0007FA" "0007IN" "0008AR" "0008FA" "0008IN" "0009AR" "0009FA" "0009IN" "0010AR" "0010AV" "0010FA" "0010IN" "0011AR" "0011AV" "0011FA" "0011IN" "0012CV" "0012FA" "0012IN" "0013AV" "0013CV" "0013FA" "0014AR" "0014CV" "0014FA" "0014IN" "0015FA" "0015IN" "0016CV" "0016FA" "0016IN" "0017CV" "0017FA" "0017IN" "0018FA" "0018IN" "0019FA" "0019IN" "0020FA" "0020IN" "0021FA" "0021IN" "0022FA" "0022IN" "0023IN" "0024IN" "0025AV" "0025FA" "0026FA" "0026IN" "0027FA" "0027IN" "0028IN" "0029FA" "0029IN" "0030FA" "0030IN" "0031FA" "0031IN" "0032AR" "0032CV" "0032FA" "0032IN" "0033CV" "0033FA" "0034AR" "0034IN" "0035AR" "0035IN" "0036IN" "0037AR" "0037FA" "0038FA" "0038IN" "0040AR" "0040CV" "0040FA" "0041FA" "0041IN" "0042FA" "0043AD" "0044AD" "0046IN" "0051IN" "0052AD" "0052AV" "0052IN" "0056AD" "0058AV" "0058IN" "0061CV" "0061IN" "0062AD" "0063AR" "0064AR" "0066AR" "0067AR" "0068AR" "0069AR" "0070AR" "0071CV" "0072AR" "0073CV" "0075IN" "0076FA" "0077AR" "0077FA" "0078FA" "0079FA" "0081AR" "0082AV" "0082FA" "0084FA" "0087IN" "0089CV" "0091CV" "0094FA" "0101AV" "0108AD" "0123AV" "0158AV" "0159AV" "0160AV" "0187IN" "0212AV" "0214AV" "0223AV" "0227AV" "0228AV" "0229AV" "0319FA" "0320FA" "0321FA" "0325IN" "0327IN" "0333FA" "0377FA" "0501AV" "0501IN" "0502AV" "0502IN" "0503IN" "0504IN" "0505IN" "0506IN" "0507IN" "0508IN" "0509IN" "COBCOR" "CORPAG" "CORPAM" "CORPAQ" "CORPCA" "CORPCH" "CORPCM" "CORPCY" "CORPEN" "CORPFI" "CORPJA" "CORPMI" "CORPMP" "CORPOD" "CORPQM" "CORPSC" "CORPSO" "CORPTC" "PSYOPS" |
| **39** | INV_SAPRDT | Drop | "1999-05-14" "1999-07-16" "1999-09-17" "2000-02-11" … "2023-09-19" "2023-09-29" "2023-10-25" "9121-01-29" |
| **40** | INV_SCOA | Converted to Binary: "Active Original": R=TRUE | " " "G" "R" "T" "V" "Z" |
| **41** | INV_SEX | Gender=sum(INV_SEX=="M"); mutate(Gender_Ratio = Male/INV) | "F" "M" |
| **42** | INV_SRECDT | Checked if a CPT had any information in their earliest entry. If not blank, then earlySep_ppw=TRUE | "1999-05-06" "1999-07-12" "1999-09-09" "1999-11-30"… "6111-01-20" "7030-10-20" "7031-01-23" "9030-01-21" |

| | | | |
|---|---|---|---|
| **44** | INV_OCN_CEDG-1 | Grouped categories into education levels. Then made a factor for highest education level achieved: "Associate", "Bachelor", "Masters", "Doctorate" | " " "AA " "AAA " "AAS " "AD " "ADC " "BA " "BAD " "BAE " "BAL " "BAR " "BBA " "BD " "BE " "BES " "BFA " "BGS " "BHA " "BIS " "BL " "BM " "BME " "BOA " "BOB " "BPA " "BPS " "BPT " "BS " "BSE " "BSN " "BSW " "BUS " "CE " "CIVE" "DBA " "DCD " "DCH " "DDS " "DDV " "DM " "DMUS" "DPA " "DPT " "DVM " "EDD " "EDS " "EJD " "ENS " "JD " "JSD " "LLB " "LLM " "LOG " "MA " "MAA " "MAG " "MAM " "MAS " "MBA " "MBI " "MBIO" "MBV " "MCI " "MCN " "MCVE" "MD " "MDMP" "MDV " "ME " "MEG " "MEN " "MEO " "MEVS" "MFA " "MFS " "MGA " "MHA " "MHC " "MHI " "MHR " "MHS " "MIA " "MIE " "MIM " "MIPA" "MIPP" "MLS " "MMAS" "MMB " "MMC " "MME " "MMEN" "MMO " "MMS " "MNE " "MNR " "MOM " "MOPM" "MOSW" "MOT " "MPA " "MPAF" "MPAS" "MPH " "MPM " "MPP " "MPS " "MPT " "MPY " "MRD " "MS " "MSA " "MSC " "MSE " "MSEN" "MSIM" "MSIT" "MSM " "MSPA" "MSS " "MSSI" "MSSM" "MSST" "MST " "MSTI" "MSW " "MT " "MTRS" "OT " "OTD " "PDD " "PHD " "PMD " "SCD " "SED " "THD " "THM " "YYYY" "ZZZZ" |
| **45** | INV_SREQDT | Drop | "2000-03-21" "2000-06-12" "2001-02-10" "2001-05-29" … "2108-05-01" "8073-10-20" "9013-01-20" "9121-10-20" |
| **46** | INV_SSUBDT | Drop | "1999-04-15" "1999-06-21" "1999-08-23" "1999-11-10" … "2023-08-02" "2023-08-05" "7060-10-20" "9052-01-21" |
| **47** | INV_STBR | See INV_STHRED | "" " " "00" "01" "02" "04" "05" "06" "08" "09" "10" "11" "12" "13" "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39" "40" "41" "42" "44" "45" "46" "47" "48" "49" "50" "51" "53" "54" "55" "56" "60" "66" "69" "72" "74" "78" "89" |
| **48** | INV_STHRED | Track change in birth to current State of Record: CHG_ST = sum(INV_STHRED==INV_STBR)<br><br>Individual model: made into binary indicator on if birth state was significantly different than reference category (TX) | " " "01" "02" "04" "05" "06" "08" "09" "10" "11" "12" "13" "15" "16" "17" "18" "19" "20" "21" "22" "23" "24" "25" "26" "27" "28" "29" "30" "31" "32" "33" "34" "35" "36" "37" "38" "39" "40" "41" "42" "44" "45" "46" "47" "48" "49" "50" "51" "53" "54" "55" "56" "60" "66" "69" "71" "72" "74" "78" "AE" "CA" "FL" "FN" "GA" "LA" "MD" "MN" "NC" "NY" "OK" "TN" "TX" "VA" "WA" |
| **49** | LOSS_DT | loss = sum(!is.na(LOSS_DT))<br><br>used as binary indicator for individual model response variable | "2008-12-01" "2009-01-01" "2009-02-01" "2009-03-01" … "2022-07-01" "2022-08-01" "2022-09-01" "2022-10-01" |
| **50** | LOSS_TYPE | | "" "INDUCED" "NATURAL" |
| **51** | MSV_OFF_QY | drop: conflated with time | |
| **52** | SOC | Factor | "OCS_COL" "OCS_INS" "OTHER" "ROTC-NON" "ROTC-SCH" "USMA" |
| **53** | TIG_QY | avg_time_CPT = mean(MSV_OFF_QY)<br><br>dropped from individual model: conflated with time | Integer: 0 - 400 |

## Appendix C: Exploratory Analysis

**Figure C1:** Source of Commission

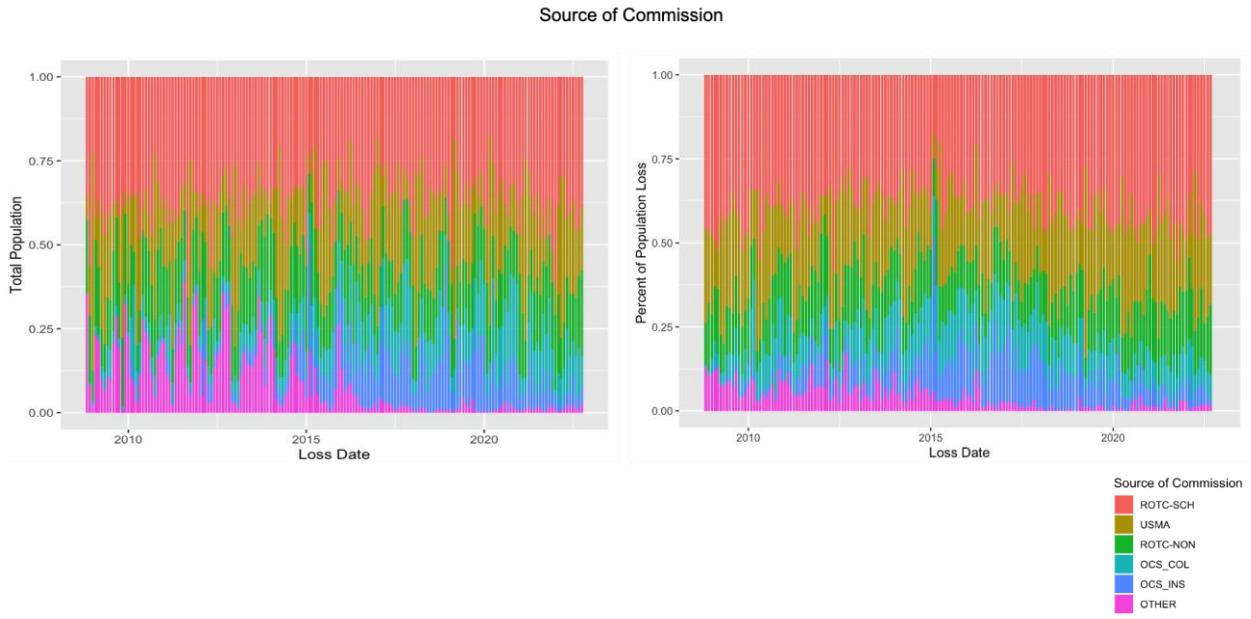



**Figure C2:** Career Emphasis

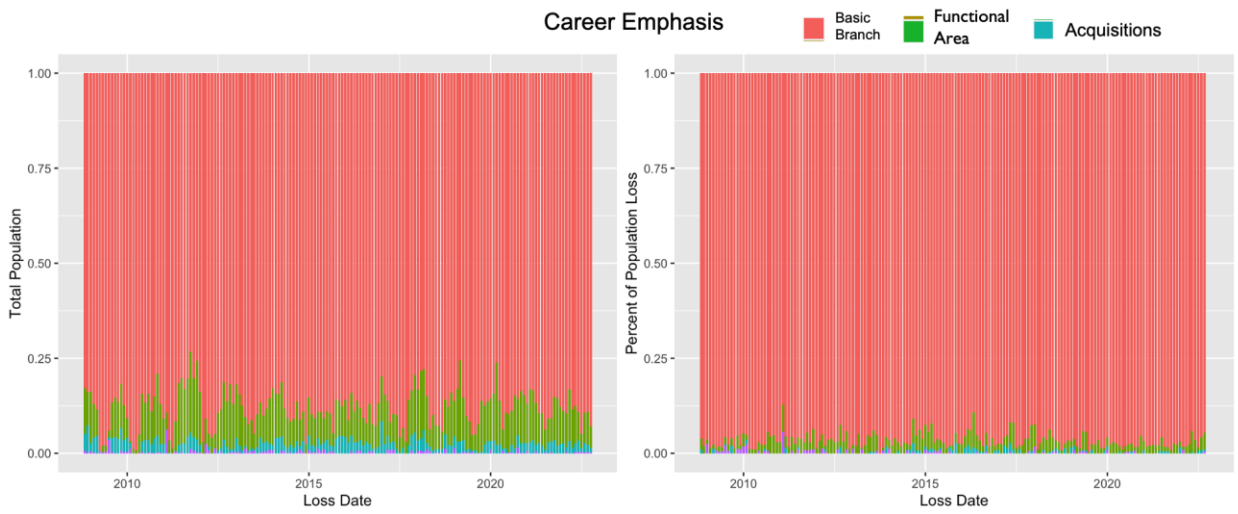


**Figure C3:** Gender

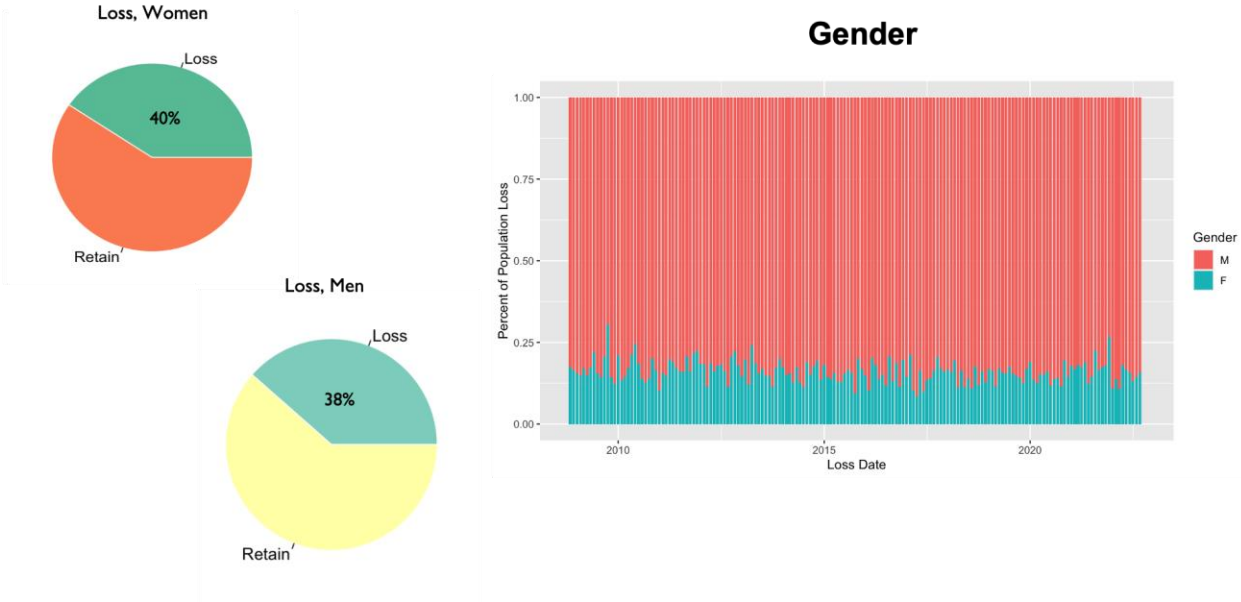



**Figure C4:** Marital Status

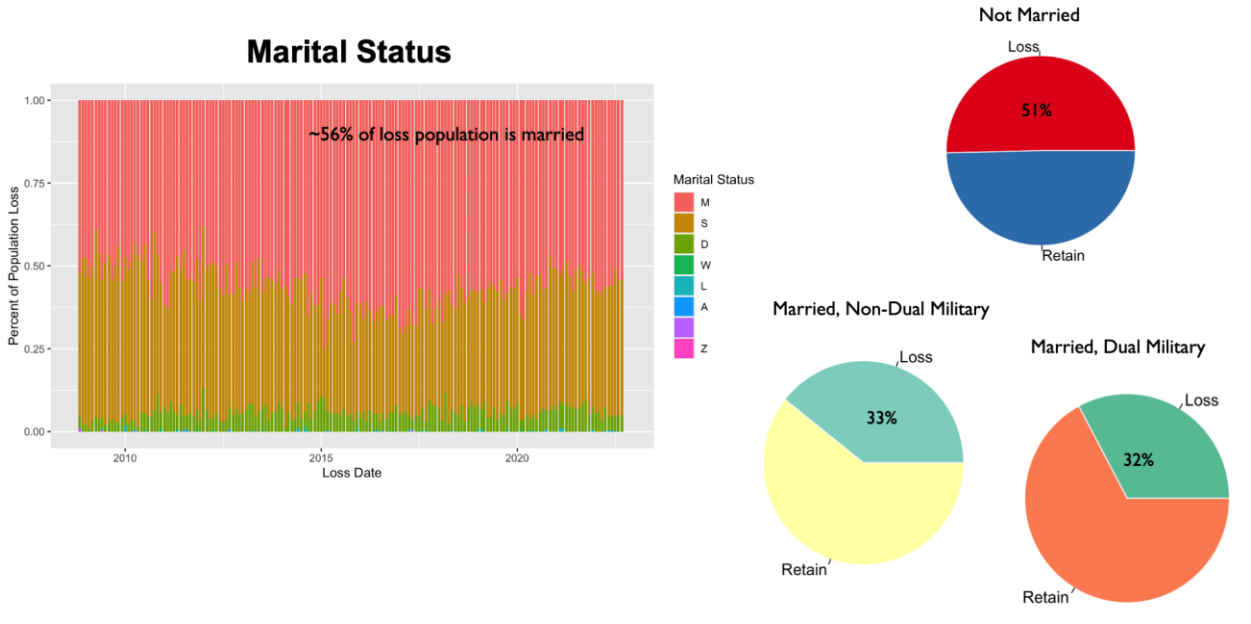



**Figure C5:** Consumer Confidence Index

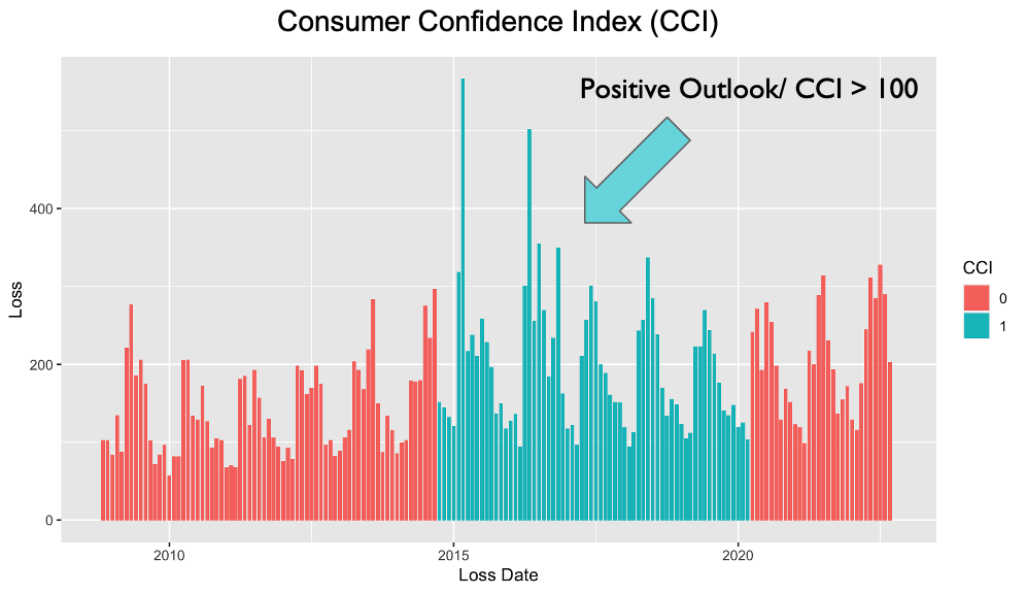



**Figure C6:** Vacancy to Unemployment Ratio

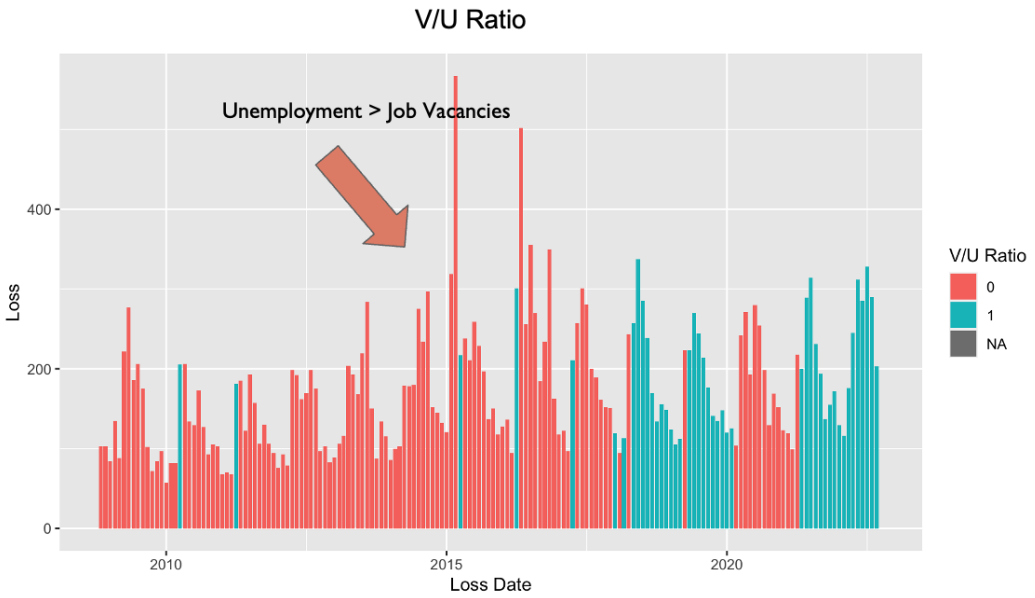



**Figure C7**: G-1 Model-Forecasted versus actual loss rates from 2016 to 2020

## Appendix D: Individual and Summary Model

**Figure D1:** ROC Curve for Logistic Regression

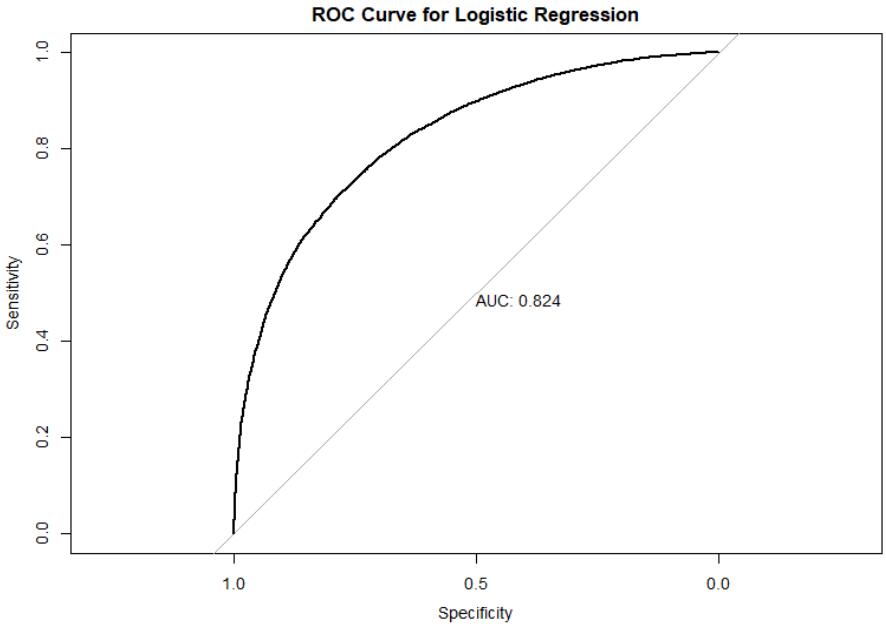
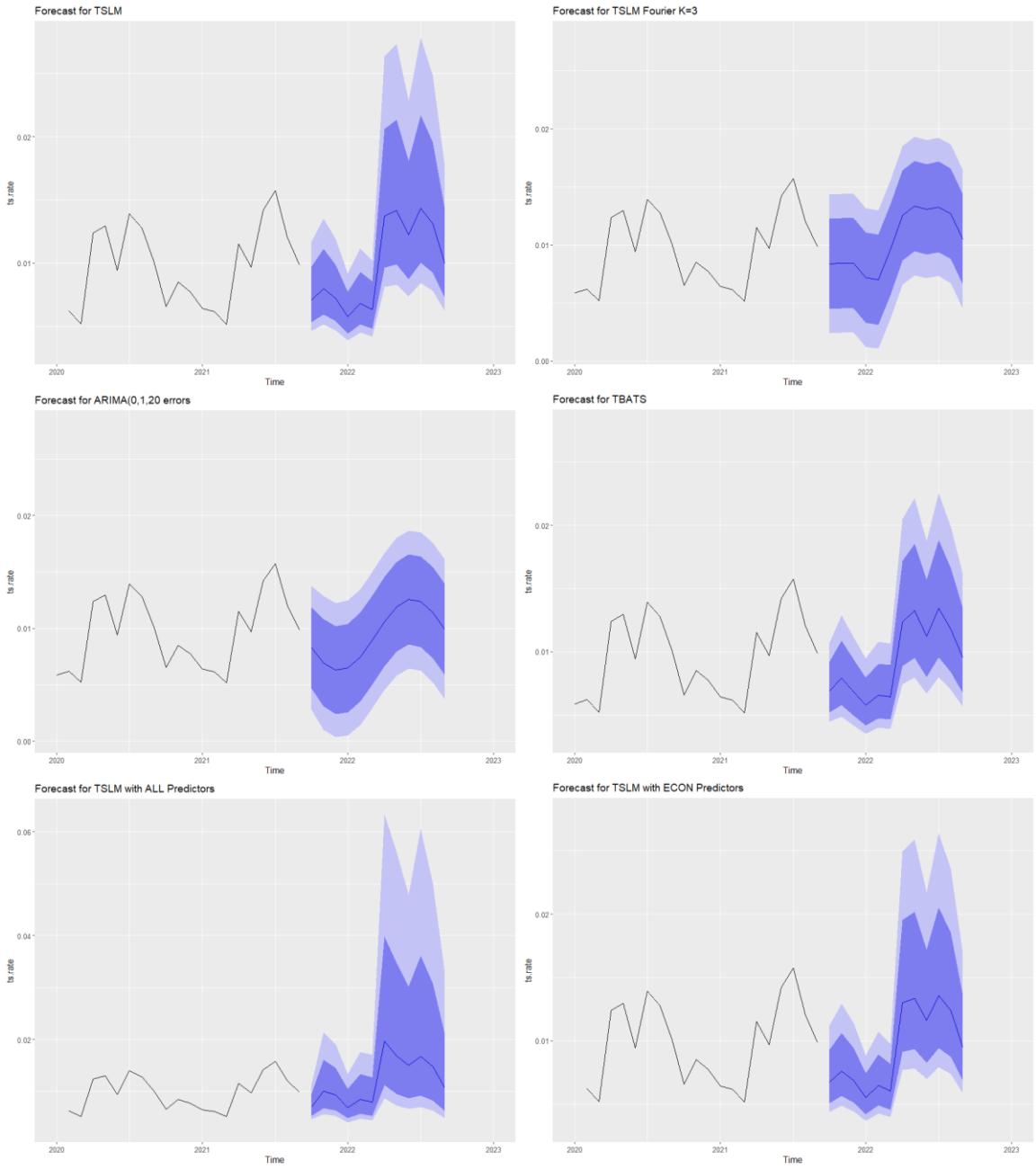



**Figure D.2:** Predictor Variables for TSLM model

| Predictor | COEFF | P-val | Code |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Inventory (INV) | 6.142e-04 | | n_distinct(ID) |
| Gen_Ratio | -1.290014e+02 | | sum(INV_SEX=="F")/INV |
| ASCO | 2.303001e+01 | | [sum(INV_ASCO1 %in% c("C3","C8"),na.rm=TRUE) +sum(INV_ASCO2 %in% c("C3","C8"),na.rm=TRUE) +sum(INV_ASCO3 %in%c("C3","C8"),na.rm=TRUE)] /INV |
| EFMP_r | 9.500828e+00 | | [sum(INV_ASCO1=="D5",na.rm=TRUE) +sum(INV_ASCO2=="D5",na.rm=TRUE) +sum(INV_ASCO3=="D5",na.rm=TRUE)] /INV |
| ROTC | -7.224340e+01 | 0.000784 | [sum(SOC %in% c("ROTC_SCH","ROTC_SCH"))] /INV |
| USMA | -3.421074e+01 | | sum(SOC=="USMA",na.rm=TRUE)/INV |
| OCS | -7.932344e+01 | 0.009561 | [sum(SOC %in% c("OCS_INS","OCS_COL"))] /INV |
| MARITAL | -4.830209e+01 | | sum(INV_MARST %in% c("D","S","A","L"))/INV |
| PHYS | -2.593402e+00 | | sum(!is.na(INV_PHYC))/INV |
| ORSTT | 3.961446e+00 | | sum(!is.na(INV_ORSTT))/INV |
| ETH_HR | 2.135635e+02 | 0.015869 | sum(INV_ETHGRP %in% c("X","Y","1","4","S","5","G","Q","D")) / INV |
| PRDVEM | 6.844983e+00 | | sum(INV_PRDVEM!="1") / INV |
| DOD_SP | -7.341795e+01 | | DOD_SPOUSE = sum(!is.na(INV_DODSPS)) / INV |
| YG_OUT | -1.355945e+01 | | sum(year(DATA_DT)-INV_FSA >10) / INV |
| BRCH_DTL | 4.772121e+01 | | sum(!is.na(INV_DTLBR)) / INV |
| SEP_PROJ | -1.354457e+01 | | sum(!is.na(INV_ASEPDT)) / INV |
| SEP_REQ | 3.690294e+01 | | sum(!is.na(INV_SRECDT)) / INV |
| V_U_RATIO | 1.284937e-01 | | case_when(V_U_RATIO>=1~1,TRUE~0), |
| CCI | 4.073960e-01 | | case_when(CCI>=100~0,TRUE~1)) |
| badYear | 2.103453e+00 | 0.000696 | case_when(year %in% c(2015,2016)~1,TRUE~0)) |

## Forecast Graphs



## References


Bureau of Labor Statistics. *Job Opening and Labor Turnover (JOLTS) Dataset.* (September 2022). https://data.bls.gov/PDQWeb/jt.

Bureau of Labor Statistics. *Current Population Survey Dataset.* (September 2022). https://data.bls.gov/PDQWeb/ln.

Hyndman, R.J., & Athanasopoulos, G. (2018) *Forecasting: principles and practice*, "Evaluating the Regression Model," 2nd edition, OTexts: Melbourne, Australia. OTexts. https://otexts.com/fpp2/.

Hyndman, Rob J. "Forecasting time series with complex seasonal patterns using exponential smoothing." *Hyndsight Blog*, 31 Dec. 2011, https://robjhyndman.com/publications/complex-seasonality/.

Hyndman, Rob J. "Forecasting Weekly Data." *Hyndsight Blog*, 5 Mar. 2014, https://robjhyndman.com/hyndsight/forecasting-weekly-data/.

Office of the Deputy Chief of Staff for Personnel, Department of the Army. *Army Captain Attrition Dataset.* (October 2022).

Organization for Economic Cooperation and Development. *Consumer Confidence Index Survey Dataset.* (October 2022). https://data.oecd.org/leadind/consumer-confidence-index-cci.htm.

"Time-Series Forecasting Using TBATS Model." Big Data Analytics. Tenth Planet, November 23, 2020. https://blog.tenthplanet.in/time-series-forecasting-tbats/.

Wickham H, François R, Henry L, Müller K (2022). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.