DA485 Capstone

Modeling Employee Attrition:

Identify Important Features That Contribute to Turnover

Authors:

Alex Strosahl, Braydon Gemar, Garrett Kelly, Joseph Havas

Intro

Although it is difficult to identify specific factors that influence employee attrition, it is a necessary function of any healthy organization in order to reduce high turnover costs, enhance productivity, and profitability. Therefore, employee retention greatly affects organizational performance, both positive and negative. While it varies by sector and position, it can take anywhere from 3-8 weeks' worth of wages to properly onboard new employees [1]. Additionally, the overall acclimation period before a new hire is fully independent in their role can extend even longer -- which further contributes to higher hiring costs and lost productivity.

In this study, we modeled attrition using Support Vector Machine, Random Forest, and a Logit Link function. There were stark differences in accuracies between Random Forest and Support Vector Machine at 83.63% and 97.12% respectively, after reducing the number of variables included in the model to combat over fitting. The Logit Link function implemented on our data set produced an average accuracy rate of 89% post variable reduction using Stepwise Fisher's variable selection method.

In defining final models, we discovered a major mistake in our methodology that needs to be acknowledged wherein we neglected to factorize Attrition (our response) and transformed our Overtime attribute to binary operators – rather than using their original Yes/No format – which had resulted in an SVM model that was only 85.3-percent accurate. Unfortunately, these mistakes were included in our presentation and were not identified until consolidating our R-scripts into a master document and preparing this report.

After transforming the Attrition variable into binary factor levels, and factorizing Over Time, the SVM model was rebuilt resulting in an increased accuracy rate from the original 85.3% as reported in our presentation to 97.12% percent accuracy rate which will be outlined in

this report. Furthermore, improvements to the Logit model were made increasing average attrition prediction accuracy rates from 88% to 89%, and our Random Forest dropped to 83.63% accuracy. Additionally, we are able to reduce the overall number of attributes needed in our SVM from twelve to six predictors, and our most valuable dimension changed from Job Role to Gender.

After fixing this problem and rebuilding our models, the accuracy of our SVM model jumped from the 85.3-percent that was reported in our presentation to the 97.12-percent accurate model outlined in this report – an increase of more than 12-percent. We also managed to improve our GLM model to achieve an accuracy of 89%, and our Random Forest dropped to 83.63% accuracy. Additionally, we are able to reduce the overall number of attributes needed in our SVM from twelve to six predictors, and our most valuable dimension changed from Job Role to Gender.

Review of Existing Literature

In reviewing other studies and literature on the subject of workplace satisfaction and attrition, we found that the work environment and how the employees felt about it was the most important determinant of satisfaction and attrition. In addition, gender proved to be an important proponent for workplace stress and retention as a result. This is consistent with our finding that gender is the most important dimension for modeling attrition, and our finding that overall employee satisfaction is an important determinant for employee attrition. Incentives also proved to be a good predictor of attrition according to our research, which is consistent with how monthly income is a very important factor in determining if an employee will attrit or not. In addition, it has been found that overall employee satisfaction can be linked to firm performance -- which could potentially link attrition to the long-term performance of a firm. Overall, the message that other studies convey is that the work environment itself is the most important part in determining whether someone attrits or not.

Data Introduction & Preparation

The data used in this study was a fictitious dataset provided by IBM for the purpose of training classification models for Employee Attrition. It includes survey data in addition to values recorded by a Human Resources department for the purposes of record keeping organization. Table 1 below provides a full summary of our dataset, which includes the variables that were removed pre-exploratory analysis due to being static or and redundant. Table 2 provides an exploratory analysis summary for the remaining attributes. Table 3 provides our feature selection methodology.

Note: *See Appendix for non-normal histograms & box plots with outliers.*

Note: *refer to master R-script for all histograms, box plots, and frequency tables.*

Table 1. Raw Data Summary

ATTRIBUTE	DATA TYPE & FORMAT	RANGE OR # OF LEVELS (IF QUALITATIVE)	NOTES
AGE	Numerical Predictor	18 - 60	NA
ATTRITION	Binary Response	Yes/No	NA
BUSINESS TRAVEL	Dimension	3 Levels	NA
DAILY RATE	Numerical Predictor	102 - 1499	Removed (redundant)
DEPARTMENT	Dimension	3 Levels	NA
DISTANCE FROM HOME	Numerical Predictor	1 - 29	NA
EDUCATION	Numerical Predictor	1 - 5	NA
EDUCATION FIELD	Dimension	6 Levels	NA
EMPLOYEE COUNT	Primary Key	1	Removed (static)
EMPLOYEE NUMBER	Primary Key	1 - 1470	Removed
ENVIRONMENT SATISFACTION	Numerical Predictor	1 - 4	NA
GENDER	Dimension	2 Levels	NA
HOURLY RATE	Numerical Predictor	30 - 100	NA
JOB INVOLVEMENT	Numerical Predictor	1 - 4	NA
JOB LEVEL	Numerical Predictor	1 - 5	NA
JOB ROLE	Dimension	9 Levels	NA
JOB SATISFACTION	Numerical Predictor	1 - 4	NA
MARITAL STATUS	Dimension	3 Levels	NA
MONTHLY INCOME	Numerical Predictor	1009 - 19999	NA
NUMBER OF COMPANIES WORKED	Numerical Predictor	0 - 9	NA
OVER 18?	Binary Predictor	Yes	Removed (static)
OVERTIME	Binary Predictor	Yes / No	NA
PERCENT SALARY HIKE	Numerical Predictor	11 - 25	NA
PERFORMANCE RATING	Numerical Predictor	3 - 4	NA
RELATIONSHIP SATISFACTION	Numerical Predictor	1 - 4	NA
STANDARD WORKING HOURS	Numerical Predictor	80 - 80	Removed (Static)
STOCK OPTION LEVEL	Numerical Predictor	0 - 3	NA

TOTAL WORKING YEARS	Numerical Predictor	0 - 40	NA
TRAINING TIMES LAST YEAR	Numerical Predictor	0 - 6	NA
WORK LIFE BALANCE	Numerical Predictor	1 - 4	NA
YEARS AT COMPANY	Numerical Predictor	0 - 40	NA
YEARS IN CURRENT ROLE	Numerical Predictor	0 - 18	NA
YEARS SINCE LAST PROMOTION	Numerical Predictor	0 - 15	NA
YEARS WITH CURRENT MANAGER	Numerical Predictor	0 -17	NA

Table 2. Exploratory Analysis

ATTRIBUTE	DISTRIBUTION	OUTLIERS
AGE	Normal, Right Skew	No
DAILY RATE	Uniform	No
DISTANCE FROM HOME	Strong Right Skew	No
EDUCATION	Normal	No
ENVIRONMENT SATISFACTION	Left Skew	No
HOURLYRATE	Uniform	No
JOB INVOLVEMENT	Left Skew	No
JOB LEVEL	Right Skew	No
JOB SATISFACTION	Left Skew	No
MONTHLY INCOME	Right Skew	Yes
MONTHLY RATE	Uniform	No
NUMBER FOR COMPANIES WORKED	Right Skew	Yes
PERCENT SALARY HIKE	Right Skew	No
PERFORMANCE RATING	Right Skew	Yes
RELATIONSHIP SATISFACTION	Left Skew	No
STOCK OPTION LEVEL	Right Skew	Yes
TOTAL WORKING YEARS	Right Skew	Yes
TRAINING TIMES LAST YEAR	Normal	Yes
WORK LIFE BALANCE	Normal	No
YEARS AT COMPANY	Right Skew	Yes
YEARS IN CURRENT ROLE	“W” Shaped	Yes
YEARS SINCE LAST PROMOTION	Right Skew	Yes
YEARS WITH CURRENT MANAGER	“W” Shaped	Yes

Table 3. Feature Creation

NEW FEATURE	METHODOLOGY	REMOVED ATTRIBUTES
SATISFACTION	Sum of Environmental Satisfaction, Job Involvement, Job Satisfaction, Relationship Satisfaction, Work Life Balance	• Environmental Satisfaction • Job Involvement • Job Satisfaction • Relationship Satisfaction • Work Life Balance

Table 4. Final Data Summary

ATTRIBUTE	DATA TYPE	DISTRIBUTION	OUTLIERS
ATTRITION	Response	NA	NA
BUSINESS TRAVEL	Dimension	NA	NA
DEPARTMENT	Dimension	NA	NA
EDUCATION FIELD	Dimension	NA	NA
GENDER	Dimension	NA	NA
JOB ROLE	Dimension	NA	NA
MARITAL STATUS	Dimension	NA	NA
AGE	Predictor	Normal, Right Skew	No
DAILY RATE	Predictor	Uniform	No
DISTANCE FROM HOME	Predictor	Strong Right Skew	No
EDUCATION	Predictor	Normal	No
HOURLY RATE	Predictor	Uniform	No
JOB LEVEL	Predictor	Right Skew	No
MONTHLY INCOME	Predictor	Right Skew	Yes
MONTHLY RATE	Predictor	Uniform	No
NUMBER FOR COMPANIES WORKED	Predictor	Right Skew	Yes
PERCENT SALARY HIKE	Predictor	Right Skew	No
PERFORMANCE RATING	Predictor	Right Skew	Yes
SATISFACTION	Predictor	Normal	Yes
STOCK OPTION LEVEL	Predictor	Right Skew	Yes
TOTAL WORKING YEARS	Predictor	Right Skew	Yes
TRAINING TIMES LAST YEAR	Predictor	Normal	Yes
YEARS AT COMPANY	Predictor	Right Skew	Yes
YEARS IN CURRENT ROLE	Predictor	“W” Shaped	Yes
YEARS SINCE LAST PROMOTION	Predictor	Right Skew	Yes
YEARS WITH CURRENT MANAGER	Predictor	“W” Shaped	Yes

Core Model Methodology

We implemented SVM by using Backwards Elimination based on Random Forest feature selection results. We tested all kernels and tuned the cost and gamma of our model by implementing a cross-fold validation technique that tests all possible combinations of cost and gamma. We validated our model by implementing 10-fold Cross-Validation. SVM proved to be the best method for predicting Attrition due to the nature of hyperplanes and how they interact with binary classification.

With Random Forest, we implemented the same feature selection methods as we did for SVM. Random Forest did not prove to be a very effective modeling method for our problem due to the computation cost.

Due to the low attrition prediction accuracy rates upon first implementation of SVM and Random Forest, we implemented h2o in R to discover which application would best suit our data set to accurately predict employee longevity. Results from h2o indicated that a General Linear Model would best fit our data set to make accurate predictions regarding employee attrition based on a logit link classification probabilistic function. The variables used in the probabilistic function were determined using Stepwise Fisher's variable selection method wherein the model with the lowest AIC, highest prediction accuracy rate, and highest H-statistic was implemented to determine which dimensions and factors most greatly influenced employee attrition.

Core Model Results

SVM

Support Vector Machine works by using hyperplanes to classify the data based on the predictors. It's a lightweight modeling method that works well for binary classification but can be applied to higher dimensional classification problems by incorporating more hyperplanes. It proved to be computationally cheap and the most effective method for classifying Attrition. We tuned SVM with a cost of 0.6, a gamma of 0.1, and used a Polynomial Kernal.
This was able to attain an accuracy of 97.12%.

Factors Used: Gender, Monthly Income, Satisfaction, Age, Job Involvement, Job Level, Overtime

Random Forest

Random Forest utilizes decision trees for a probabilistic approach to classification. This is a robust approach for our dataset due to its ability to adapt to new levels, with models showing an average accuracy of 83.63%. It was consistently able to classify observations with a "no" Attrition response, but it struggled to correctly classify observations which didn't attrit. Additionally, it was computationally expensive to implement compared to SVM, so the reduction in accuracy is not worth the robustness that Random Forest offers.

Random Forest places more importance on higher variance data points, so the dimension with more levels is likely to be recognized as having higher information gain. This resulted in the

dimensions with lower levels being less important for the Random Forest model, and the only dimension that had a positive impact on model accuracy was Job Role. This is inconsistent with Gender being the most important dimension for the other modeling methods applied in this study.

Factors Used: Gender, Monthly Income, Satisfaction, Age, Job Involvement, Job Level

Fischer's Stepwise Variable Selection Method in a Probabilistic General Linear Model

The General Linear Model applied in this analysis is a binomial distribution Logit Link function based on log-odds probability. This application of a GLM reduces the impact of overfitting by penalizing a model with a high AIC but increases the computational cost when there are more variables and observations. Specifically, the logit link model benefits from qualitative data when there is abnormality present in the distribution and a binary probability in a response variable therefore, the IBM employee attrition data set containing a great deal of ordinal and dimensional data appeared to be ideal. Our final model was implemented with the most dimensions used of all other methods while attaining a classification and prediction accuracy of 89%.

Final Model:

Attrition = 3.89 - 0.03(Age)+ 2.02(BusinessTravelTravel_Frequently)+ 1(BusinessTravelTravel_Rarely) + 0.05(DistanceFromHome) - 0.8(EducationFieldLife Sciences) - 0.34(EducationFieldMarketing) - 0.891928(EducationFieldMedical) - 0.87(EducationFieldOther) + 0.23(EducationFieldTechnical Degree) + 0.33(GenderMale) -

1.15(JobInvolvement2) - 1.46(JobInvolvement3) - 2.17(JobInvolvement4) - 1.75483(JobLevel2) -0.478625(JobLevel3) - 1.63(JobLevel4) + 0.36(JobLevel5) + 0.2(JobRoleHuman Resources) + 0.61(JobRoleLaboratory Technician) - 0.84(JobRoleManager) + 0.2(JobRoleManufacturing Director) - 2.06(JobRoleResearch Director) - 0.45(JobRoleResearch Scientist) + 1.21(JobRoleSales Executive) + 0.98(JobRoleSales Representative) + 0.21(NumCompaniesWorked) + 2(OverTimeYes) - 1.49(StockOptionLevel1) - 1.33(StockOptionLevel2) - 0.96(StockOptionLevel3) - 0.06(TotalWorkingYears) - 0.19(TrainingTimesLastYear) + 0.15(YearsAtCompany) - 0.14(YearsInCurrentRole) - 0.12(YearsWithCurrManager) - 0.35(Overall_Satisfaction)

The Logit function used in this model is the natural log odds that Attrition is 1, indicating that an employee remained at a company where parameters are described in terms of probability ratio to predict if an employee attrits.

The parameters defined in the model above are chosen from the model with the lowest AIC in the Stepwise Fisher's selection method with an AIC of 902.42. The AIC estimates the in-sample prediction error functioning similar to an adjusted R-square effectively penalizing a model that uses too many variables. Furthermore, the goodness of fit and classification accuracy of this model was determined by implementing statistical testing methods, classification tables, and ROC curve plotting.

The statistical test we implemented to determine goodness of fit used was the Hosmer-Lemeshow goodness of fit test which is calculated by dividing the data set according to

their predicted probabilities based on their estimated parameter values, as seen in the equation above. The probability that attrition = 1 is calculated for each observation in the sample based on each observation's covariates. Our final H-statistic for the model using every observation in the data set was 0.067, greater than the p-value of 0.05 indicating poor fit.

It's important to note that large p-values do not strictly indicate good fit, because, lack of evidence against a poor fit is not sufficient for declaring good fit in favor of the alternative hypothesis. Because of our large sample size, and a moderate p-value greater than 0.05, our results from the test are not a consequence of the test having lower power to detect misspecification of a model, rather it is indicative of good fit. Therefore, with a p-value of .067, it is indicated that this model is sufficient in making predictions regarding the probability of an employee remaining at a company.

To model was also tested by plotting the ROC curve and the C-value Area Under the Curve. The ROC measures in this analysis were sensitivity, 1-Specificity, False Positive, and False Negative and sensitivity measuring the goodness of accuracy & specificity measures the weakness of the model.

The plot of accuracy and specificity measures is displayed in the appendix on the concave plot indicating that as sensitivity is increasing 1-specificity is increasing but at a diminishing rate. The C-value Area Under the Curve, known as the AUC, or the value of the concordance index calculates the measure of the area under the ROC curve. According to

these statistical measures testing goodness of fit if C-value is .5 or less, it would indicate that the model cannot accurately discriminate between 0 and 1 in the dependent variable. This suggests that the model cannot appropriately predict which employees stay and which employees leave based on given parameters.

The c-value AUC is much greater than 50%, at 88% indicating that our model can accurately discriminate between 0 and 1 concluding that our model is sufficient for making predictions for employee attrition based on the model parameters.

With the evidence of significant variables in our model, Hosmer-Lemeshow goodness of fit test H statistic p-value of .067, and 88% area under the curve, our model was further tested by creating classification accuracy tables that was then implemented into a for loop to calculate the average classification attrition prediction accuracy at 89%.

This indicates that if an organization gives attention to these variables and makes effort in controlling for them, they have a greater chance of keeping an employee and making changes in the organization to prevent high turnover. It is indicated from our final model that the most significant variables in predicting attrition are Age, Business Travel, Distance From Home, Gender, Job Involvement, Job Level factor level 2 and 4, Job Role Research Director factor level, Job Role Sales Executive factor level, Number of Companies Worked, Overtime, Stock Option Level, Total Working Years, Training Times Last Year, Years at Company, Years In Current Role, Years With Current Manager, and Overall Satisfaction with all p-values less than .05.

β_0, is the intercept parameter sometimes called the constant term which describes the relationship that Attrition is a function of Age, Business Travel, Distance From Home, Education Field, Gender, Job Involvement, Job Level, Job Role, Number of Companies Worked, Over Time, Stock Option Level, Total Working Years, Training Times Last Year, Years at Company, Years in Current Role, Years with Current Manager, and Overall Satisfaction. The Intercept is the mean value of the response variable when all of the predictor variables in the model are equal to zero. Indicating that when all explanatory variables are zero, the probability that an employee attrits is 4%.

β_1, the coefficient of Age, is the coefficient of a probabilistic logit model that is the difference in the log odds, that a one unit increase in Age changes the expected probability of employee attrition by 97% per one unit change in age while all other factors are fixed.

β_2, the coefficient of Business Travel at the factor level of Travel Frequently, is the coefficient of a probabilistic logit model indicating that a one unit increase in this variable at this factor level changes the expected probability of employee Attrition increasing the likelihood that an employee attrits if they are able to travel frequently 3 times more than if they were to travel rarely.

β_3, the coefficient Business Travel at the factor level of Travel Rarely in our probabilistic logit model increases the expected probability of employee attrition when measuring the effect of this variable at this factor level on attrition holding all other factors fixed. Meaning

that a one unit change in this variable at this factor level increases the likelihood of employee attrition twice more than any other factor level of business travel.

β_4, the coefficient of Distance From Home is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee lives within a specified range of their place of work, they are more likely to stay longer at a company.

β_5, β_6, β_7, β_8, β_9, are the coefficients of Education Field on the factor levels of Life Sciences, Marketing, Medical, Other, and Technical Degree. These coefficients in our probabilistic logit model change the expected probability of employee attrition while all other factors are fixed. Specifically, a one unit change in any of these factor levels decreases the expected probability of employee attrition by 50% if they studied life sciences, 81% if studied marketing, 45% if they studied medicine, 44% if their studies were "other" oriented, and reduced the probability of attrition by 91% if they had a technical degree.

β_10, the coefficient of Gender at the factor level male is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed every one unit change in gender. This indicates that if an employee is male, they are two times more likely to be retained at an organization.

β_11, β_12, β_13, are the coefficients of Job Involvement on the factor levels of 2, 3, and 4. These coefficients in our probabilistic logit model change the expected probability of employee attrition while all other factors are fixed. Specifically, a one-unit change in any of

these factor levels decreases the expected probability of employee attrition by 46% if they are involved at a level 2, 33% if they were involved at a level 3, and 20% if they are involved at a level 4.

β_14, β_15, β_16, β_17, are the coefficients of Job Level on the factor levels of 2, 3, 4, and 5. These coefficients in our probabilistic logit model change the expected probability of employee attrition while all other factors are fixed. Specifically, a one-unit change in any of these factor levels decreases the expected probability of employee attrition by 3% if they are involved at a level 2, 48% if they were involved at a level 3, 14% if they are involved at a level 4, and increased their probability of retention by 22% if they were at the job level 5.

β_18, β_19, β_20, β_21, β_22, β_23, β_24, β_25, are the coefficients of Job Role on the factor levels of Human Resources, Laboratory Technician, Manager, Manufacturing Director, Research Director, Research Scientist, Sales Executive, and Sales Representative. These coefficients in our probabilistic logit model change the expected probability of employee attrition while all other factors are fixed. Specifically, an employee is 4 times more likely to stay on if they work in human resources or are lab techs, however, if they are a manager, they are 17 times less likely to attrit, and 10 time more likely to stay on if they work as a director in manufacturing. If the employee works as a research director they are 35% less likely to attrit twice as likely to leave if they work as a research sciences, 3 times more like to stay on if they work in sales as an executive, and 8 times more likely to stay if they work as a sales representative.

β_26, the coefficient of Number of companies worked is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee has worked within specified range of years, they are more likely to stay longer at a company.

β_27, the coefficient of overtime is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee works overtime, they are 4 times more likely to stay on than someone who doesn't work overtime.

β_28, β_29, β_30, are the coefficients of Stock Option Level on the factor levels of 1, 2, 3. These coefficients in our probabilistic logit model change the expected probability of employee attrition while all other factors are fixed. Specifically, a one-unit change in any of these factor levels decreases the expected probability of employee attrition by .34 times if at stock level 1, .25 times less likely if at stock level 2, and .66 times less likely if at stock level 3.

β_31, the coefficient of Total Working Years is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee has been working for a specified period of times they are 93% more likely to attrit.

β_32, the coefficient of Training Times Last Year is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee has been working for a specified period of times they are 93% more likely to attrit.

β_33, the coefficient of Years At Company is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee has been working for a specified period of times, they are 12% more likely to attrit.

β_34, the coefficient of Years In Current Role is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee has been working for a specified period of times they are 87% more likely to attrit.

β_35, the coefficient of Years With Current Manager is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed if an employee has been working for a specified period of times they are 89% more likely to attrit.

β_36, the coefficient of Overall Satisfaction is the coefficient of a probabilistic logit model that changes the expected probability of employee attrition, while all other factors are fixed

if an employee has been working for a specified period of times they are 71% more likely to attrit.

After examining evidence that reports our model to be valid, useful, and accurate, we will utilize the observations made in this portion of our analysis to create to make predictions regarding employee longevity for future analysis.

The Akaike Information Criterion (AIC) values used in variable selection of the most significant variables to our model at 902.42, identical to the other full model used. Therefore, the variables outlined above are reliable in making predictions of employee attrition given the parameters previously defined.

This indicates that an individuals age, if they are male, if they travel for business, how far away they live from home, level of job involvement, specific job role, how many companies they've worked over the years, if they have over time, stock option level, how long they've worked in their live, how long they've been at the company, how long they've been in their specific role, how long they've worked with their specific manager and overall satisfaction in life significantly aid in predicting whether an employee attrits or not.

Validation

For validation of model accuracy, we utilized a 10-fold cross validation method. We accomplished this by creating a vector with the same length as our data frame, and then randomly assigned the vector values between 1 and 10. We then implemented a for loop and utilized a tenth of the observations as the testing subset before averaging the 10 results for our validated mean accuracy.

Table 5. Overall Model Results

METHOD	ACCURACY	PROS	CONS
SVM	97.12%	• Highest Accuracy • Computationally Cheap • Great for binary classification	• Complex • Not robust Would struggle beyond binary classification
RANDOM FOREST	83.63%	• Robust	• Computationally Expensive • Least accurate • High variance bias.
GLM	90%	• Fantastic for identifying the impact of individual factor levels.	• Computationally Expensive Incredibly complex interpretation

Individual Models

Departments

Departments proved to be relatively unimportant compared to our other dimensions. For SVM and Random Forest it was removed from the model due to decreased accuracy and the risk of overfitting. For GLM however, there were a couple levels that were important for the decisions that the GLM model was making in regards to Attrition. The Sales and Research departments were both significant and increased the probability of attrition by more than 40% (70% in the case of the Sales department). Due to most levels not being important, Department was not implemented in even the GLM.

Marital Status

Marital Status was not an important variable in any of the models. This is probably due to it being a low-value dimension only having 3 levels, which were Single, Married, and Divorced. Referring to figure (Random Forest Importance), we can see that Marital Status did not score very highly, it performed worse in importance than most of the other attributes and dimensions. The only important breakpoint in marital status is whether or not someone is single. Single people are much more likely to roll with the punches and stay than someone married or divorced.

For GLM it was still important, however it wasn't important for either SVM or Random Forest. And within the GLM only being single was significant factor as to whether or not someone stayed.

Job Role

The only important dimension for the Random Forest model proved to be Job Roles, with every other dimension either having no effect or a negative one. This may have been due to the increased variance of the dimension since it had 9 levels. Referring to figure (GLM importance one), there were multiple Job Role levels that were incredibly important to the distribution of Attrition. This is not consistent with our SVM model, where Gender proved to the most important dimension. The importance of job role by our metrics came purely from how Random Forest prioritized it as mentioned earlier, in our actual models it proved to be largely useless.

For the GLM there were several Job Roles that were important for predicting Attrition, such as Sales Representative which had an increase towards them choosing to attrit. There were several other roles that had similar probabilities for attrition such as Research Director, Laboratory Technician, and Human Resources. These roles all had a high likelihood of staying with the company, while the other roles had a far higher chance of leaving. The values for this can be found in the Job Role section of Appendix D.

Education Field

Education Field wasn't important to any of the models. In the GLM none of its levels mattered, in Random Forest it scored worse than most in terms of importance, in SVM it was also not significant. It is consistently insignificant in every single model. It's likely not important to this field because the field of education is very vague, and often doesn't matter to the field someone works in. Someone who is working in software development may have been studying in the field of archaeology, and be working in a completely different field of their study. One's studied field does not guarantee a job in said field.

Gender

We initially found that Gender was not important, this was before we fixed our mistake. After we fixed our mistake, it became the most important dimension for the data set as a whole. In SVM it allowed us to improve our model to 97.12% from around 94%. This is consistent with GLM where we found that men are 55% more likely to attrit than women, this can be observed in the Gender section of Appendix D. Gender became the only dimension used for our SVM and Random Forest Model, as the other dimensions would only contribute to overfitting for minimal increases in accuracy.

Business Travel

Business Travel was one of the least significant factors in Random Forest and SVM, its impact on those models were insignificant. However, this dimension was significant in the logit model where if an individual traveled frequently, they were four times more likely to remain at a company than an individual who didn't travel or traveled rarely. This may be due to the fact that, when travelling, there's a lot of things that you wouldn't be doing in a normal job setting, additionally it could be seen as a small break and as a result it may be a perk to many.

Key Findings

Individually, many of the dimensions were not useful for SVM or Random Forest. However, most dimensions indicated a higher joint importance when implemented in the logit link model as well as high significance individually. This is due to the factor levels in each of the dimensions and the function of the GLM probabilistic classifier. There were also clear levels that were more important than others according to figure (GLM deep learning graph) where the minimal levels may have contributed to the model's greater reliability in accurately identifying employees who would attrit out of the many who would not. However, the dimensions also had levels that were not important to the decision of the model and showed specific breakpoints in the Overall Satisfaction variable at 13 and 4, which is indicative of lower or high satisfaction being important for an employees decision to quit or attrit.

There is potential for overfitting our GLM model, but this may also be due to interactions between the dimensions and other factors that were not properly accounted for. We were able to bring our SVM model down from 12 factors to merely 6, with Gender being the sole dimension used. These were also employee metrics, indicating that the employee perspective is more important for attrition than employer based metrics such as performance rating.

Conclusion

Attrition is a hard problem to model, especially with survey data being inconsistent in nature. The most important dimension for our data proved to be Gender, despite our initial findings that it was not very important. This shows that the stress imparted on each gender is different, which was also consistent with other studies[4](gender source). Gender based stress will contribute to a lack of attrition in women, this can lead of a severe brain drain in a company if not combated.

We also found that the corporate metrics for employee success such as Performance Rating were largely irrelevant, and the employee metrics had more significance. When combined as a sum, the satisfaction values were incredibly important and showed how the general stress level of an employee will impact their decision to leave a company. If we were to apply the techniques in this study to a real HR data set, it would be worth splitting the metrics between employee and employer metrics as we did to observe the effects.

In terms of modeling performance, the Support Vector Machine modeling method proved to be the most accurate and it utilized the fewest variables to attain that accuracy with the least chance of overfitting. It was able to attain a 97.12% accuracy, compared to Random Forest which was below 85% and the GLM method which was 90%. This is consistent with our initial expectations that SVM would be a good modeling method due to hyper planes, and due to the less complex and cheaper implementation it proves to be the best model by far.

Adding monetary value to attrition would be a good route for further study on the subject, to observe the cost-benefit ratio of solving any important factor during analysis. The cost to train a new employee is the hard cost for failure to attrit, there are also the soft costs of lost productivity for a worker who has become acclimated to the working environment. We found that employee satisfaction was an important predictor for attrition, and another study found that

higher employee satisfaction can also lead to more profits for a company on average [Di Miceli da Silveira]. So, a high turnover rate can not only have immediate costs of hiring a new employee but also the long-term profit loss caused by lower employee satisfaction that is indicative of a high turnover environment.

Correcting Our Mistakes

During our modeling phase of the experiment, we made a crucial mistake with the factorization of our variables. This led to our SVM model not being able to accurately predict Attrition as our hyperplanes were not properly tuned. This also resulted in Random Forest having a higher accuracy than it realistically should due to overfitting, and it resulted in our GLM algorithm slightly underperforming. This also led to us giving misleading results in our presentation for the project. But we identified the problem and found a massive increase in the accuracy of our models.

References

Alsheref, Fahad Kamal, et al. "Automated Prediction of Employee Attrition Using Ensemble Model Based on Machine Learning Algorithms." Computational Intelligence & Neuroscience, June 2022, pp. 1–9. EBSCOhost, https://doi-org.ezproxy.bellevuecollege.edu/10.1155/2022/7728668.

Blatter, Marc, et al. "Hiring Costs for Skilled Workers and the Supply of Firm-Provided Training." Oxford Economic Papers, vol. 68, no. 1, 2016, pp. 238–57. JSTOR, http://www.jstor.org/stable/43772907. Accessed 17 June 2023.

Di Miceli da Silveira, Alexandre. "The Employee Is Always Right: Employee Satisfaction and Corporate Performance in Brazil." RAC - Revista de Administração Contemporânea, vol. 23, no. 6, Nov. 2019, pp. 739–64. EBSCOhost, https://doi-org.ezproxy.bellevuecollege.edu/10.1590/1982-7849rac2019190224.

Fenech, Angel Ellul, et al. "Gender-based Exclusionary Practices in Performance Appraisal." Gender, Work & Organization, vol. 29, no. 2, Mar. 2022, pp. 427–42. EBSCOhost, https://doi-org.ezproxy.bellevuecollege.edu/10.1111/gwao.12768.

Hunsucker, John L. "Another Look At A Payroll Savings Incentive Plan as An Aid In Curbing Employee Attrition." Industrial Management, vol. 24, no. 3, May 1982, p. 23. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=4989458&site=ehost-live.

Jayathilaka, Pabasara, and Prabhu Subasinghe. "Work Stress and Job Performance in Relation to Gender and Years of Employment." Sri Lanka Journal of Population Studies, vol. 15/16, Apr. 2016, pp. 89–102. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=116136593&site=ehost-live

KESAVAN, L., and S. DHIVYA. "A Study On Causes Of Employee Attrition." Journal of Pharmaceutical Negative Results, vol. 13, Oct. 2022, pp. 479–83. EBSCOhost, https://doi-org.ezproxy.bellevuecollege.edu/10.47750/pnr.2022.13.S08.62.

OSAZEVBARU, HENRY OSAHON, and PETER EGWAINIOVO AMAWHE. "Empirical Narratives on Workplace Environment and Employees' Performance Nexus: New Evidence from the Knowledge Industry." Journal of Academic Research in Economics, vol. 13, no. 3, Dec. 2021, pp. 422–41. EBSCOhost, search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=153990516&site=ehost-live.

Redelinghuys, Kleinjan, et al. "Workplace Flourishing: Measurement, Antecedents and Outcomes." SAJIP: South African Journal of Industrial Psychology, vol. 45, Jan. 2019, pp. 1–11. EBSCOhost, https://doi-org.ezproxy.bellevuecollege.edu/10.4102/sajip.v45i0.1549.

Rustam, Furqan, et al. "Review Prognosis System to Predict Employees Job Satisfaction Using Deep Neural Network." Computational Intelligence, vol. 37, no. 2, May 2021, pp. 964–90. EBSCOhost, https://doi-org.ezproxy.bellevuecollege.edu/10.1111/coin.12440.

Soriano, Aida, et al. “Employees’ Work Patterns–Office Type Fit and the Dynamic Relationship Between Flow and Performance.” Applied Psychology: An International Review, vol. 70, no. 2, Apr. 2021, pp. 759–87. EBSCOhost, https://doi-org.ezproxy.bellevuecollege.edu/10.1111/apps.12251.

“AUTOML: Automatic Machine Learning¶.” *AutoML: Automatic Machine Learning - H2O 3.40.0.4 Documentation*, docs.h2o.ai/h2o/latest-stable/h2o-docs/automl.html. Accessed 16 June 2023.

Cimentada, Jorge. “Producing Stargazer Tables with Odds Ratios and Standard Errors in R.” *Github*, cimentadaj.github.io/blog/2016-08-22-producing-stargazer-tables-with-odds-ratios-and-standard-errors-in-r/producing-stargazer-tables-with-odds-ratios-and-standard-errors-in-r/. Accessed 16 June 2023.

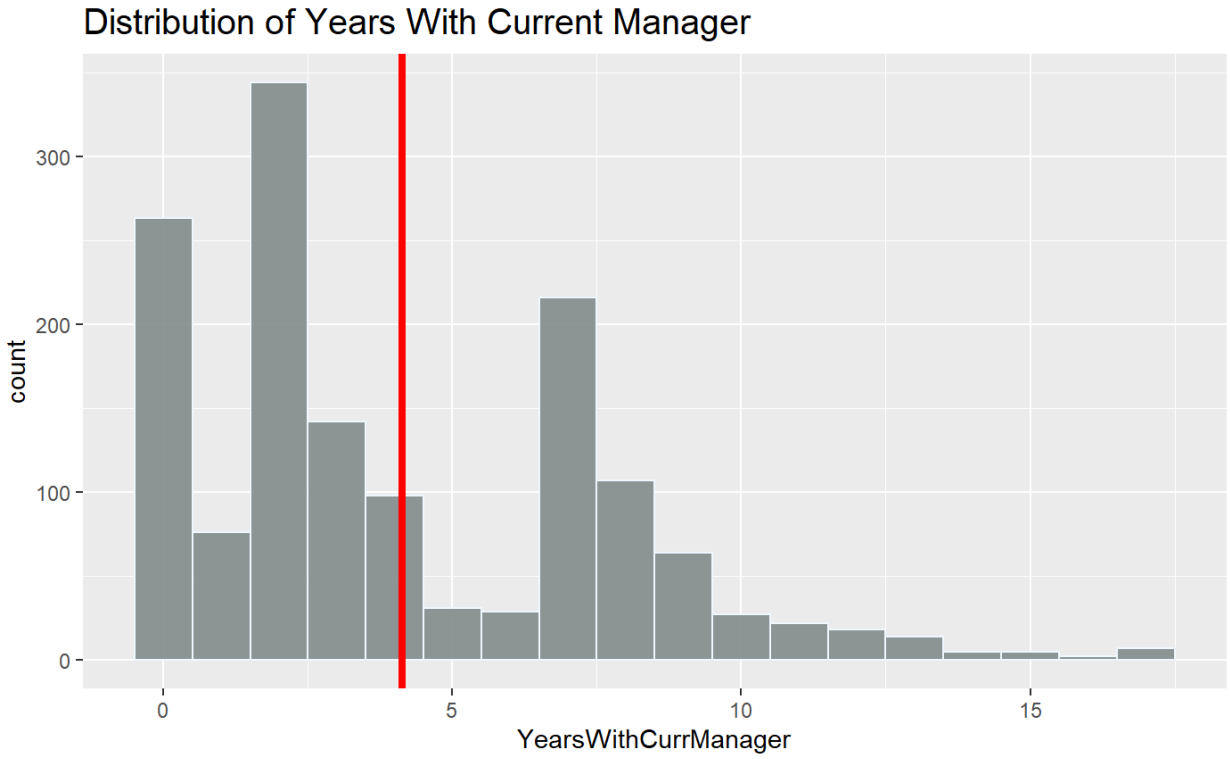
Datalab, Analyttica. “Hosmer-Lemeshow Goodness-of-Fit Test.” *Medium*, 28 Jan. 2019, medium.com/@analyttica/hosmer-lemeshow-goodness-of-fit-test-65b339477210.

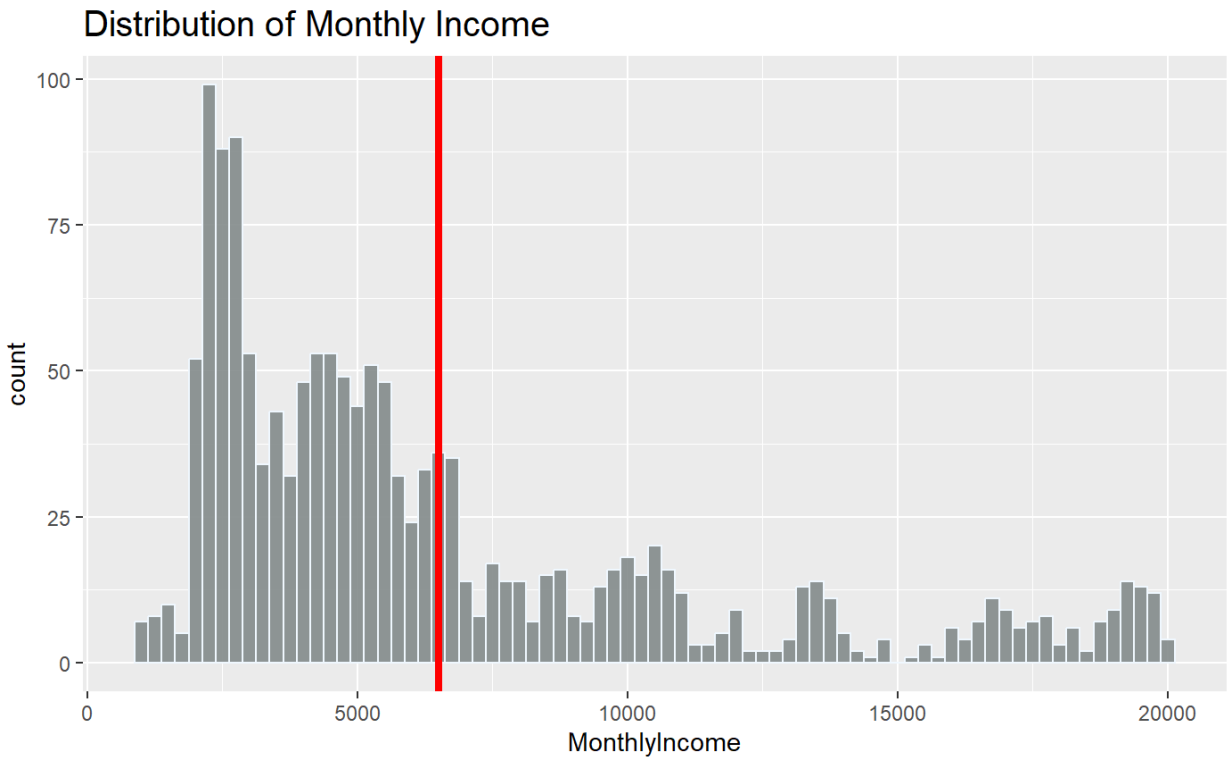
“Show Approval Color Palette Image Format.” *" Blue " SchemeColor.Com*, www.schemecolor.com/show-approval.php. Accessed 16 June 2023.

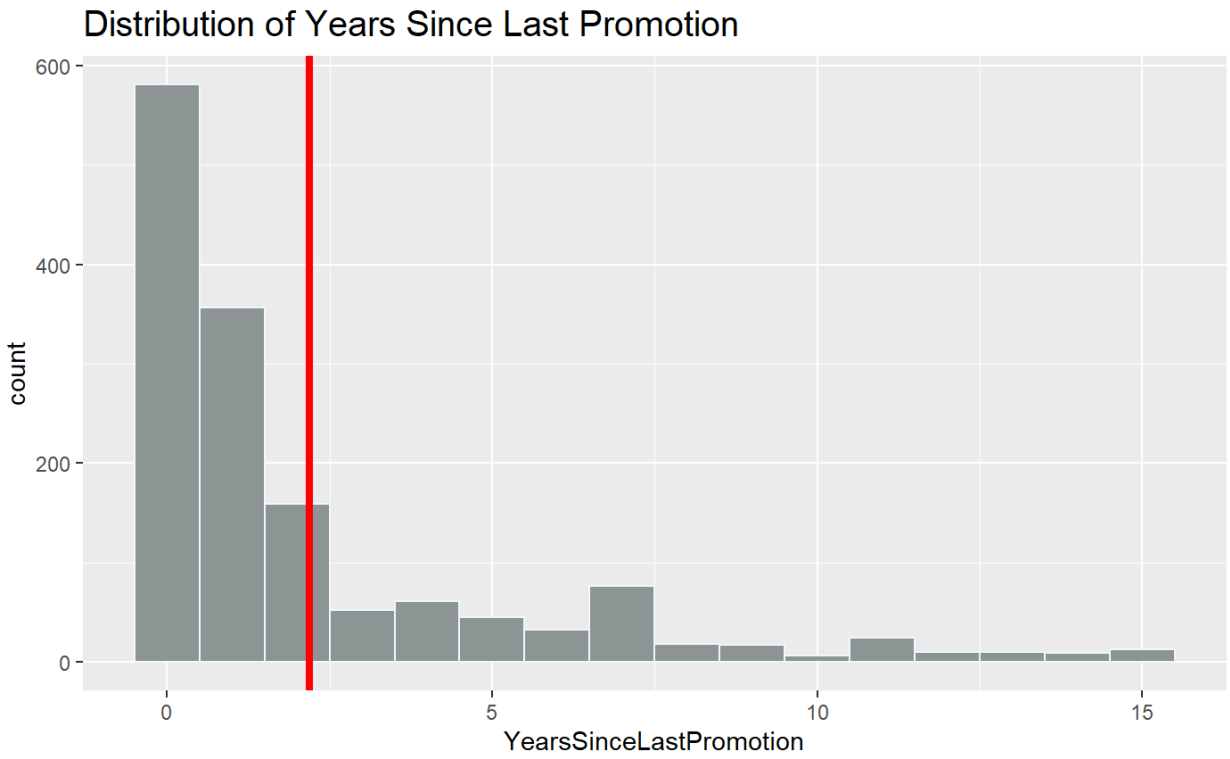
Stephanie. “Hosmer-Lemeshow Test: Definition.” *Statistics How To*, 14 Dec. 2020, www.statisticshowto.com/hosmer-lemeshow-test/.

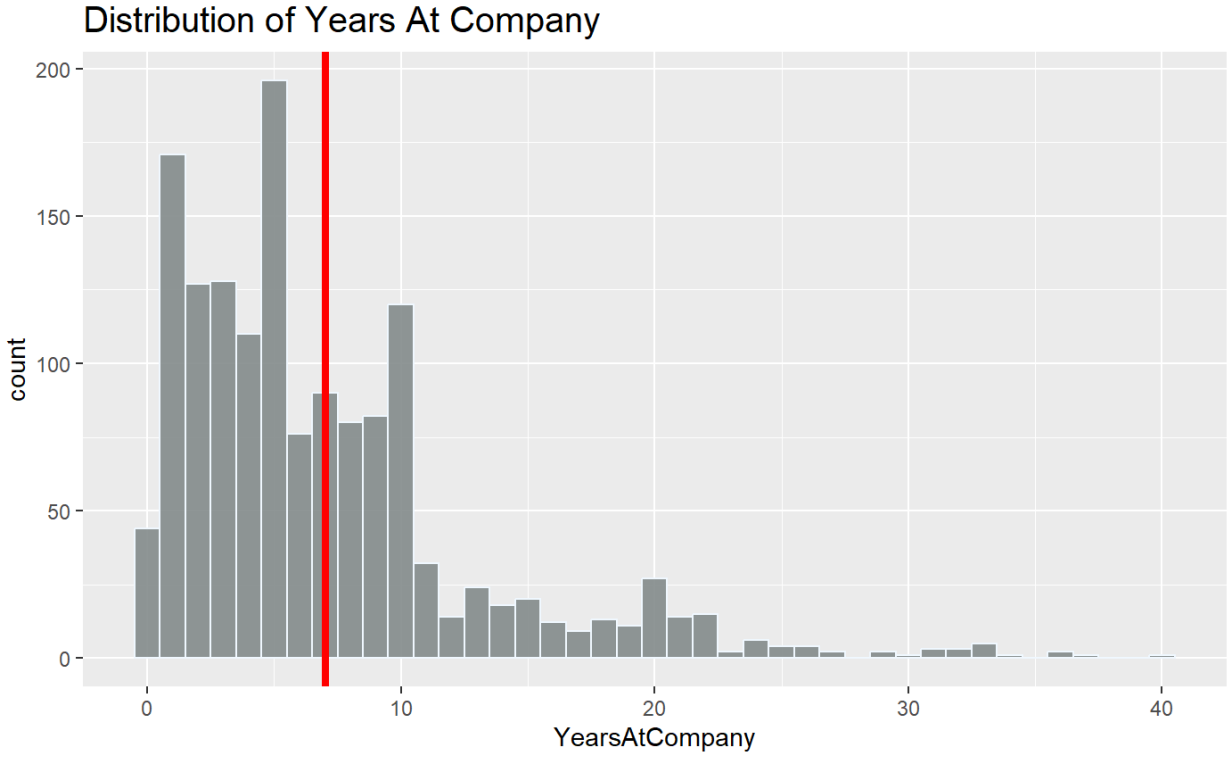
Appendix

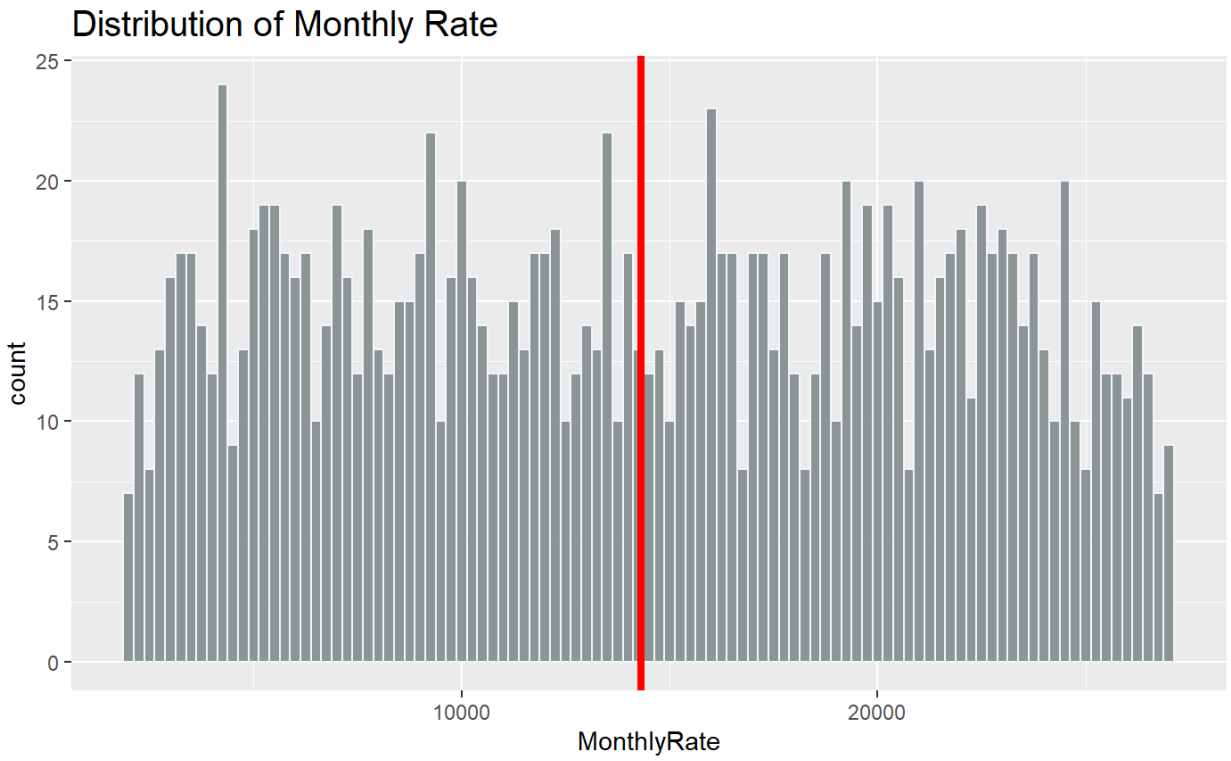
A. Histograms of Abnormal Factors

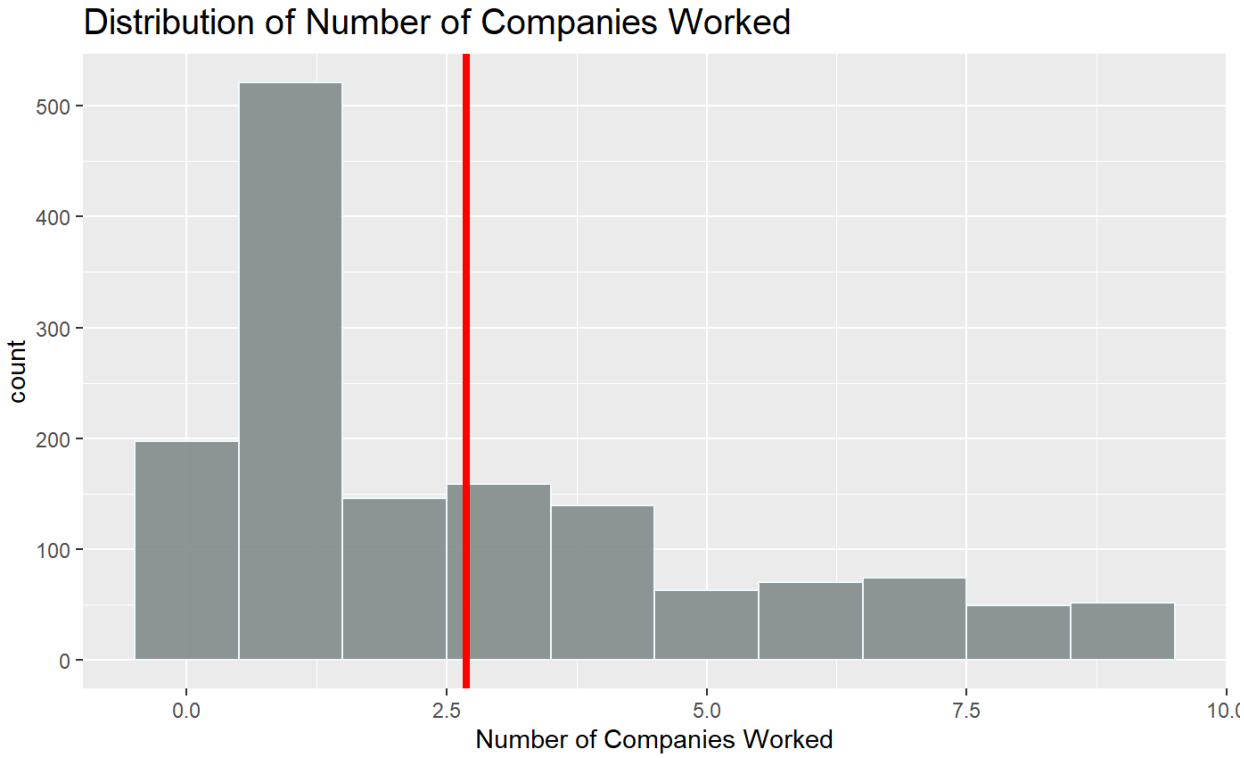


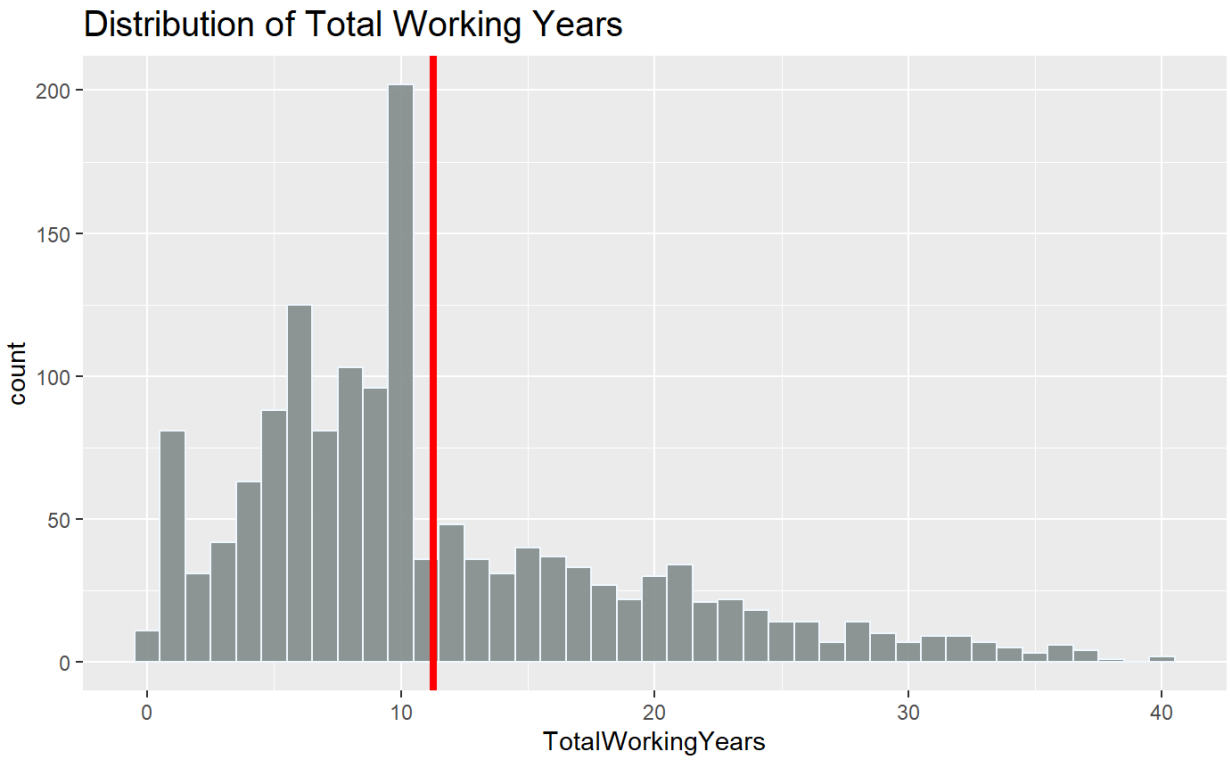



Distribution of Years Since Last Promotion
600
400
200
0
count
0
5
10
15
YearsSinceLastPromotion

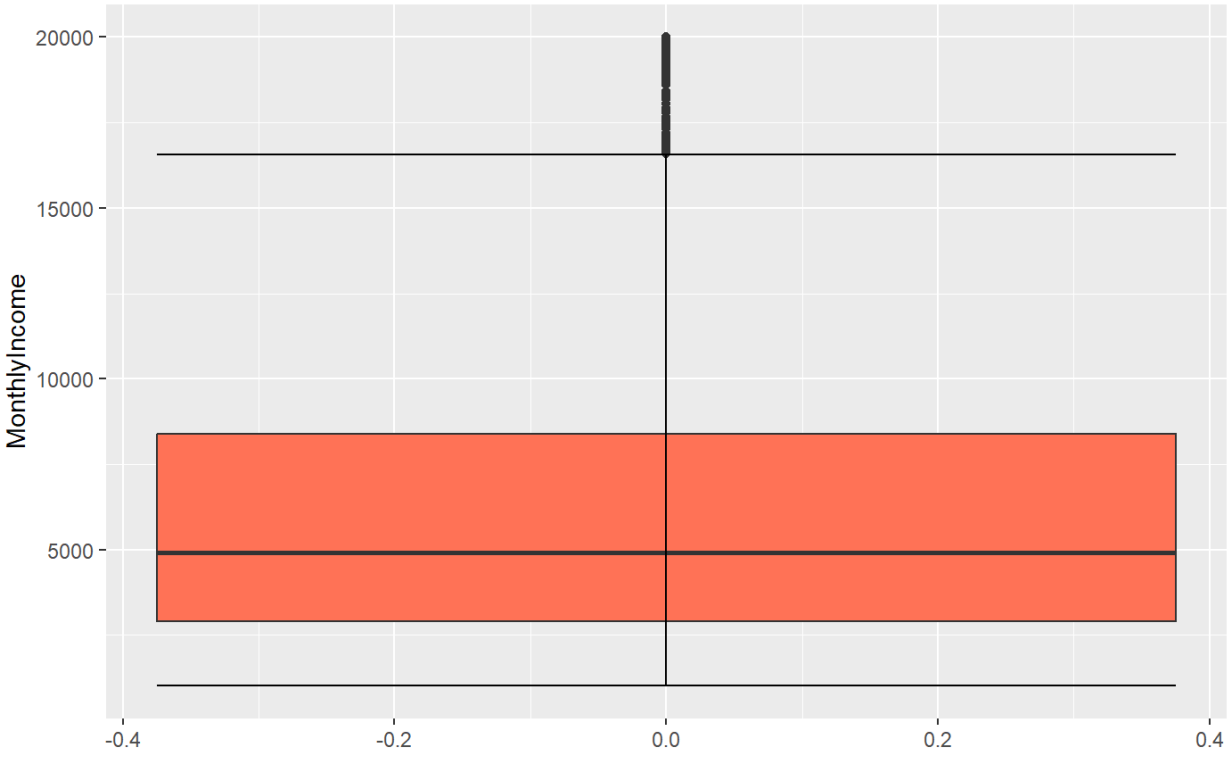

Distribution of Years At Company
200
150
100
50
0
count
0
10
20
30
40
YearsAtCompany

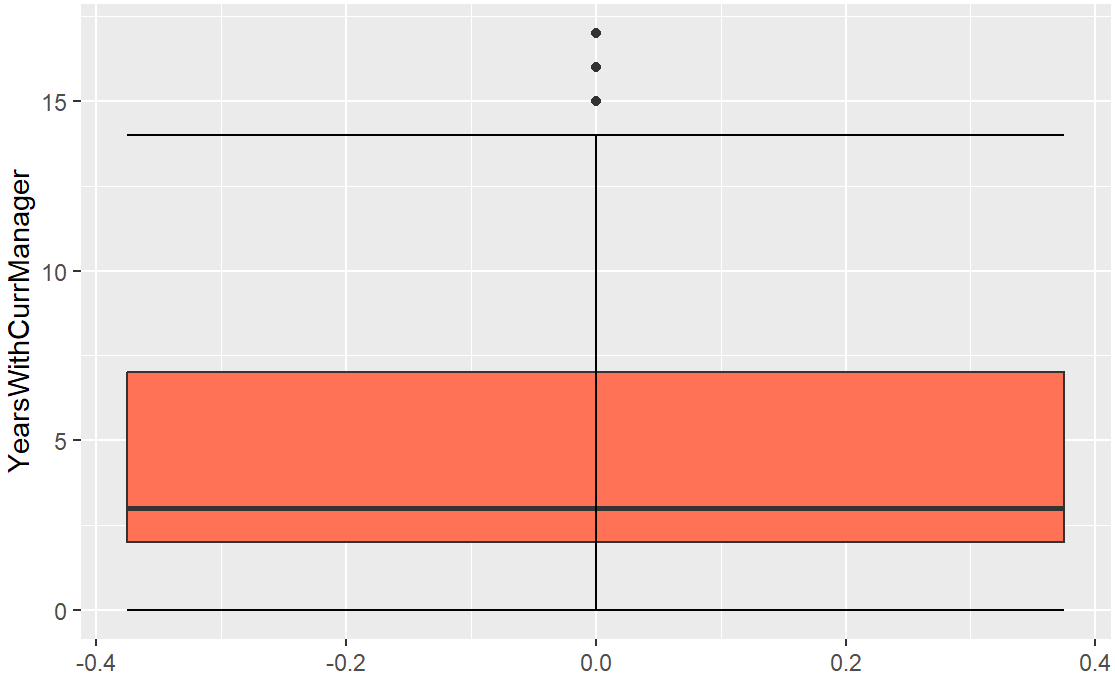

Distribution of Monthly Rate
25
20
15
10
5
0
count
10000
20000
MonthlyRate

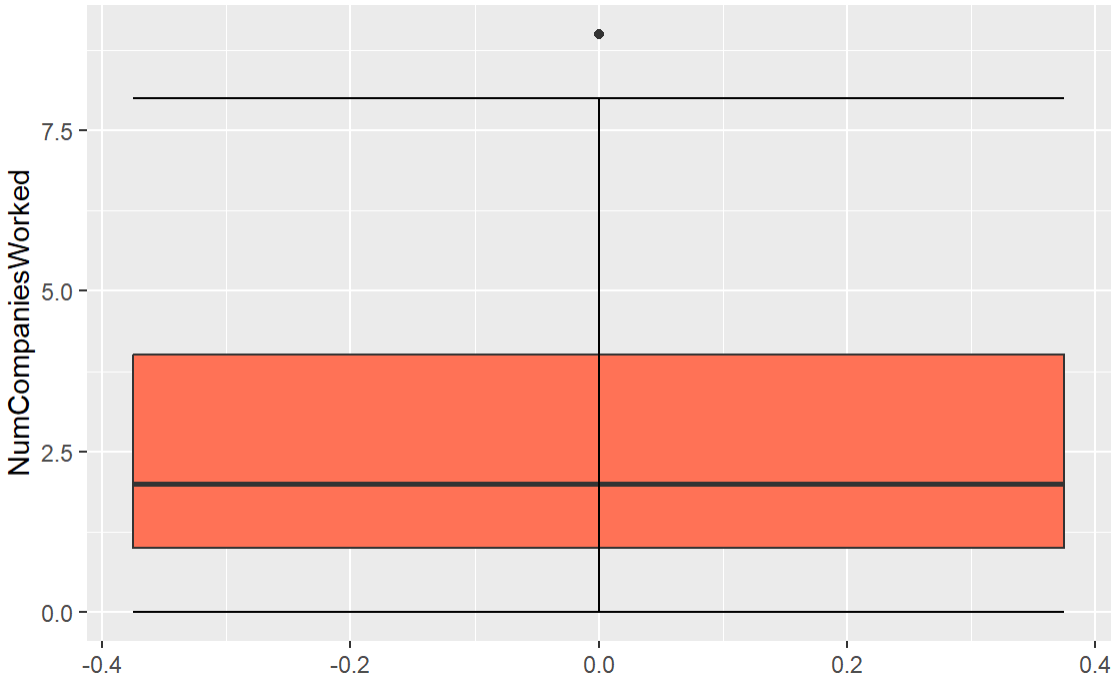

Distribution of Number of Companies Worked
500
400
300
200
100
0
count
0.0
2.5
5.0
7.5
Number of Companies Worked

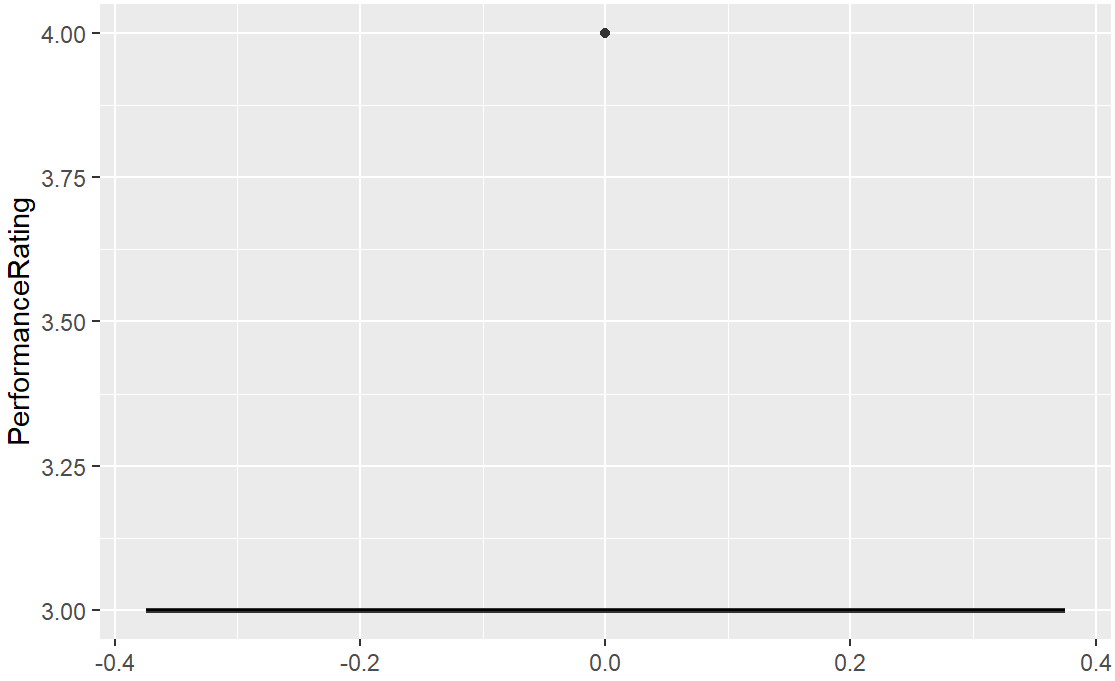

Distribution of Total Working Years
200
150
100
50
0
count
0
10
20
30
40
TotalWorkingYears

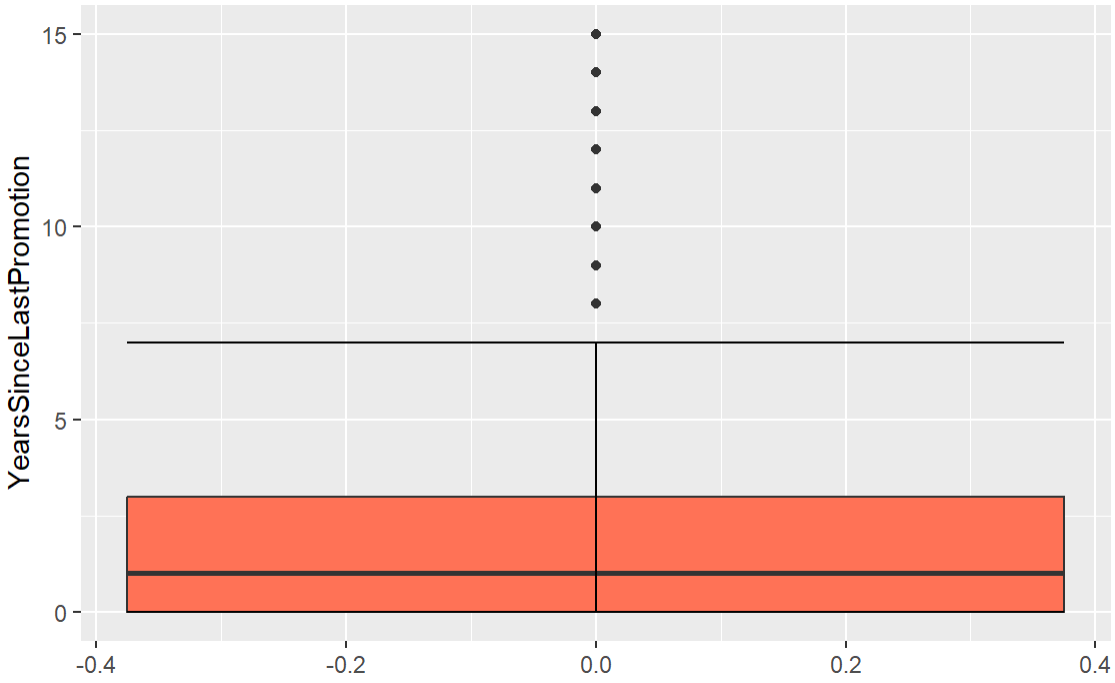
B. Box-Plots of Outliers

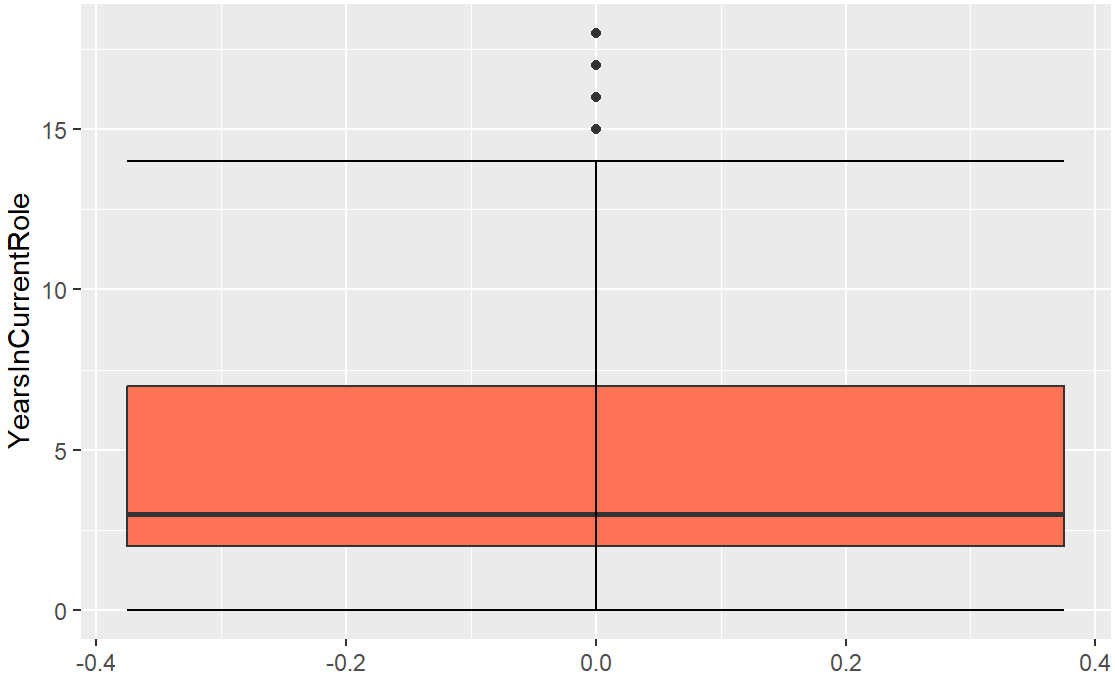


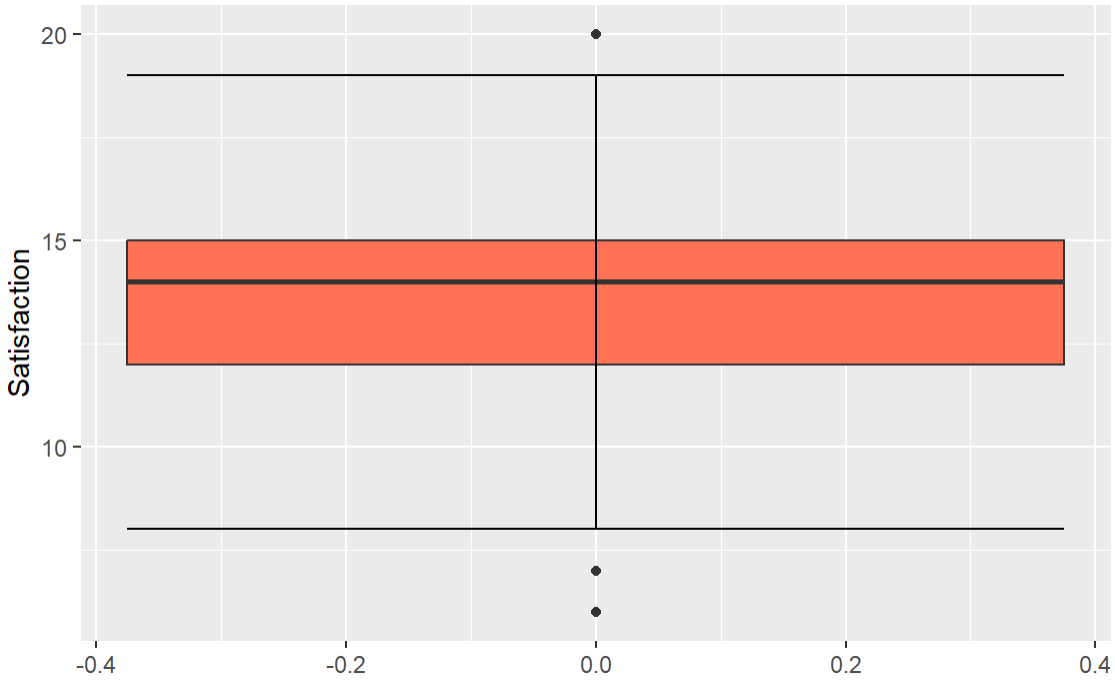

YearsWithCurrManager
15
10
5
0
-0.4
-0.2
0.0
0.2
0.4

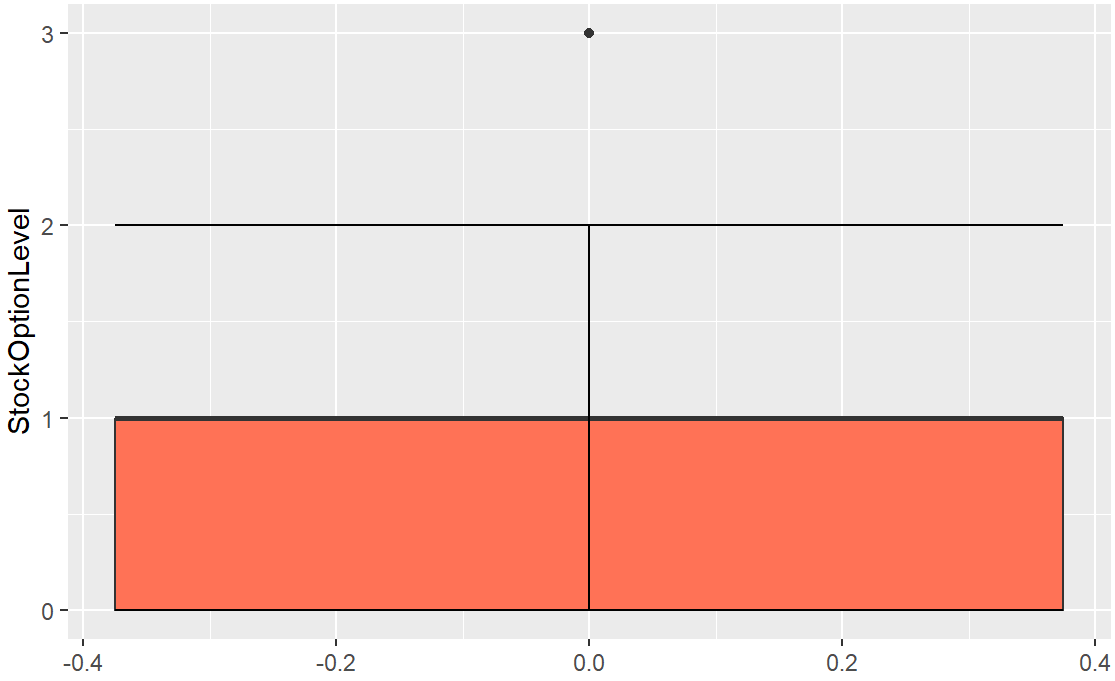

NumCompaniesWorked
7.5
5.0
2.5
0.0
-0.4
-0.2
0.0
0.2
0.4

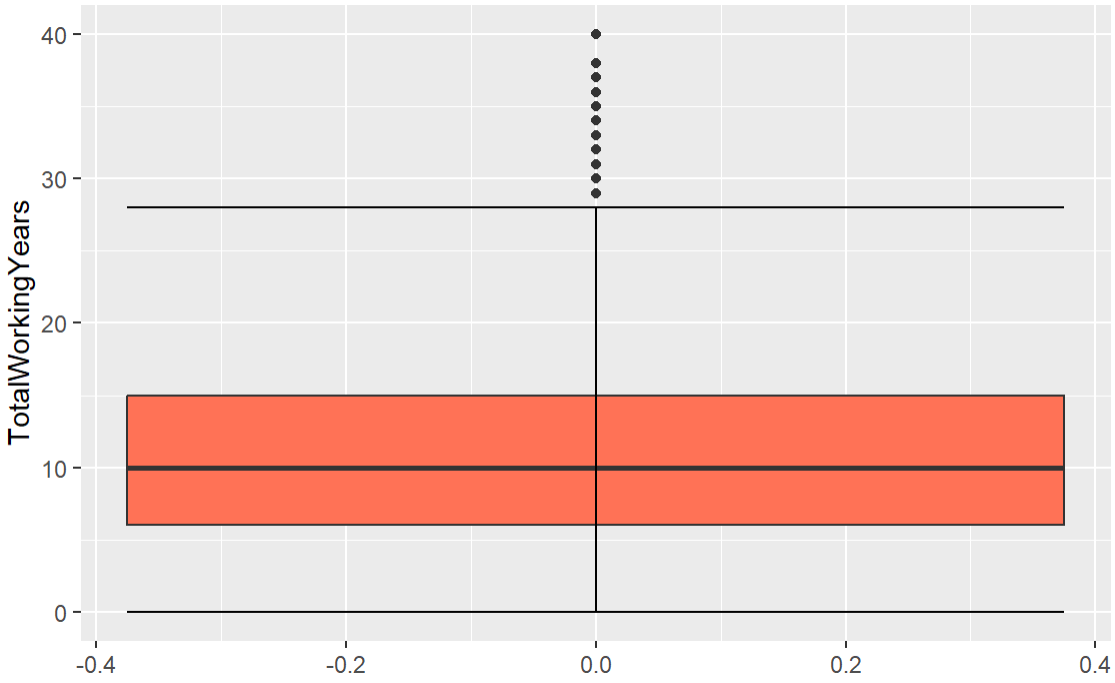

PerformanceRating
4.00
3.75
3.50
3.25
3.00
-0.4
-0.2
0.0
0.2
0.4

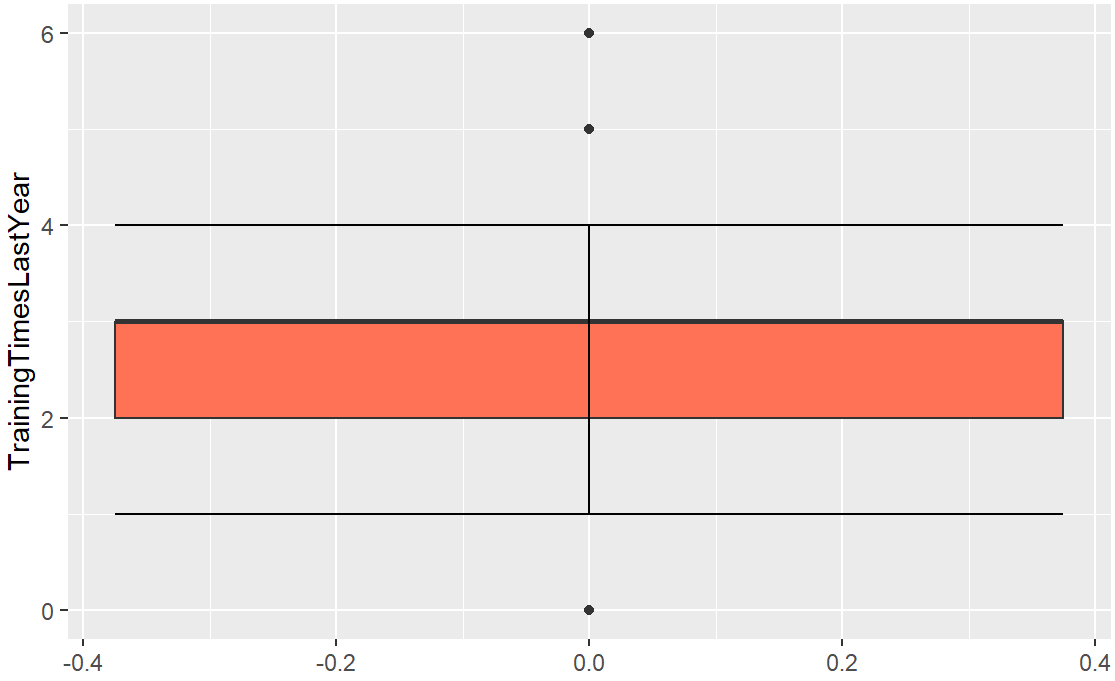

YearsSinceLastPromotion
15
10
5
0
-0.4
-0.2
0.0
0.2
0.4

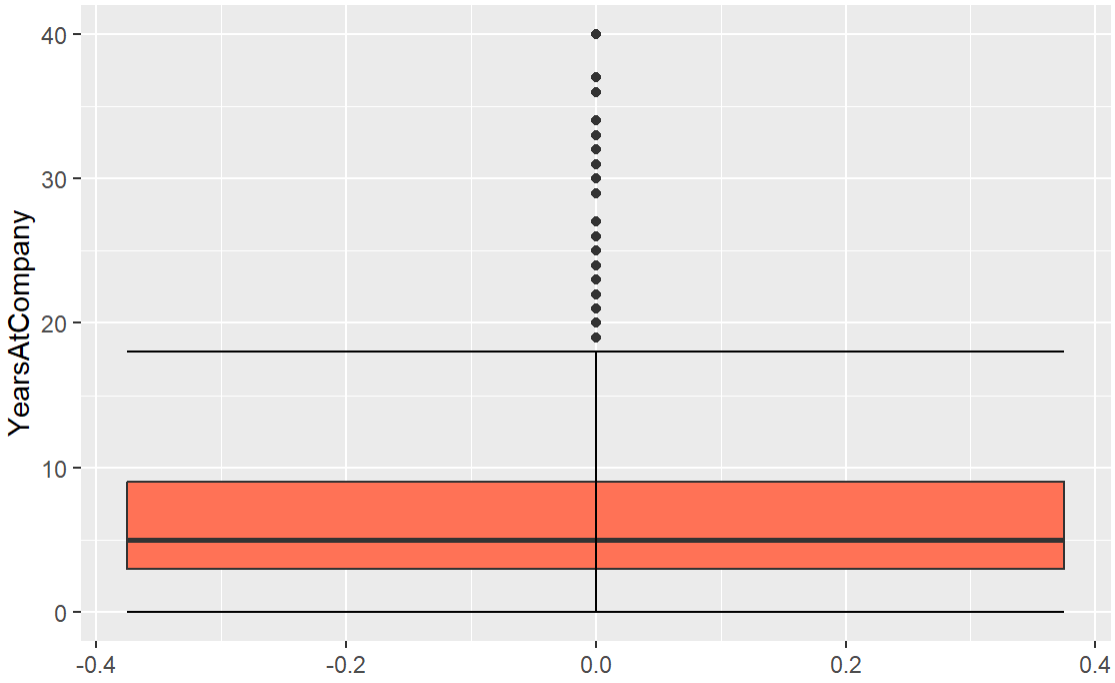

YearsInCurrentRole
15
10
5
0
-0.4
-0.2
0.0
0.2
0.4


Satisfaction
20
15
10
-0.4
-0.2
0.0
0.2
0.4


StockOptionLevel
3
2
1
0
-0.4
-0.2
0.0
0.2
0.4

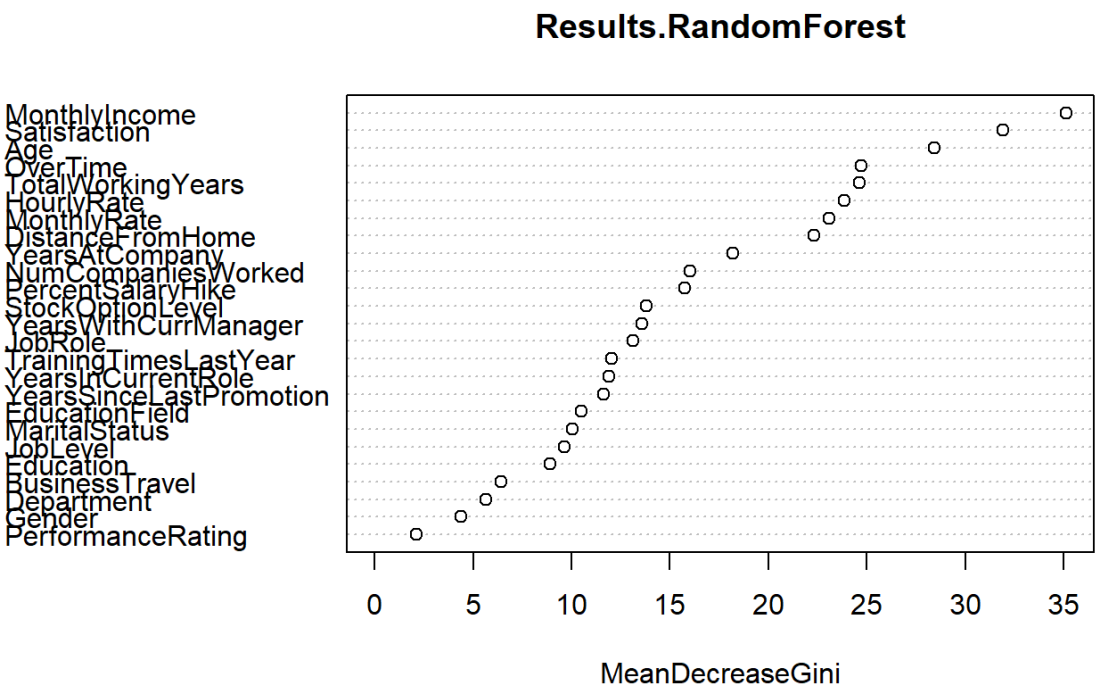

TotalWorkingYears
40
30
20
10
0
-0.4
-0.2
0.0
0.2
0.4


TrainingTimesLastYear
6
4
2
0
-0.4
-0.2
0.0
0.2
0.4

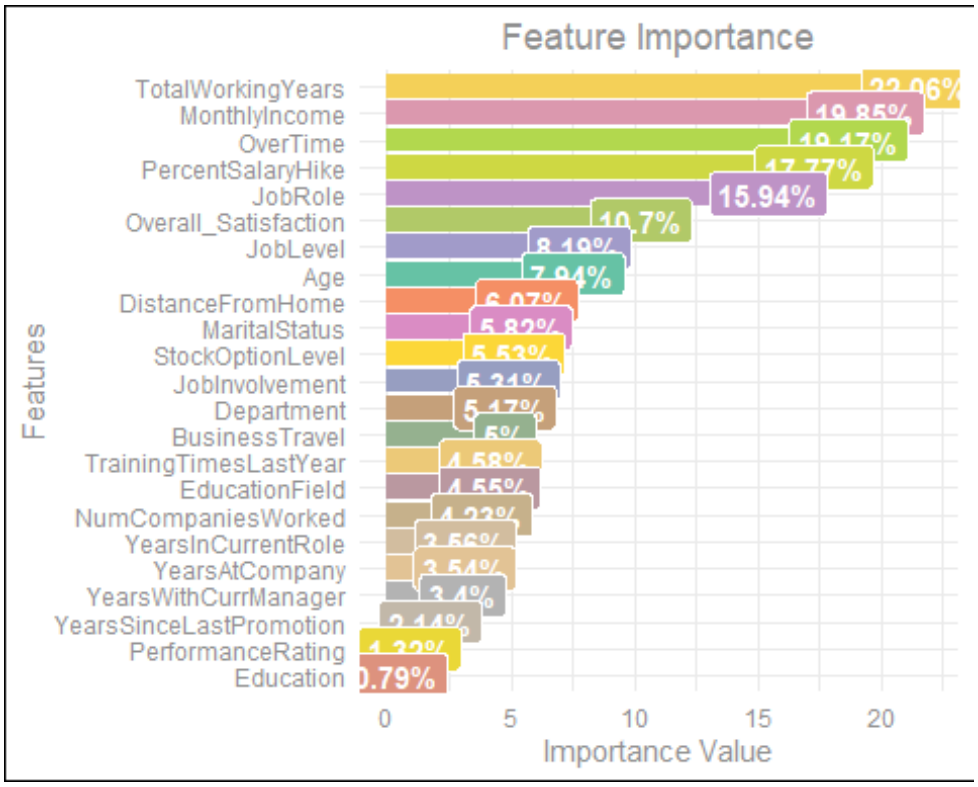

YearsAtCompany
40
30
20
10
0
-0.4
-0.2
0.0
0.2
0.4

C. Feature Importance

Random Forest



GLM


Feature Importance
TotalWorkingYears
MonthlyIncome
OverTime
PercentSalaryHike
JobRole
Overall_Satisfaction
JobLevel
Age
DistanceFromHome
MaritalStatus
StockOptionLevel
JobInvolvement
Department
BusinessTravel
TrainingTimesLastYear
EducationField
NumCompaniesWorked
YearsInCurrentRole
YearsAtCompany
YearsWithCurrManager
YearsSinceLastPromotion
PerformanceRating
Education
15.94%
0.79%
Features
0
5
10
15
20
Importance Value

D. GLM LogOdds Coefficients

Department

```
stargazer(department,type = "text")
```

```
##
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                Attrition
## ------------------------------------------------
## DepartmentResearch           Development
##                                (0.334)
##
## DepartmentSales                 0.099
##                                (0.342)
##
## Constant                       -1.447***
##                                (0.321)
##
## ------------------------------------------------
## Observations                    1,470
## Log Likelihood                -644.046
## Akaike Inf. Crit.             1,294.092
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Business Travel

```
stargazer(bizT,type = "text")
```

```
##
## ============================================================
##                                      Dependent variable:
##                                  ---------------------------
##                                         Attrition
## ------------------------------------------------------------
## BusinessTravelTravel_Frequently          1.339***
##                                           (0.331)
##
## BusinessTravelTravel_Rarely               0.704**
##                                           (0.313)
##
## Constant                                 -2.442***
##                                           (0.301)
##
## ------------------------------------------------------------
## Observations                               1,470
## Log Likelihood                           -637.411
```

```
## Akaike Inf. Crit.                          1,280.822
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

Education Field

```
stargazer(EduF,type = "text")
```

```
##
## ============================================================
##                                    Dependent variable:
##                                ---------------------------
##                                          Attrition
## ------------------------------------------------------------
## EducationFieldLife Sciences                -0.710
##                                            (0.454)
##
## EducationFieldMarketing                    -0.215
##                                            (0.479)
##
## EducationFieldMedical                      -0.801*
##                                            (0.460)
##
## EducationFieldOther                        -0.815
##                                            (0.546)
##
## EducationFieldTechnical Degree             -0.090
##                                            (0.484)
##
## Constant                                  -1.050**
##                                            (0.439)
##
## ------------------------------------------------------------
## Observations                                1,470
## Log Likelihood                            -641.841
## Akaike Inf. Crit.                         1,295.683
## ============================================================
## Note:                            *p<0.1; **p<0.05; ***p<0.01
```

Gender

```
stargazer(MF,type = "text")
```

```
##
## ================================================
##                        Dependent variable:
##                    ---------------------------
##                              Attrition
```

```
## ------------------------------------------
## Gender1                     0.166
##                            (0.147)
##
## Constant                   -1.751***
##                            (0.116)
##
## ------------------------------------------
## Observations                1,470
## Log Likelihood             -648.649
## Akaike Inf. Crit.          1,301.297
## ==========================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Job Role

```
stargazer(JR,type = "text")
```

```
##
## ========================================================
##                                  Dependent variable:
##                              ---------------------------
##                                        Attrition
## --------------------------------------------------------
## JobRoleHuman Resources                 1.403***
##                                         (0.477)
##
## JobRoleLaboratory Technician           1.451***
##                                         (0.375)
##
## JobRoleManager                          -0.358
##                                         (0.574)
##
## JobRoleManufacturing Director            0.004
##                                         (0.476)
##
## JobRoleResearch Director                -1.057
##                                         (0.795)
##
## JobRoleResearch Scientist               0.956**
##                                         (0.380)
##
## JobRoleSales Executive                 1.055***
##                                         (0.375)
##
## JobRoleSales Representative            2.191***
##                                         (0.412)
##
## Constant                               -2.607***
```

```
##                                             (0.345)
##
## ------------------------------------------------------
## Observations                                 1,470
## Log Likelihood                              -604.837
## Akaike Inf. Crit.                           1,227.674
## ======================================================
## Note:                           *p<0.1; **p<0.05; ***p<0.01
```

Marital Status

```
stargazer(MS,type = "text")
```

```
##
## ================================================
##                          Dependent variable:
##                      --------------------------
##                                Attrition
## ------------------------------------------------
## MaritalStatusMarried             0.239
##                                 (0.217)
##
## MaritalStatusSingle             1.117***
##                                 (0.212)
##
## Constant                        -2.187***
##                                 (0.184)
##
## ------------------------------------------------
## Observations                     1,470
## Log Likelihood                  -627.292
## Akaike Inf. Crit.               1,260.583
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```