Capstone Project:

Classification Models: Predicting Employee Churn

(Final Project Feedback: Part F)

Team 1:

Shaurya Mehta and Jay Floresca

Northeastern University

Table of Contents

[Introduction 3](#_Toc9000126)

[Analysis 3](#_Toc9000127)

[Part A: Exploratory Data Analysis 3](#_Toc9000128)

[1. Independent variables 7](#_Toc9000129)

[2. Dependent variable (Churn Status) 7](#_Toc9000130)

[3. Churn status and predictor variables. 8](#_Toc9000131)

[Part B: Data segregation: 10](#_Toc9000132)

[Part C: Logistic Regression Model 10](#_Toc9000133)

[1. Logistic Regression Models 10](#_Toc9000134)

[1.1 Logistic regression model 1: Figure 11 10](#_Toc9000135)

[1.2 Logistic regression model 2: Figure 12 12](#_Toc9000136)

[1.3 Final Logistic regression model: Figure 14 14](#_Toc9000137)

[1.4 Probability threshold tuning over training data set 15](#_Toc9000138)

[1.5 Fitting final model over test dataset. 17](#_Toc9000139)

[Part D: Random Forest Model 18](#_Toc9000140)

[1. RF (Random Forest) Model: Churn Status with the 9 Independent variables. 18](#_Toc9000141)

[2. RF parameter fine-tuning: 18](#_Toc9000142)

[3. Optimal Forest Size 19](#_Toc9000143)

[4. Number of Splits 20](#_Toc9000144)

[5. Number of features 20](#_Toc9000145)

[6. Random Forest Final Model: 21](#_Toc9000146)

[7. RF Final Mode: Training data (the year 2017 and below) 21](#_Toc9000147)

[8. RF Final Model: Test data (2018) 22](#_Toc9000148)

[Part E: Comparing two models 22](#_Toc9000149)

[Part F: CONCLUSION 24](#_Toc9000150)

[Part G: RECOMMENDATION 24](#_Toc9000151)

[**Part H: PROJECT FEEDBACK FROM Q & A** 24](#_Toc9000152)

# Introduction

In this paper, we will be leveraging our Dataset for the capstone project. The dataset is from Glassdoor that contains employee reviews from six companies namely; Amazon, Apple, Facebook, Microsoft, Google and Netflix with 67525 observations and 19 variables. The paper is divided into five parts; Part A showcases Exploratory Data Analysis (EDA) i.e. describing the data by means of statistical and visualization techniques. This helped in surfacing the important aspects of the data which were focused for further analysis. This involved looking at the data and summarizing it without making any assumptions about its contents. Based on the discoveries made, cleaning and dividing the data into training and testing sets were performed. Part B shows how the dataset is divided into training and test data using the years as a threshold. Part C shows the development of the logistic regression model. Part D shows the development of the Random Forest model. Part E compares the two models.

The paper objective is to predict the churn status of the employees by utilizing the machine learning algorithms: Logistic Regression and Random Forest. We will also compare the two models using ROC/AUC and confusion matrix outputs.

Logistic Regression: Logistic regression works by finding the best fitting model to describe the relationship between the binomial or dichotomous characteristic of interest (response or dependent variable, i.e. churn status) and a set of independent or predictor variables. It generates the coefficients of a formula to predict a logit transformation of the probability of the presence of the characteristic of interest.

Random Forest: Random Forest works by creating a forest of Decision Trees, which are trained with the bagging i.e. a combination of learning models which results in overall increases or betterment of the result.

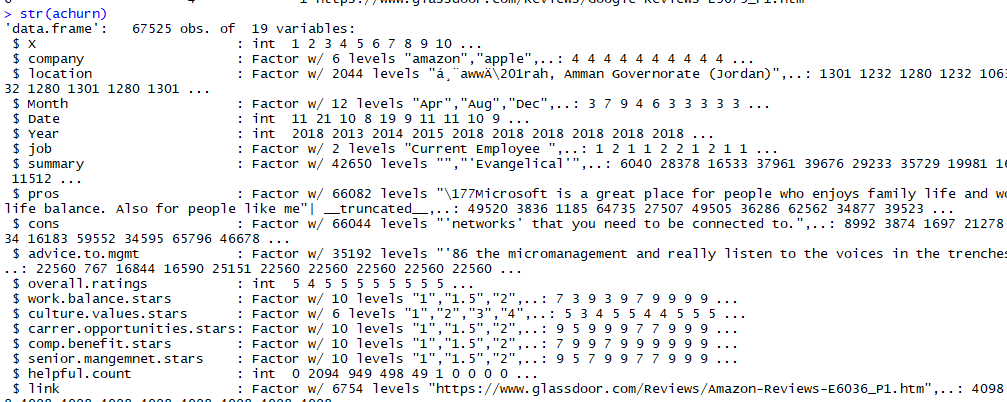
Note: Employee review dataset used for the project will be referred to as "dataset". Multiple transformations have been performed on the same that including renaming it with many miscellaneous names during coding. Figures or tables inserted for reference may contain those.

# Analysis

## Part A: Exploratory Data Analysis

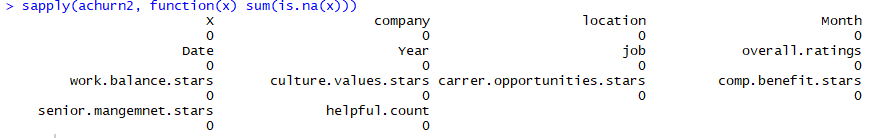
The dataset contains 67529 observations with 17 variables. The table below lists the structure of the dataset. An important part of the Survey uses various ratings of 1 to 5, from lowest to highest. All the columns are all independent even the overall rating is independently scored like the rest of the parameters. There are also narratives on the pros and cons, which could be a part of future models and investigation on this dataset. *(Figure 1).*

*Figure 1*



We have decided to drop some of the variables or columns that we will not be utilized in the model creations and will not contribute to its development. Most of these variables have large texts which created issues on the processing of the dataset as well. We dropped the following columns; link, pros, cons, summary, advice.to.mgmt. After dropping these columns, we checked for missing values, specifically; "NA". The results show there is no ‘NA's'. *(Figure 2)*

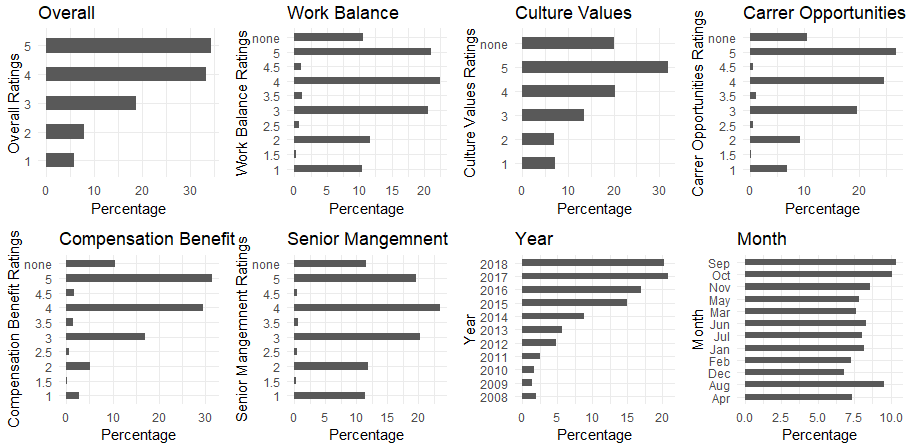
*Figure 2*



Upon further investigation of the dataset, ‘none’ and decimal rating values are discovered for the rating variables.

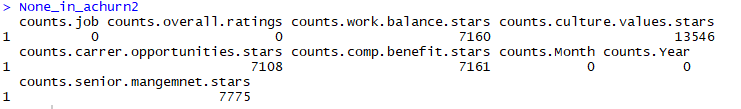
Work balance, career opportunities, compensation benefits, senior management variables had both none values and decimal ratings. Cultural values had none values. *(Figure 3)*

*Figure 3*



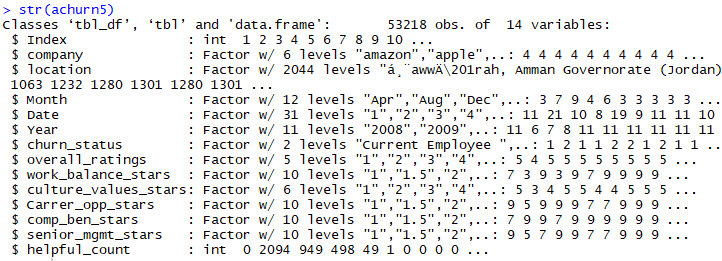
We checked how many rows have ‘none’ to validate if it will substantially affect our data once we develop our models. *(Figure 4)*

*Figure 4*



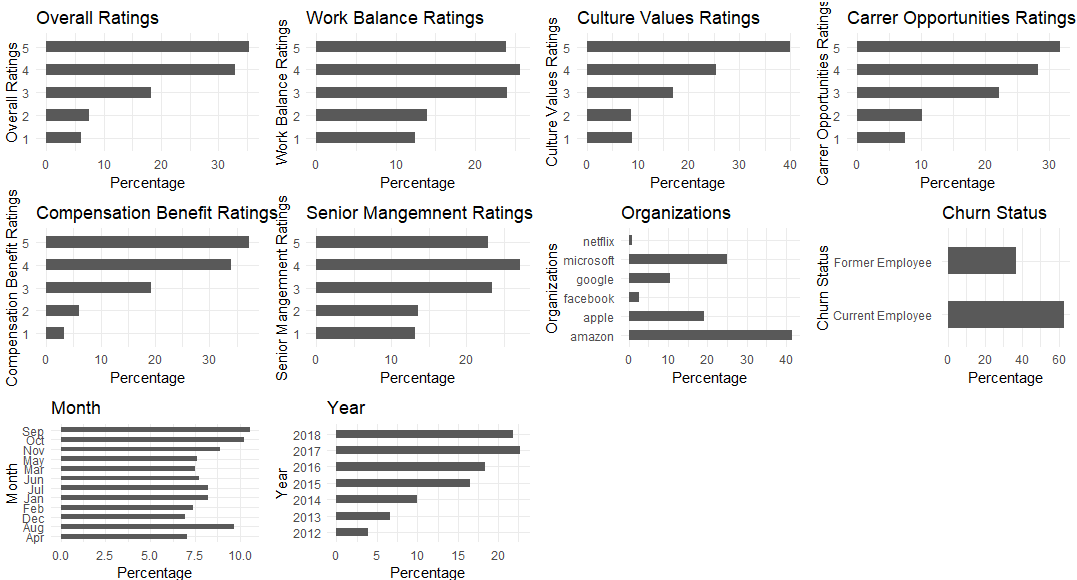
There was a total of 7160 values on work balance column, 13546 values on culture, 7108 values on career opportunities, 7161 values on compensation. As all the ratings are individually filled by the employees and there is no dependency of calculating a missing value for from other, these ratings are removed from the dataset as replacing the same with other mean or median ratings could have introduced bias in the developed model. Partial or decimal rating values are also removed at the same time as they were residing adjacent to the none values. The cleaned dataset now has 53218 observations of 14 variables. Thereby we reduced the number of observations by approximately 14k rows or by approximately 20%. *(Figure 5)*

*Figure 5*



Bar plots were recreated once the data cleaning was completed to visualize the dataset and distribution of the variables. *Figure 6* displays the distribution of the variables (dependent and independent) that are feasible and are included in the model. The backward approach is followed and variables with no or less contribution were dropped going forward.

*Figure 6*



## Independent variables

Overall Ratings- This is an independent survey question and reviewers will have to rate the company overall. In the data set, we see that the majority of the reviews were given on a scale of 4 to 5. We can also see that overall rating from current and former employees that are almost a third that is within ratings of 1-3.

Work Balance Ratings- From the bar graph, we can see that the ratings from 3-5 are almost the same. They don't have wide variations with the rating 4 slightly ahead of the 3 and 5 ratings.

Culture Values- This is the metric which got 40% with a rating of 5 and the next rating of 4 is 25% of the ratings on culture. This the highest ratings all the six companies got among the different variables.

Career Opportunities- The rating for 4 and 5 are 60% of the ratings on this variable with the rating 1-3 comprising just 40%.

Compensation and Benefits- This shows the six companies as rated having good compensation and benefits. The ratings for 4 and 5 combined is about 70% which means that the compensation across the jobs and between former and current employee rates them as very high.

Senior Management- The rating for senior management almost mirrors how the work-life balance was rated by reviewers. The 3-5 ratings are almost the same and the rating 1-2 are on the same percentage as well. This indicated that raters either have a high view of their senior management or has a low view. There is no variability in the middle ranges.

Organizations- Amazon along with Microsoft employees contributed for more than 50% of the reviews while the remaining review was contributed by the current or former employees of Apple, Google, Facebook, and Netflix.

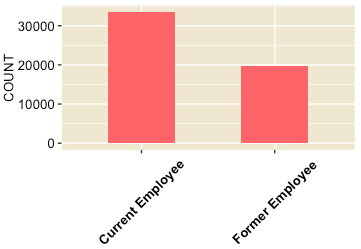
Year- Around 60 % of the reviews are done from the year 2016 to 2018 while the remaining reviews are of the year 2012 to 2015

Month- All the months have even distribution of the number of reviews filled in each of them.

## Dependent variable (Churn Status)

The ratings were given by current and former employees of the six companies. We can see that 60% were given by current employees and 40% by former employees. (Figure 7)

*Figure 7*

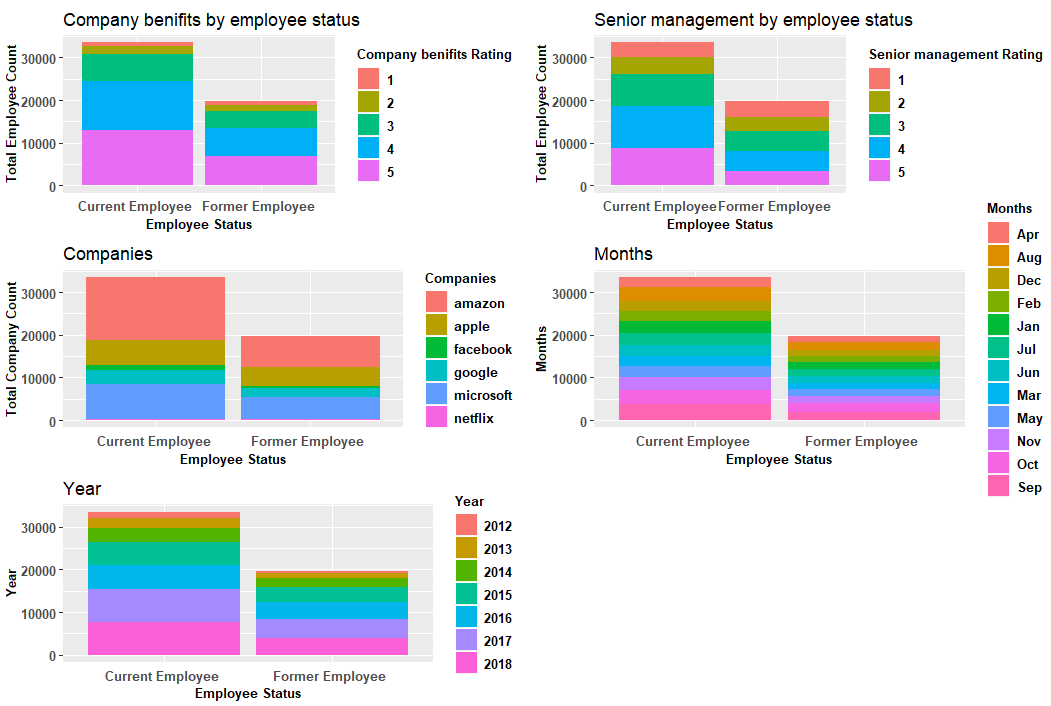


## Churn status and predictor variables.

We overlaid the ratings for each of the variables and what the employee status is. This shows the rating provided by the employee on the various ratings. We can see the pattern of how the two groups rate their companies based on the following variables. *(Figure 8)*

*Figure 8*





## Part B: Data segregation:

Prior to proceeding with the models, cleaned dataset set has been divided into training and testing dataset. The testing data set is segregated with the condition on the Year i.e. it has the observations for the year 2018. This test dataset has 11615 observations which are about 20% of the dataset. *(Figure 9)*

*Figure 9*



Training data set has 41603 observations which belong to the years less than 2018. This training dataset consisted of almost 80% of the dataset. *(Figure 10)*

*Figure 10*

.

## Part C: Logistic Regression Model

## Logistic Regression Models

On this part of the paper, we are developing logistic regression models with 9 independent variables Overall Ratings, Work Balance Ratings, Culture Values ratings, Career Opportunities, Compensation and Benefits, Senior Management, Organizations, Year, Month. Churn status is our dependent variable which we predicted or classified based on the mentioned independent variables. The backward approach is followed, and irrelevant variables have been dropped after evaluating the prepared models.

## Logistic regression model 1: Figure 11

This model is developed on all the independent variables.

The intercept (b0) -.743123 which is the mean value of the log-odd and does not depends on the number of variables included in the model.

The next part of the output shows the coefficients for variables, their standard errors, the z-statistic (sometimes called a Wald z-statistic), and the associated p-values. The logistic regression coefficients give the change in the log odds of the outcome for a one unit increase in the predictor variable.

For example every 1 unit change in a year, the log odds of an employee being a current employee (versus former employee) increases by .099.

The indicator variables for the rest of the independent variables have a slightly different interpretation. For example, being an employee of the apple company, versus an employee of Microsoft, changes the log odds of churn status by .606. In similar way coefficients of each type of ratings listed here showcases the change in the log odds of churn status when compared to the rating 1.

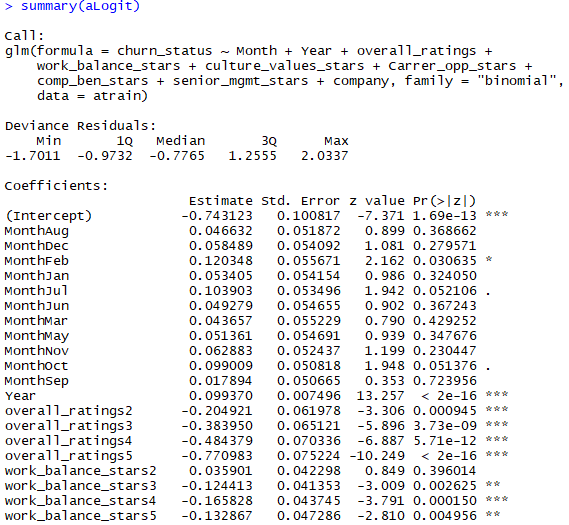
Based on the significance level of the variables which are calculated on the z statistic/p-values following series of the variables is drawn, moving from most important to least important variables (Lesser the p-value more significant the variable is.)

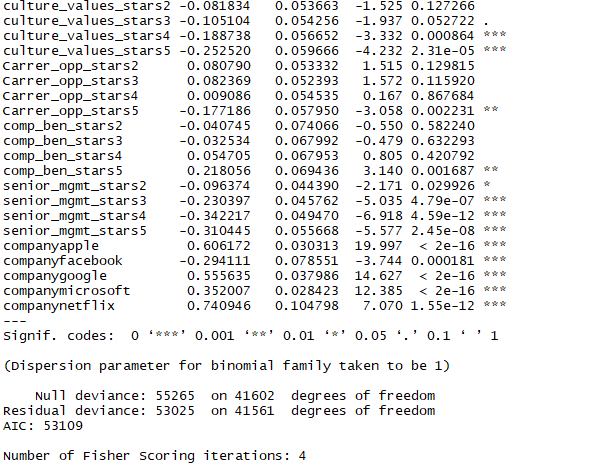
*Most Significant* Company > overall ratings > Year > senior mgmt. stars > work balance stars > culture values stars > Month > comp ben stars > Career opp stars *Least Significant*

Null deviance (55265) is a value representing the fit when we don’t have any variable and only has intercept.

Residual deviance (53025) is a value representing the fit of the model with adjusting the degree of freedom for the included predictors. Lesser the value of deviance means a better fit. Difference between Null deviance and Residual deviance tells us that the model is a good fit.

*Figure 11*



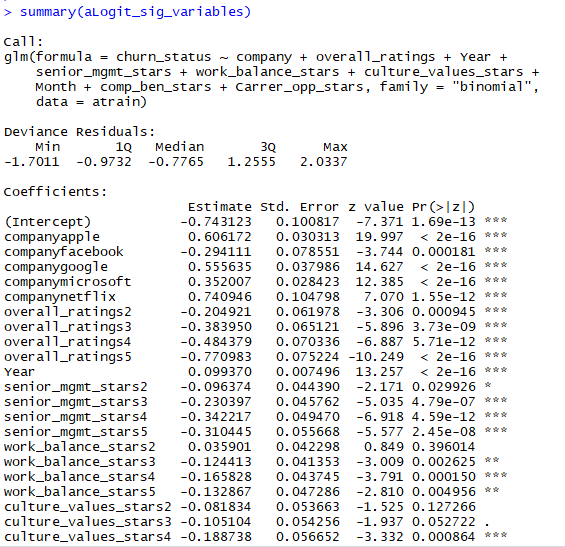


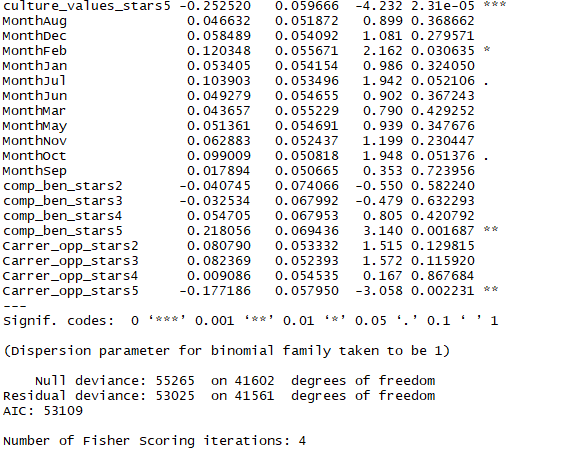
## Logistic regression model 2: Figure 12

In this model number of dependent variables remained the same i.e. all 9 but the key difference here is the sequence of the variables plugged in. The sequence of the variables is from most significant to the least significant variables. The sequence is important for ANOVA (Analysis of variance) test as ANOVA test showcases the variance in the models by adding 1 variable to the next developed model.

The sole purpose of this model is to feed the independent variable with the highest significance first followed by lesser significant variables.

*Figure 12*





ANOVA test on above model (Logistic Regression Model 2)

Analysis of variance test is performed on the model 2 (model with variables fed from highest to lowest significance). Residual deviances are obtained for various models with an increasing number of variables (Figure 13). Following tables explains and lists the results from the ANOVA test, how the addition of a variable to the model is affecting the Residual deviance (Table 1)

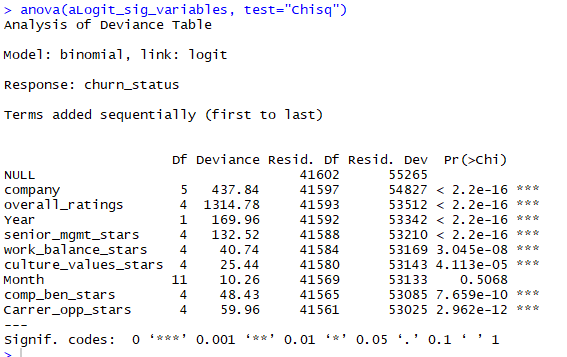
It becomes significant from the table (populated from the ANOVA results) that after the iteration where cultural value stars are added, further models don't showcase much difference in terms of decrease (means improvement in the model) in Residual Deviance. So we restricted our final model with the top 5 variables which makes most significant difference to the model i.e. company + overall\_ratings +Year + senior\_mgmt\_stars + work\_balance\_stars +culture\_values\_stars.

Thus, ANOVA test helped in comparing various models possible with the most significant to least significant variables.

*Table 1*

|  |  |
| --- | --- |
| Variables in Model | Model Residual Deviance |
| None | 55265 |
| Company | 54827 |
| Company + overall rating | 53512 |
| Company + overall rating + Year | 53342 |
| Company + overall rating + Year + Senior management stars | 53210 |
| Company+ overall rating+ Year + Senior management stars + Work balance stars | 53169 |
| Company+ overall rating+ Year + Senior management stars + Work balance stars + cultural value stars | 53143 |
| Company+ overall rating+ Year + Senior management stars + Work balance stars + cultural value stars + Month | 53133 |
| Company+ overall rating+ Year + Senior management stars + Work balance stars + cultural value stars + Month + comp benfit stars | 53085 |
| Company+ overall rating+ Year + Senior management stars + Work balance stars + cultural value stars + Month + comp benfit stars + carrer opp stars | 53025 |

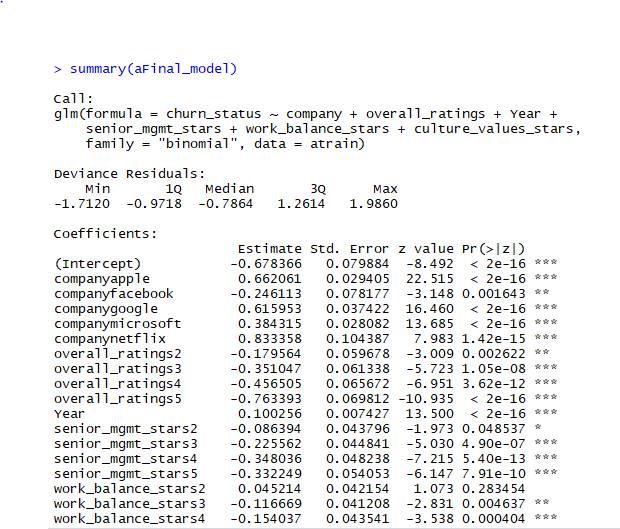
*Figure 13*

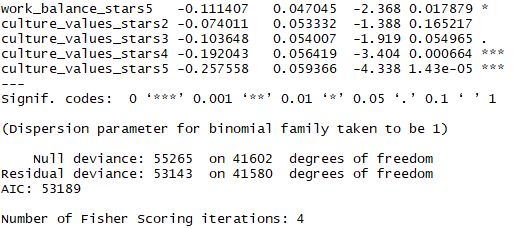


## Final Logistic regression model: Figure 14

A final model is developed with variables (company + overall\_ratings +Year + senior\_mgmt\_stars + work\_balance\_stars +culture\_values\_stars.) selection after the ANOVA test. The model has top variables only with the most significance and almost stabilized Residual Deviance of 53143.

*Figure 14*

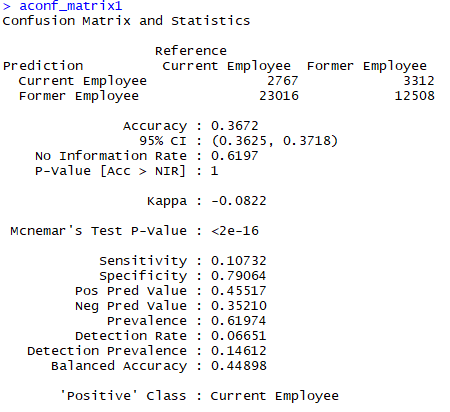




## Probability threshold tuning over training data set

With the default threshold set to .5 i.e. probability of greater than .5 will be identified as Current Employee and less than .5 will be identified as Former Employee we have an accuracy of 36% when checked on training dataset (having approximately 41k observations). Sensitivity is just .10 stating model is low or not good in predicting the correct current employees over the actual number of current employees. While Specificity is high i.e. .79 indicating model is highly capable of predicting Former employees correctly over the actual former employees. *Figure 15*

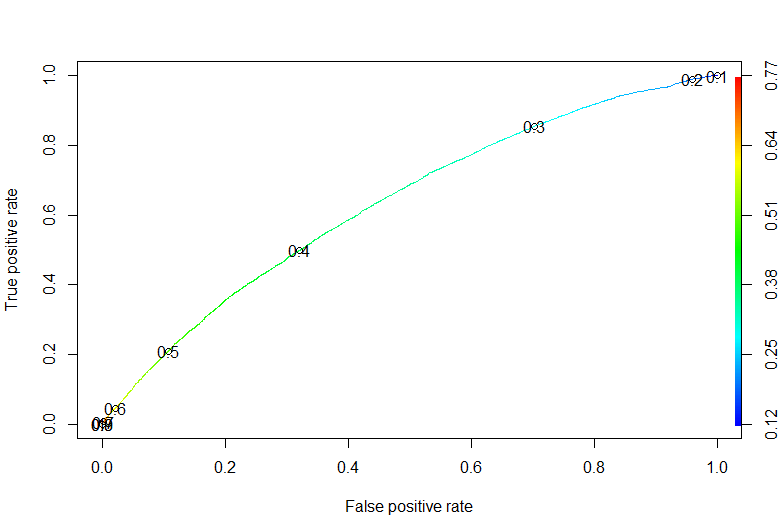
*Figure 15*



Plotting a curve between True positive rate (Sensitivity) against False positive rate we have picked

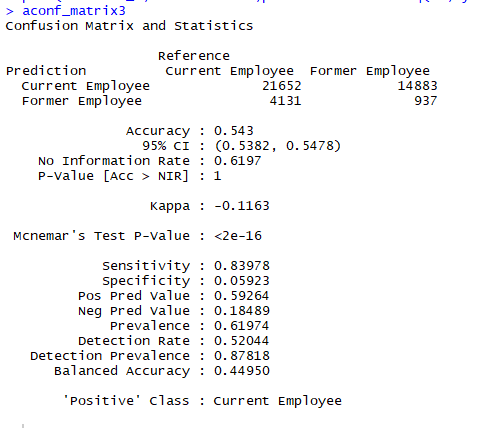
a value of .25 to retune the for the betterment of the results of the identifying True Positive i.e. Current Employees. *Figure 16*

*Figure 16*



With a probability threshold of .25 on training data we can see that Sensitivity increased and changed to .83 from .10, while specificity decreased to .05. This implies that our model can now predict Positive true values over the actual true values with much perfection than compared to Specificity which degraded to mere .05 implying poor prediction of Former employee. Accuracy of the model which is a =measure of how the prediction of true positives and true negatives has also increased from 36% to 54%. *Figure 17*

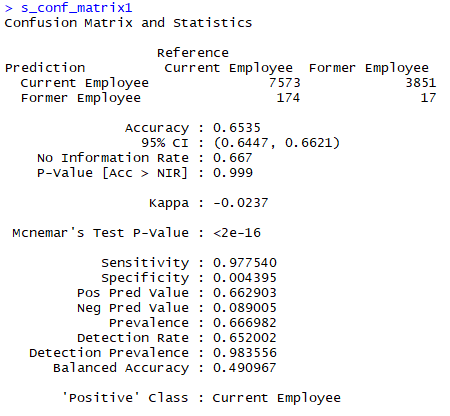
*Figure 17*



## Fitting final model over test dataset.

We now fit the developed and tuned model over the test dataset which is having approximately 11k observations. We see that the accuracy over the test dataset further increased to 65% that was around 54% over the training dataset. The sensitivity that is predicting the current employees increased to .97 implying we are capturing almost all true positives while the specificity went to a low of .004. This implies that we are not properly capturing former employees correctly that is truly negative. Overall, there is a tradeoff between the true positives and the true negatives. Based on the business requirement we can tune the threshold parameter. *Figure 18*

*Figure 18*



## Part D: Random Forest Model

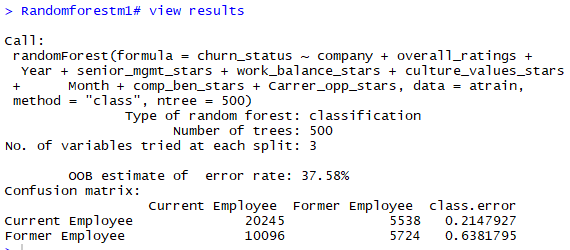
On this part of the paper, a Random forest model is considered to provide a prediction on the dataset and if the model can classify to a certain degree with accuracy those that are a current and former employee. The independent variables on the previous models are used and the data segmentation in Part B of the paper.

## RF (Random Forest) Model: Churn Status with the 9 Independent variables.

The first RF Model developed takes into consideration the churn status and the 9 independent variables on the dataset. The N Tree is set to 500 and the number of variables tried at each split is 3. The out of the bag error rate of this model is calculated at 37.58%. This represents the fraction of the number of incorrect classifications over a number of out of bag samples. From this, the accuracy is computed as 62.42%. *(Figure 19)*

From the RF Model 1 confusion matrix, we have a high error rate in classifying former employee and a lower error rate in classifying our positive outcome which is current employees.

*Figure 19*



## RF parameter fine-tuning:

This section explores further on how we can fine tune our Random Forest model by ascertaining the optimal forest size, a number of splits and the paring down of our attributes to produce a parsimonious model. This will identify the optimal values for these parameters which enables our model to provide the most accurate outcome given the variables on the data set.

## Optimal Forest Size

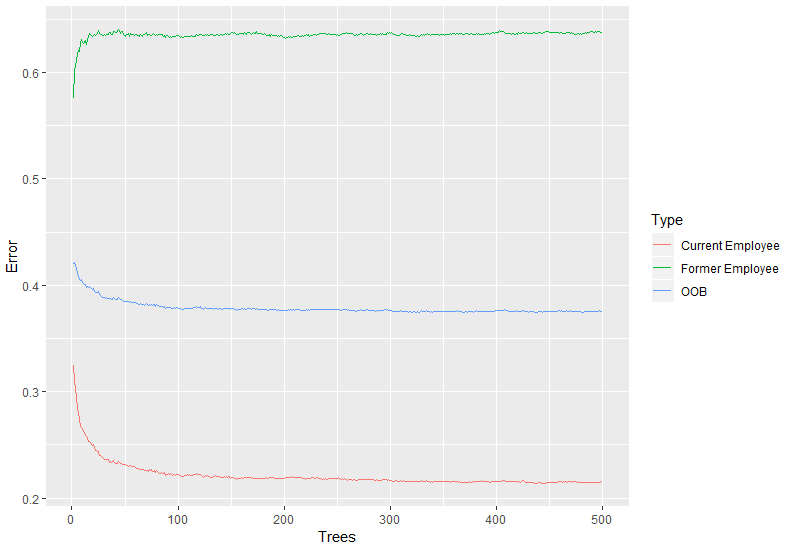
Finding the optimal forest size helps in identifying the number of trees to incorporate in the model to provide efficient computation. This will balance the classification performance with the time to produce the model itself.

The figure below shows the misclassification rate of the training data in the RF Model 1 developed above which is the blue line for our OOB, the current employee in the red line, the former employee the green line. This shows that our positive outcomes have a better classification and that our overall misclassification rate is around 37.58%. *(Figure 20)*

The former employee error rate shows that with a larger number of trees that its strength is just before 50 trees before it plateaus. The current employee rate shows that it is much higher at around 150 trees before it plateaus. In the same way our OOB rate. This shows that the difference in stability after 150 trees is very small.

Therefore, we can say that the optimal forest size for our Random Forest model is 150 trees. This will be incorporated into the final model.

*Figure 20*

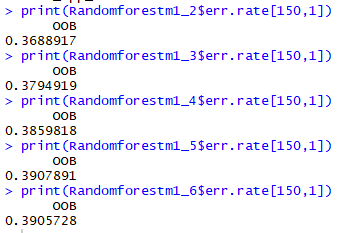


## Number of Splits

The figure below shows Random Forest models OOB error rate with varying number of splits with values of 2, 3, 4, 5, 6. This shows that the more split we add our OOB error rate increased and thereby decrease our accuracy for our Random Forest model that we are building. The ideal number of splits should show the least error rate in order to optimize the accuracy of the model that we are building. The number of splits will dictate how a tree can recover from a bad split as well as if a branch splits on noise, which will allow for termination of a bad branch early on. That is why we are selecting the number of splits with the least error rate. *(Figure 21)*

Therefore, based on the OOB error rate, the number of splits to be considered in the final model is 2.

*Figure 21*



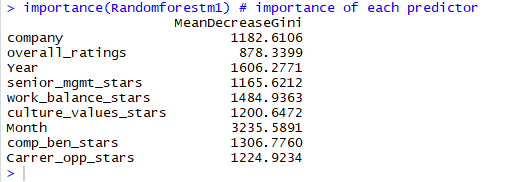
## Number of features

As random forests is an aggregation of decision trees, a node represents conditions on a single feature. In this study, these nodes are a total of 9 variables. It is designed to split the dataset into two splits so that we have the same values in the same set. This is measured in terms of the optimal condition is the Gini. This value is represented by the mean decreased gini shown below. It shows these values for each of the variables. *(Figure 22)*

In the feature selection, the mean decrease gini is used in order to ascertain which variables are more significant than others. The mean decreased gini means that the accuracy of the random forest decreases with the exclusion of a variable. This also means that the exclusion is made on more important variables which decrease the accuracy. Thus, the mean decrease is a measure of how the variable contributes to the homogeneity of the splits.

The higher mean decrease gini suggests that the variable is more significant than others. Based on the below table, the variables used are pared down into 5 and will be incorporated into the final model. The variables chosen are; Year, work balance stars, month, comp ben stars, career opp stars.

*Figure 22*



## Random Forest Final Model:

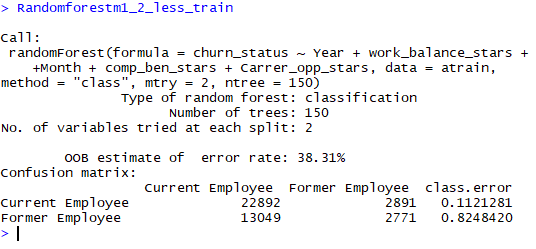
This part of the papers takes into consideration the fine-tuning parameters explored in the preceding parts.

The Random Forest Final model is built on the basis of feature selection, we classify churn status based on the 5 selected variables (Year, work balance stars, month, comp ben stars, career opp stars) with 150 trees and with the number of variables tried at each split at 2.

## RF Final Mode: Training data (the year 2017 and below)

The model below shows an error rate of 38.31% is built on the training data. It shows a higher misclassification rate with the former employee and a lower misclassification rate with the current employee. This means that our classification of current employees in the training data (2012-2017) is at 61.69% accurate. *(Figure 23)*

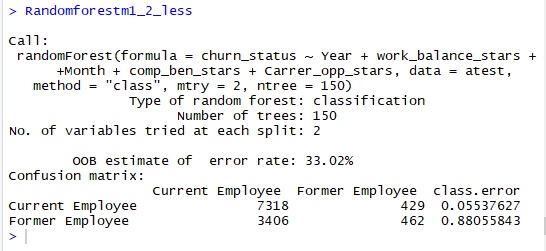
*(Figure 23)*



## RF Final Model: Test data (2018)

The model below shows an error rate of 33.02% which is built on the test data. It shows a higher misclassification rate with the former employee and a lower misclassification rate with the current employee. This means that our classification of current employees in the test data (2018) is at 66.98% accurate. *(Figure 24)*

*(Figure 24)*



Overall, there was a slight improvement with the test data set in terms of overall accuracy. The error rate on former employees increased by .06 while the error rate for current employees decreased by .06. In this case, our model for the test is accurate by more than 5.29% than our training data.

## Part E: Comparing two models

The table below shows a summary of the metrics selected to evaluate the two models that were developed.

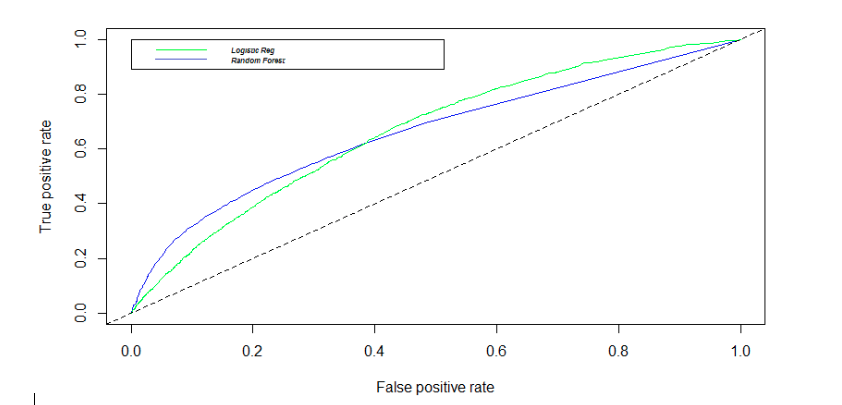
The Random forest has a small difference in how the model worked on the training and test data based on two metrics: accuracy and sensitivity. This implies that Random forest is the more stable of the two models in classifying current and former employees. The random forest also performs better on error rate with a small variation than Logistic Regression. Random Forest also was able to classify negative outcomes better than logistic regression, though both performed with the same difference form training and test data. *(Table 2)*

*Table 2*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Logistic Regression | | Random Forest | |
| Metrics/ Data | Training Data | Test Data | Training Data | Test Data |
| Error Rate | 0.460 | 0.350 | 0.390 | 0.340 |
| Accuracy | 0.543 | 0.653 | 0.610 | 0.660 |
| Sensitivity | 0.840 | 0.977 | 0.880 | 0.940 |
| Specificity | 0.059 | 0.004 | 0.170 | 0.120 |

The AUC for Logistic Regression shows that higher true positive rate and the area under the curve is .665. For the AUC for Random Forest the area under the curve is .661. The AUC was calculated to the default of .50 threshold for logistic regression which shows it has a better sensitivity of predicting true positive rate. Though when we chose the final model as per the tuning parameters we forego a higher sensitivity rate for higher specificity and overall accuracy as per the tuning done on logistic regression and random forest.As per the ROC Plot below we can see that both models are comparable to each other and the logistic regression is slightly better in predicting true positive than random forest.*(Figure 25)*

*(Figure 25)*



For the variable selection, we have the following for logistic regression, company, overall\_ratings, Year, senior\_mgmt\_stars, work\_balance\_stars, and culture\_values\_stars. Meanwhile, we have, (Year, work balance stars, month, comp ben stars, career opp stars). The two models overlap in using; year, work balance as both highly predictive in classifying current and former employees.

## Part F: CONCLUSION

In comparing the two models we found that the Random Forest Model performs better to answer the paper objective of predicting the churn status of employees from the six companies included in the dataset. In terms of classifying whether the employee is with the company or not we get a lower misclassification rate and a stable prediction with the random forest model. By applying the finalized logistic regression and random forest model we found that random forest has a lower difference in predicting current from former employees between the training and test data set. We look at the difference in terms of accuracy, the logistic regression model has a jump of .11 from the training to the test data, while, Random forest jump .05 from the training to the test data. Data segregation is based on the year that the survey was submitted. We can say that even with the different years taken into consideration our Random forest model performs better than logistic regression.

As already stated above, the tuning parameters for random forest is easily navigable and can be shows itself as a better classification model than logistic regression. For our objective in this paper we are able to achieve an accuracy high of 66% when identifying employees that will stay with the company rather than leave the company. Comparing the two models we see that there are two variables that are significant in the two model namely; year and work life balance. We can say that years would capture any trend that is prevailing during that time, ie, economic slowdown. The work life balance variable becomes a very significant predictor if an employee will stay in the company or would rather look for other opportunities.

## Part G: RECOMMENDATION

We have touched the most important variables on the data set regarding employee status for the initial phase. For the next phase we want to further drill down into the pros and cons column, scrape negative and positive words that can predict whether the employee stay or not. We can do this by taking in inventory of the positive and negative words and create new variables for these word counts which can further contribute to the response variable to provide better prediction.

## Part H: PROJECT FEEDBACK FROM Q & A

1. Which performance metric is the most important for your analysis specificity or sensitivity?

Developing the models, we were able to see that the models that we were more focused on predicting the positive outcome which is sensitivity, which represents the current employees. After the feedback from the class presentation we realized that the prediction for the project should focus on specificity which is the negative outcome, which is former employees. Since our project goal is to predict employee churn.

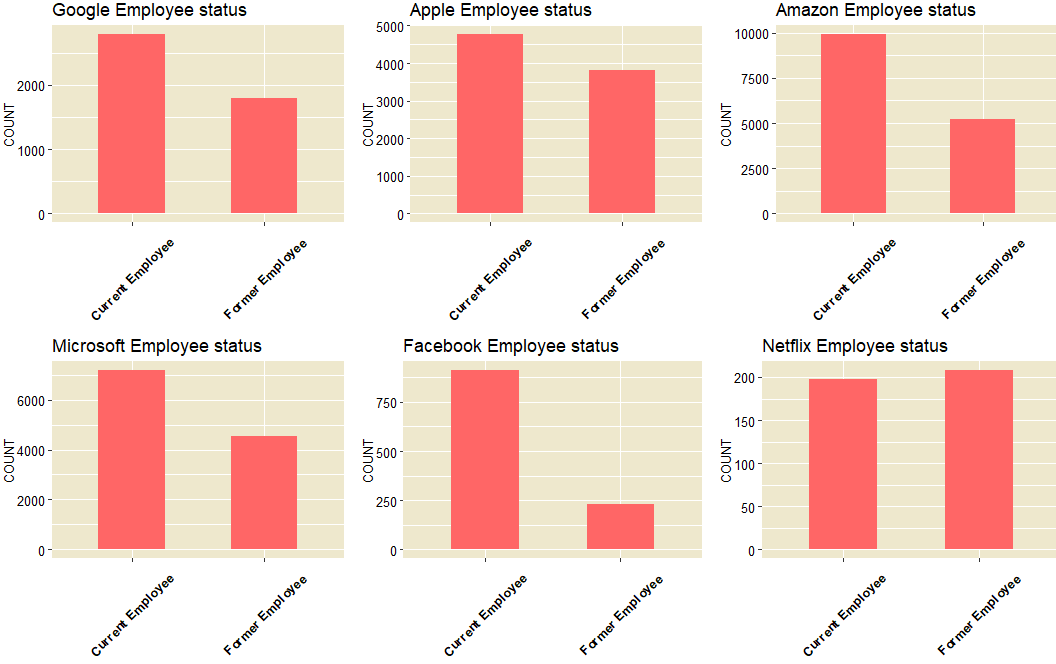
As per our developed model we see that we have a very low specificity rate which could be the consequence of class imbalance. This should be taken cared for developing models and is a good learning for us.

1. Please address the class imbalance and consider different sampling options (oversampling, under sampling, etc.) as a future improvement.

As discussed above, the reason why our specificity is low due to the class imbalance. As presented in the class we can develop models that will have the different sampling options such as under-sampling, over-sampling, both, or SMOTE. Then we can choose among the developed models how they are going to perform given the sampling options. This will allow us to address the class imbalance model and develop more robust models rather than having a ‘lazy’ classification model.

1. Consider including historical churn count distribution per company to support why the company variable is an important variable in your model.

We went back to the dataset of the training data and found the reason for our higher sensitivity i.e. why our model was predicting true positives and weak at specificity. It is found that Amazon and Microsoft, which are having dominating numbers are having imbalance in the class. This is a leaning for us which we obtained from professor feedback. Training data requires a balance in the class by utilizing the sampling methods. This issue will be taken care in the future models.



1. In your report, please update your bar charts by comparing the percentages rather than raw counts.

All charts under EDA are in percentages.