***Human Resources Analytics Department***

***Shum Chang Wu Chen***

***CIS 3920 Semester Project***

# 1. ****Proposal****

**Dataset**: Human Resources Data Set (Kaggle)

**URL**: https://www.kaggle.com/datasets/rhuebner/human-resources-dataset **Description**:

This data set contains 312 observations and thirty-six variables. This dataset “revolves around a fictitious company and the core data set contains names, DOBs, age, gender, marital status, date of hire, reasons for termination, department, whether they are active or terminated, position title, pay rate, manager name, and performance score” (Kaggle).

**Project Goals:**

My intent with this project is to build a prediction model to predict whether an employee would stay or leave this fictitious company. The label that represents this in the data set is “Termd”, it’s a binary variable between 0 and 1, where 0 represents someone still active, and 1 representing someone who has left for various reasons. I

I will begin to do this by building a logistic regression model. I believe this will be a good starting point as well as help me understand what variables are significant. I will then plan to move to random forest to further my understanding of the variables presented, and hopefully gradient boosting. I have read that gradient boosting is best for tabular data that are structured, that is what this dataset has currently. I also don’t have much background knowledge to gradient boosting, but this is a good opportunity to learn. As for the features used in the prediction models, I will first begin with all the variables, and a subset and see whether the accuracy increases with the adjustment.

*These are some questions to help guide my project.*

*• Can I predict an employee's termination?*

*• Can I measure the attrition of this fictitious company?*

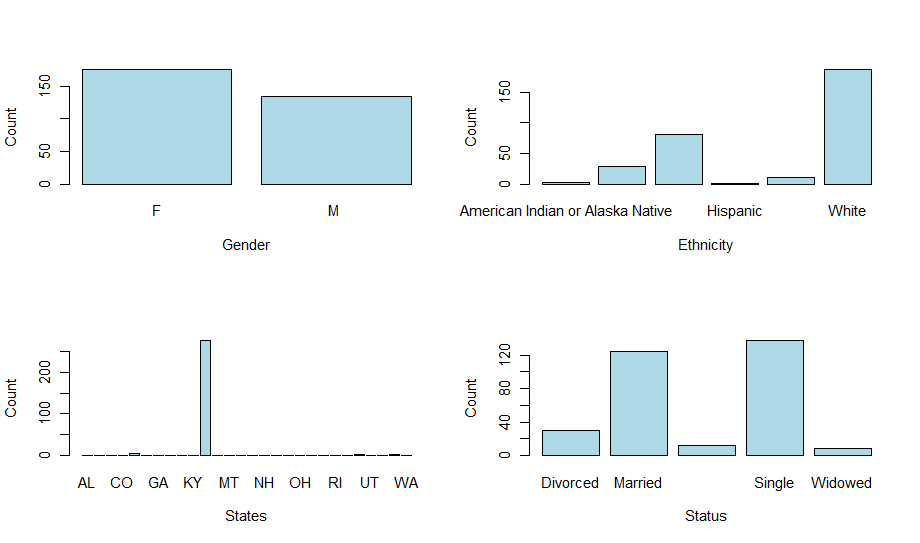
*• What factors contribute to an employee’s decision to leave?*

*• Does the manager affect the employee’s performance and happiness?*

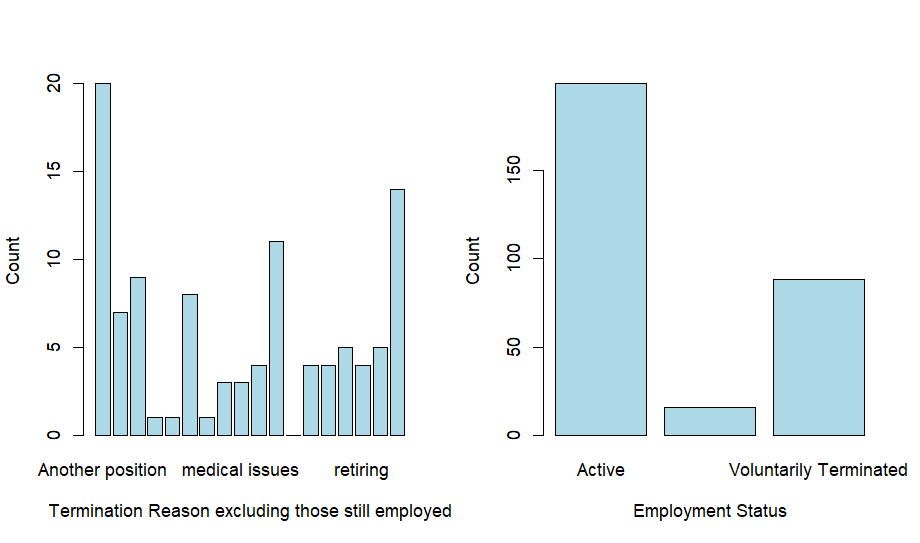
# 2. ****Data Cleaning****

Because this data is pretty cleaned, I did not have to complete much on my end. I first searched for duplicates, null, and NA values. I found a total of 8 NA’s, I decided to remove the values because it was less than 5% percent of the total data, this would have minimal effect in the study. I then normalized the integer base values like performance scores and salary to the min and max. Finally, I noticed that most of my categorical variables were read as integers. This would cause an issue, so I fixed this by converting the data type from int to factor, this done majority of the variables that were listed as “ID”.

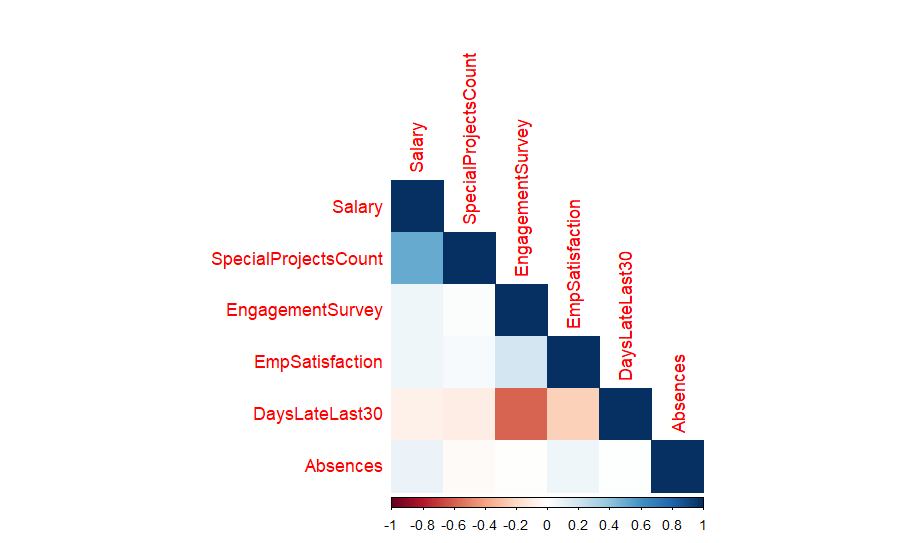
# 3. Exploratory Data Analysis



This graph provides a general understanding of our demographic. Majority are female and white, and there is equal distribution between married and single. Lastly most of our employees are from Massachusetts. My hypothesis is that these factors will not affect whether an employee will desert their position.



This graph provides some answers to employee’s job status, if they have left the company, the graph on the left provides some explanations. This includes leaving for another position, retiring, military and medication conditions, these are just a few to list. We can see from the graph on the right that there is a class distribution in balance. There are more active employees than inactive employees, which may lead to a bias in our models.



Lastly, this is a heat map of all the left-over integer variables. What is interesting here is that there is a strong negative correlation between “DaysLateLast30” and “EngagementSurvey”. This does not come as striking because you would assume someone is less engaged if they are often late. This has a higher weight in my modeling later. Another interesting point is the moderate positive correlation between “Salary” and “SpecialProjectCount”.

# 4. Modeling

4.1 Logistic Regression Model

Features: SpecialProjectsCount + Salary + EmpSatisfaction + EngagementSurvey + PerformanceScore + DaysLateLast30

A screenshot of a computer

Description automatically generatedLabel: Termd

This was the first logistic regression model I trained; the features were selected based on intuition on the basis on what makes sense realistically. This model and the following models are all built with the caret package, allowing models to be build extremely quickly with minimal code, it also easily allowed cross validation with 10 folds and 3 tries. The performance of this model is 65.57% which isn’t ideal but it’s a good start. What is interesting here is that the confusion matrix is that it predict 37 at TP and only 3 as TN. The rest of 18 at FN and 3 as FP. My assumption is that this is because of class imbalance that I noted earlier. This may cause the model to be more A screen shot of a computer

Description automatically generatedbiased to predicting 0 which represents the employers that are active.

Another assumption that I have is that the variables might play a role in the performance, as these may not this the best models. After pulling a feature importance scale using varIMP(), the most significant five were SpecialProjectCount, DayLateLast30, PerformanceScoreFullMeet, EmpSatisfactions, and PerformanceScorePIP.

4.2 Random Forest

Features: SpecialProjectsCount + Salary + EmpSatisfaction + EngagementSurvey + PerformanceScore + DaysLateLast30

A graph with a line

Description automatically generatedA screenshot of a computer screen

Description automatically generatedLabel: Termd

This model was built using the same features to maintain consistency. This one has 5-fold cross validation and 5 mtry. The accuracy of this model did not change from the last. The matrix has 39 TP 1 FP 20 FN 1 TN, this is a slight change from the logistic regression model. Moving onwards, the accuracy within this model decreases with each iteration of mtry. It starts off with 65.57 % accuracy, with each number of randomly selected predictors, the accuracy decreases to around 51% as seen above.

A diagram of a survey

Description automatically generated with medium confidenceA screen shot of a computer program

Description automatically generated

From the varImp() function the most important variable from the random forest is Salary, which is interesting, because this correlates to plot explaining reasons why employees left the company. “More money” was the 3rd most frequent response from the employees. The tree model on the was a decision tree from the random forest, where it had specialprojectcount as the first node, then salary. What is interesting here is that there again is a class imbalance located within the decision tree. There are more success variables (our 0’s) than our failure variable (1’s).

4.3 Gradient Boosting

Features: SpecialProjectsCount + Salary + EmpSatisfaction + EngagementSurvey + PerformanceScore + DaysLateLast30

A screenshot of a computer

Description automatically generatedA screen shot of a computer

Description automatically generatedLabel: Termd

Using the same parameters as before, our accuracy slightly drops to 62.3 %. Also we still see the class imbalance wthin the prediction and true values. Within this model, we see again that the salary is the most important variable, which again is relevant in everyday settings. Human beings tends to move from role to role for the pay increase and different job respsonsbilities.

# 5. Conclusion

To conclude, I don’t think these models are in a usable state as of now. Although I have done cross validation techniques to obtain more consistent results, the main problem with my prediction models is that it predicts too many false positive/negative values. My assumptions to why this occurred may be the class imbalance within the variable “TermD”, where there are 2 times more success variables than failure variables. Another assumption is that there is overfitting from the models, however random forest is one of the models that reduces it. However, in my use case, random forest and logistic regression both provide very similar confusion matrix and accuracies. Beyond these two assumptions, my final guess is that the predictors I am choosing are the ones that are causing the problem with the prediction model.

Although this model wasn’t successful, it can still be used to gain insight. We see from the 3 machine learning models that there are certain variables that have higher weights in determining the employment status of an employee, salary, special projects, and engagement. These 3 are the variables are likely to influence an employee’s decision-making process with their employment. As with real life, an ideal job is one that pays well and engages your mind and body well. These models support that statement.