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.

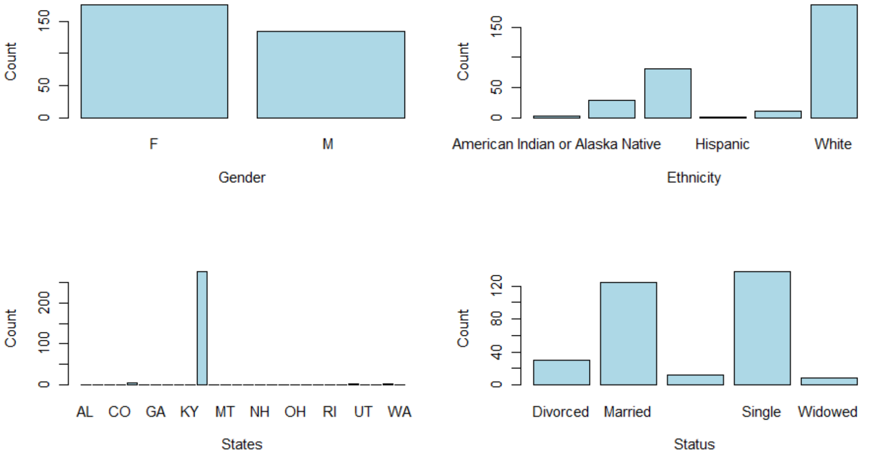
These are some questions to help guide my project.

- *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?*

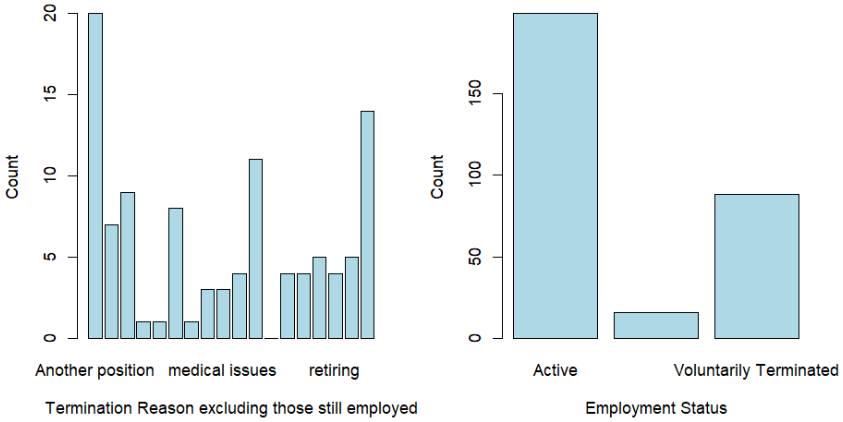
2. Data Cleaning

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".

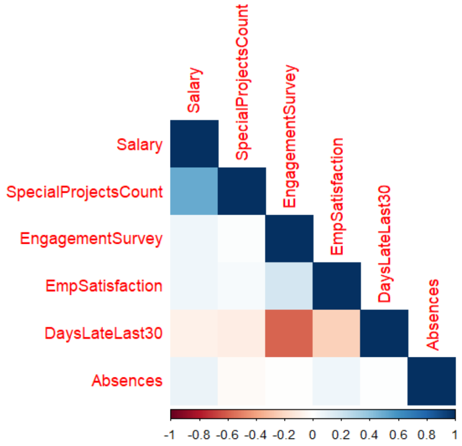
3. Exploratory Data Analysis



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.

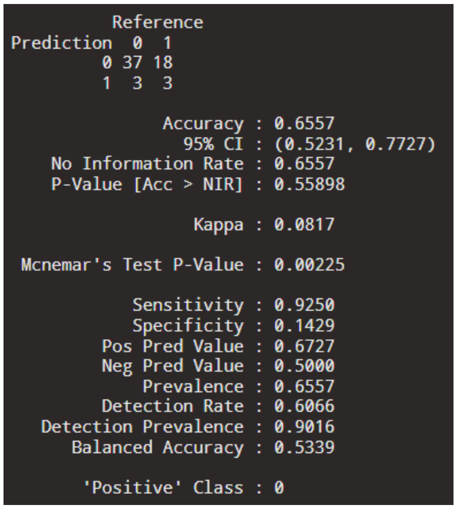


Lastly, this is a heat map of all the left-over integer variables. What is interesting here is that there is a strong negative correlation between "DaysLateLast30" and "EngagementSurvey". This does not come as striking because you would assume someone is less engaged if they are often late. This has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

4. Modeling

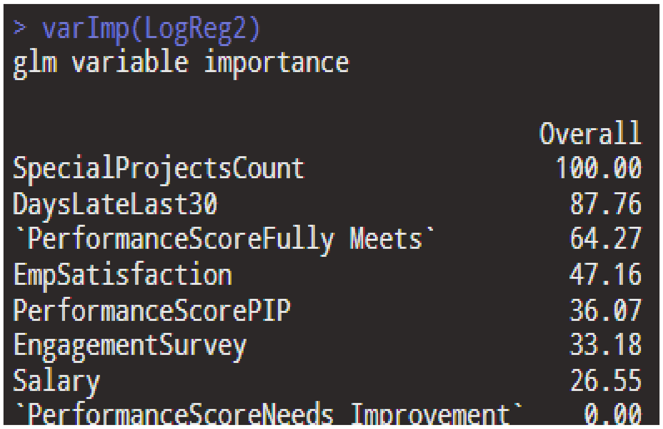
4.1 Logistic Regression Model

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



```
          Reference
Prediction  0  1
         0 37 18
         1  3  3

               Accuracy : 0.6557
                 95% CI : (0.5231, 0.7727)
    No Information Rate : 0.6557
    P-Value [Acc > NIR] : 0.55898

                  Kappa : 0.0817

 Mcnemar's Test P-Value : 0.00225

            Sensitivity : 0.9250
            Specificity : 0.1429
         Pos Pred Value : 0.6727
         Neg Pred Value : 0.5000
             Prevalence : 0.6557
         Detection Rate : 0.6066
   Detection Prevalence : 0.9016
      Balanced Accuracy : 0.5339

       'Positive' Class : 0
```

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 biased to predicting 0 which represents the employers that are active.



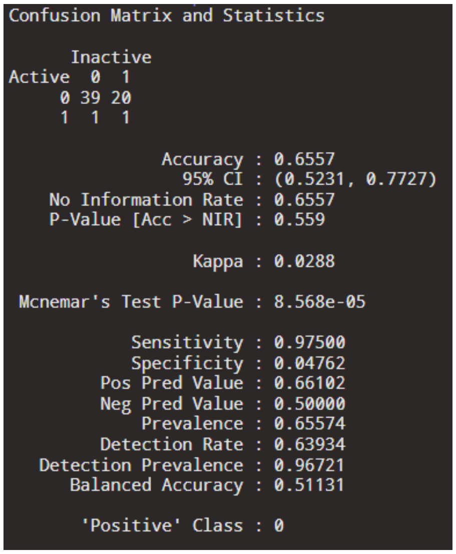
```
> varImp(LogReg2)
glm variable importance

                                    Overall
SpecialProjectsCount                 100.00
DaysLateLast30                        87.76
`PerformanceScoreFully Meets`         64.27
EmpSatisfaction                       47.16
PerformanceScorePIP                   36.07
EngagementSurvey                      33.18
Salary                                26.55
`PerformanceScoreNeeds Improvement`    0.00
```

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.

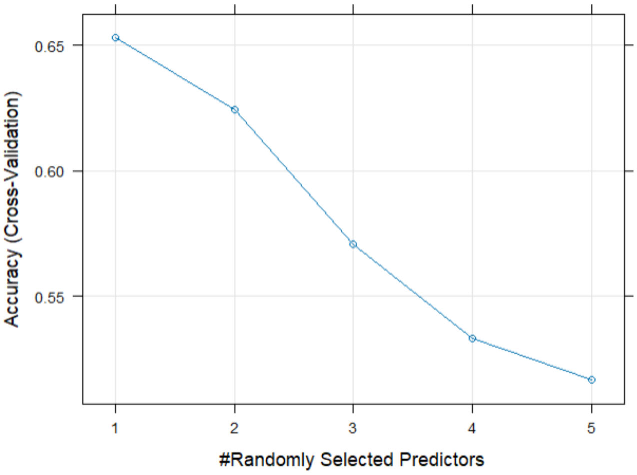
4.2 Random Forest

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

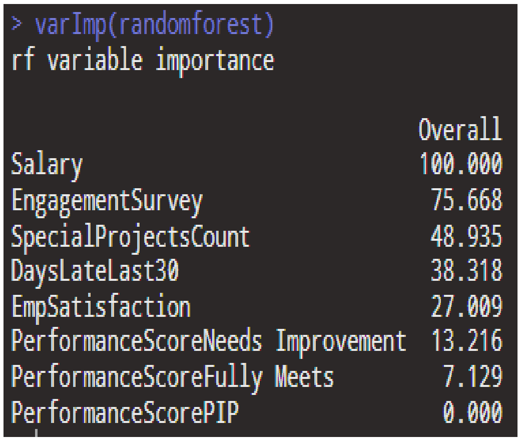
Label: Termd



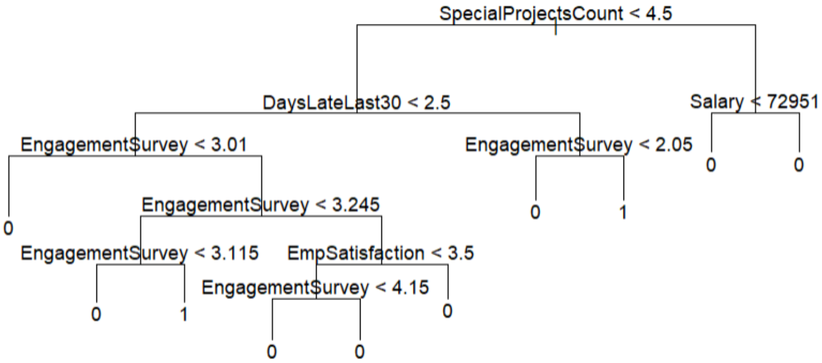
```
Confusion Matrix and Statistics

      Inactive
Active  0  1
     0 39 20
     1  1  1

               Accuracy : 0.6557
                 95% CI : (0.5231, 0.7727)
    No Information Rate : 0.6557
    P-Value [Acc > NIR] : 0.559

                  Kappa : 0.0288

 Mcnemar's Test P-Value : 8.568e-05

            Sensitivity : 0.97500
            Specificity : 0.04762
         Pos Pred Value : 0.66102
         Neg Pred Value : 0.50000
             Prevalence : 0.65574
         Detection Rate : 0.63934
   Detection Prevalence : 0.96721
      Balanced Accuracy : 0.51131

       'Positive' Class : 0
```



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.



```
> varImp(randomforest)
rf variable importance

                              Overall
Salary                        100.000
EngagementSurvey               75.668
SpecialProjectsCount           48.935
DaysLateLast30                 38.318
EmpSatisfaction                27.009
PerformanceScoreNeeds Improvement  13.216
PerformanceScoreFully Meets     7.129
PerformanceScorePIP             0.000
```

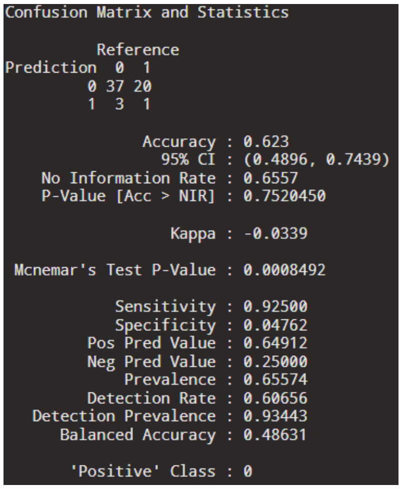


From the varImp() function the most important variable from the random forest is Salary, which is interesting, because this correlates to plot explaining reasons why employees left the company. “More money” was the 3rd most frequent response from the employees. The tree model on the was a decision tree 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).

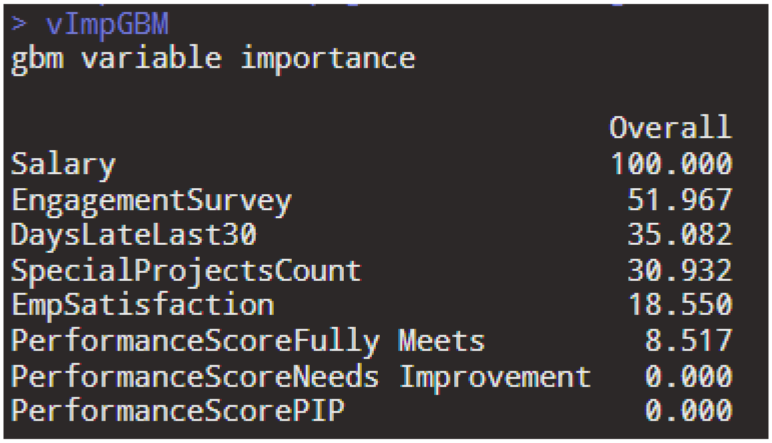
4.3 Gradient Boosting

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

Label: Termd



```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 37 20
         1  3  1

               Accuracy : 0.623
                 95% CI : (0.4896, 0.7439)
    No Information Rate : 0.6557
    P-Value [Acc > NIR] : 0.7520450

                  Kappa : -0.0339

 Mcnemar's Test P-Value : 0.0008492

            Sensitivity : 0.92500
            Specificity : 0.04762
         Pos Pred Value : 0.64912
         Neg Pred Value : 0.25000
             Prevalence : 0.65574
         Detection Rate : 0.60656
   Detection Prevalence : 0.93443
      Balanced Accuracy : 0.48631

       'Positive' Class : 0
```



```
> vImpGBM
gbm variable importance

                                  Overall
Salary                            100.000
EngagementSurvey                   51.967
DaysLateLast30                     35.082
SpecialProjectsCount               30.932
EmpSatisfaction                    18.550
PerformanceScoreFully Meets         8.517
PerformanceScoreNeeds Improvement   0.000
PerformanceScorePIP                 0.000
```

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.