Executive Summary

Employee Attrition Statistical Analysis
John Verdoza
Western Governors University
Dr. Emelda Ntinglet
February 18, 2021

1 Statement of Problem and Hypothesis

Businesses are implementing data analytics to drive efficiency and decrease their costs. One area that data analytics is being applied to is in human resources, specifically in detecting what causes employees to leave, known as attrition or churn. Previous studies by Barrick and Zimmerman (2009) detected that employee satisfaction and job performance were good indicators of employee churn. Another study by Yigit and Shourabizadeh (2017), built machine learning models using employee data to predict which employees were likely to churn. The purpose of this analysis is to build off of these foundations by constructing a logistic regression model that will predict employee churn. The hypothesis is stated as there are statistically significant factors that have an effect on employee churn.

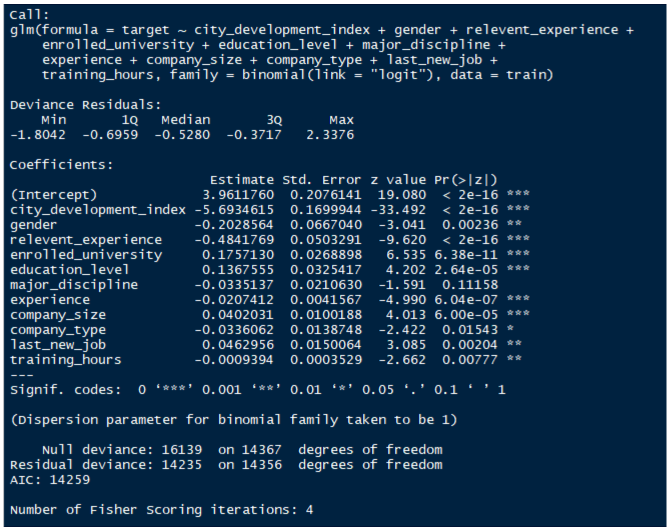
2 Summary of Data Analysis

The data set was gathered from Kaggle and consists of information gathered on candidates using a data analysis training module. Gathered information includes the candidate's education, gender and work experience. I used the statistical programming language R to handle all of the data processing and analysis. There was one variable removed, Enrollee ID, which is a unique identification number. All categorical variables were converted to numerical values in order to be used in the logistic regression model. The data was also cleaned of all missing variables by using a process called Multiple Imputation by Chained Equations, MICE.

Once the data has been processed, I moved onto analyzing the data. First, I split the data into two sets, a training set consisting of 75% of the data and the remaining 25% for testing. Next, I fit the data to a logistic regression model using all of the variables. In order to fine tune the model, I will be using stepwise selection in order to reduce features and improve the fit of the model. I will be using the Akaike Information Criterion, AIC, in order to determine which model fits the data best. As the stepwise selection method removes and adds variables, an AIC value for each model is generated. The model with the smallest value fits the data best as stated by Tuffery (2011).

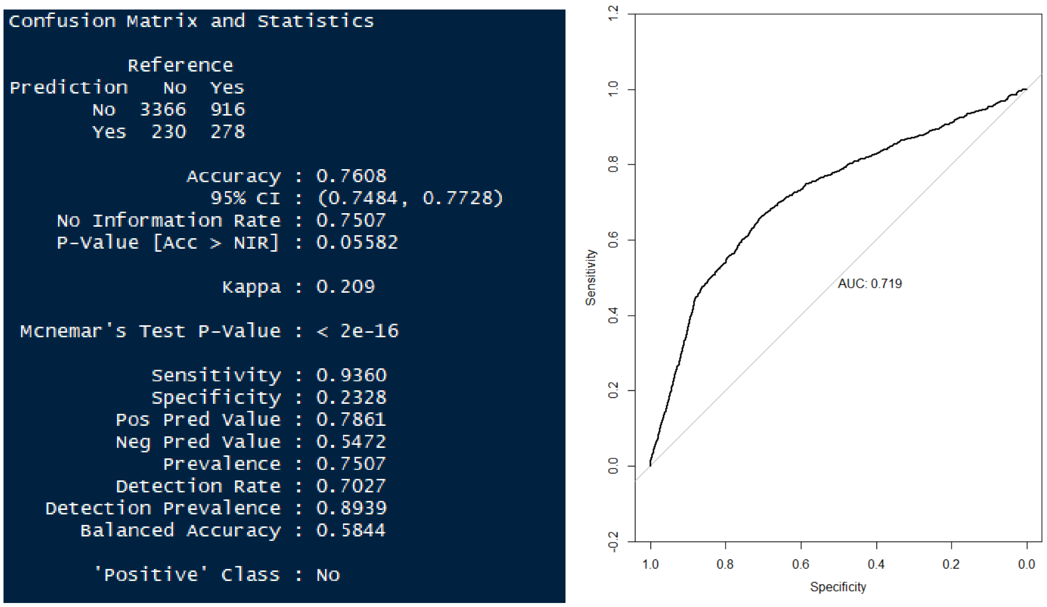
3 Outline of Findings

The final model contains the features, City Development Index, Gender, Relevant Experience, Enrolled University, Education Level, Major Discipline, Experience, Company Size, Company Type, Last New Job and Training Hours. All of these features are statistically significant except Major Discipline, which will be kept in the final model since the AIC value increased when removed. The output for the final model is seen in Figure 1. Our final model is then fitted with the testing data set. The purpose of fitting the data to the final model is to validate how well the model performs on unseen data. In the below figures we see that the model has a .7608 accuracy, a sensitivity of .936 which is the probability of predicting a true positive and a specificity of .2328 which is the probability of predicting a true negative. Both of these values can be combined to create the Receiving Operator Characteristic curve which shows the sensitivity as a function of the specificity according to Tuffery (2011). The metric that will determine how well the model performs is the area under the curve, AUC. As written by Tuffery (2011), the AUC tells how well the model discriminates between true values and false values. As the AUC approaches 1, the model is discriminating perfectly, while an AUC approaching 0 is declining in performance. In Figure 2 below, the model has an AUC of .719 which is pretty decent at determining the true positives and negatives.



```
Call:
glm(formula = target ~ city_development_index + gender + relevent_experience +
    enrolled_university + education_level + major_discipline +
    experience + company_size + company_type + last_new_job +
    training_hours, family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8042  -0.6959  -0.5280  -0.3717   2.3376

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             3.9611760  0.2076141  19.080  < 2e-16 ***
city_development_index -5.6934615  0.1699944 -33.492  < 2e-16 ***
gender                 -0.2028564  0.0667040  -3.041  0.00236 **
relevent_experience    -0.4841769  0.0503291  -9.620  < 2e-16 ***
enrolled_university     0.1757130  0.0268898   6.535 6.38e-11 ***
education_level         0.1367555  0.0325417   4.202 2.64e-05 ***
major_discipline       -0.0335137  0.0210630  -1.591  0.11158
experience             -0.0207412  0.0041567  -4.990 6.04e-07 ***
company_size            0.0402031  0.0100188   4.013 6.00e-05 ***
company_type           -0.0336062  0.0138748  -2.422  0.01543 *
last_new_job            0.0462956  0.0150064   3.085  0.00204 **
training_hours         -0.0009394  0.0003529  -2.662  0.00777 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 16139  on 14367  degrees of freedom
Residual deviance: 14235  on 14356  degrees of freedom
AIC: 14259

Number of Fisher Scoring iterations: 4
```

Figure 1: Logistic Regression Outputs



```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  3366  916
       Yes  230  278

               Accuracy : 0.7608
                 95% CI : (0.7484, 0.7728)
    No Information Rate : 0.7507
    P-Value [Acc > NIR] : 0.05582

                  Kappa : 0.209

 Mcnemar's Test P-Value : < 2e-16

            Sensitivity : 0.9360
            Specificity : 0.2328
         Pos Pred Value : 0.7861
         Neg Pred Value : 0.5472
             Prevalence : 0.7507
         Detection Rate : 0.7027
   Detection Prevalence : 0.8939
      Balanced Accuracy : 0.5844

       'Positive' Class : No
```

(a) Confusion Matrix for Final Model

(b) Recieving Operator Characteristic Curve

Figure 2: Confusion Matrix and ROC Curve

4 Limitations of Techniques and Tools

There are a few limitations encountered in the analysis. One such limitation was the computation speed of using the MICE process to impute missing values. While the process did take several minutes to run, scaling this process up to hundreds of thousands of data points could be very time consuming. This can translate to hours lost of productivity as this process runs. Another limitation that I encountered was the model itself. While logistic regression can be easy to understand, there is a lot of work to prepare the data. There are more robust models to build such as using a random forest which can take categorical variables as input with no need to convert to numerical values. This translates into increased productivity and could even lead to a more accurate model.

5 Proposed Actions and Benefits

This model can be used by the data analytics company in order to determine which candidates are likely to churn. However, it is recommended that the company acquires more data as this will increase the accuracy and also introduce new features that could uncover new insights. Some features that would be helpful are the employee's satisfaction with their current employer, their salary and also what topics they completed in their training. Once these features are implemented then a new logistic regression model can be created. Future improvements on this analysis include building and testing different models such as a decision tree or a random forest as seen in the paper by Yigit and Shourabizadeh (2017). Once these models are built then I can find which models are the most accurate.

References

Barrick, M. R., & Zimmerman, R. D. (2009). Hiring for retention and performance. *Human Resource Management*, *48*(2), 183–206. doi: 10.1002/hrm.20275

Tuffery, S. (2011). *Data mining and statistics for decision making.* Wiley-Blackwell.

Yigit, I. O., & Shourabizadeh, H. (2017). An approach for predicting employee churn by using data mining. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP).* doi: 10.1109/idap.2017.8090324