Employee Attrition Statistical Analysis

John Verdoza
Western Governors University
Dr. Emelda Ntinglet
February 18, 2021

Abstract

What causes an employee to leave a company? Are they the type of employee that leaves a job after a year or two, do they feel that their experience level is not well suited for their job or does it have to do with the size of their previous company? Is it possible to predict if an employee will leave a company given information on their education, experience and current employer? The goal of this paper is to build a model to answer this question.

1 Introduction and Research Question

Many companies are starting to use data analysis to improve business processes. One such process that has seen significant improvement is in predicting if an employee will leave their current employer also known as attrition or churn. While the employee's motivations can span many different areas such as company culture, job title, money and self-worth this analysis will only focus on information such as education, experience and gender. Previous studies have tested whether knowing someone at the company (?, ?) or employee satisfaction and job performance were useful in lowering employee churn (Barrick & Zimmerman, 2009). Another study compared the accuracy, precision and recall of six different machine learning techniques to determine factors that drive employee churn according to Yigit and Shourabizadeh (2017). While these studies looked at preventing churn and also finding factors, this report aims to build off of these studies to determine how the data analytics company can attract those that churn. While doing so, I will also be able to identify which factors have a significant influence on attrition. In order to make this prediction I will need to start by gathering data, then cleaning the data and finally building a model.

In this analysis I will be using a data set obtained from Kaggle that contains information of persons utilizing a data analysis training module. The company in charge of the training module is interested in determining what factors will influence a candidate to leave their current employer and work for the data analytics company. In essence they are looking to see what factors drive churn to attract talent for their company. Not much information is known about the company or specific details about each person as the data has been anonymized. However, upon registration, information is gathered for each person and this is what makes up the data. Each candidate supplies the city that they are from, their gender, years of experience, how long they worked for their current employer and the type of company that they working for. Lastly, each candidate is assigned their unique Enrollee ID number. Using this information I will build a logistic regression model that I will use to predict the probability that a given candidate will leave their current employer. Using this model I will find out which factors have the most influence on the model. The hypothesis for this analysis is as follows:

H_0: There are no statistically significant factors at the $\alpha = .05$ level that have an effect on a candidate leaving their current company.

H_a: There is at least one statistically significant factor at the $\alpha = .05$ level that has an effect on a candidate leaving their current company.

2 Data Collection

The data that will be used in this analysis was taken from the website Kaggle and was stored in a comma separated value file also known as a csv file. The data will be read into the statistical programming language R and stored as a data frame. I have chosen to use R because of it's ability to easily read in csv files, how easy it is to convert raw data into usable information and the simplicity of creating graphs. While R does have many advantages one disadvantage is that some of the functions can take a while to run. The process of Multiple Imputation by Chained Equations, used in this analysis did take a while to run. Once imported into R, I can use functions such as structure and summary to give me an overview of the data. The data set consists of 19,158 unique observations, thirteen independent variables and one dependent variable. Of the thirteen independent variables, ten of those are categorical variables and three are numeric variables.

The variable **Target** is the dependent variable. Each of the categorical variables will be read into R as factors to preserve their values. This will also allow me to preserve any missing values in the data set. The variables are defined as below:

- Enrollee ID: A unique id number given to each candidate.
- City: A numerical representation for the city the candidate lives in.
- City Development Index: This is a value from 0 to 1 that denotes the level of development for the given city with 1 being the highest.
- Gender: The gender of the candidate.
- Relevant Experience: If the candidate has relevant experience or not.
- Enrolled University: If the candidate is currently enrolled in a university and if they are a full or part time student.
- Education Level: The highest level of education achieved.
- Major Discipline: The candidate's area of study.
- Experience: The candidate's years of experience.
- Company Size: The size of the company that the candidate works at.
- Company Type: The type of the company that the candidate works at.
- Last New Job: The length of time that the candidate has worked at their current job.
- Training Hours: The number of training hours that the candidate completed.
- Target: Whether the candidate left their company. (1 = "Yes", 0 = "No")

3 Data Extraction and Preparation

In order to properly build the logistic regression model, I will need to prepare and clean the data. I will prepare the data by examining if there are any duplicates or missing values. The output displayed in Figure 1 shows that there are no duplicated rows found in the data set and the total number of missing values for each variable. Missing values can introduce biases to the analysis and will be handled in a later section.

```
> hire[duplicated(hire)]
data frame with 0 columns and 19158 rows
> na_count
                        na_count
enrollee_id                    0
city                           0
city_development_index         0
gender                      4508
relevent_experience            0
enrolled_university          386
education_level              460
major_discipline            2813
experience                    65
company_size                5938
company_type                6140
last_new_job                 423
training_hours                 0
target                         0
```

Figure 1: Variables and Total Missing Values

I will begin by examining each of the categorical variables: **Gender**, **Enrolled University**, **Education Level**, **Major Discipline**, **Last New Job**, **Relevant Experience**, **Company Size**, **Company Type**, **Experience** and **City**. Categorical variables can take on values called levels that are descriptive labels. The first categorical variable **Gender** contains three levels which can are *Female*, *Male* and *Other* with the majority of candidates being *Male*. Gender also contains 4,508 missing values. The second categorical variable **Enrolled University** also has three levels which are *Full Time*, *No* and *Part Time* with a majority of candidates not being enrolled in a university. There are a total of 386 missing values for **Enrolled University**. The third categorical variable **Education Level** contains five levels which are *Graduate*, *High School*, *Masters*, *Phd* and *Primary School*. There are also a total of 460 missing values and the majority of candidates have a bachelors degree. The fourth variable **Major Discipline** contains six levels. Those levels are *Arts*, *Business Degree*, *Humanities*, *No Major*, *Other* and *STEM* which stands for Science, Technology, Engineering and Math. There are also 2,813 missing values and the overwhelming majority of candidates come from a STEM background. Due to the high number of candidates with a STEM major, another question can be proposed, does the STEM majority have a significant effect on the model? The graphs for each of these can be found in Figure 2.

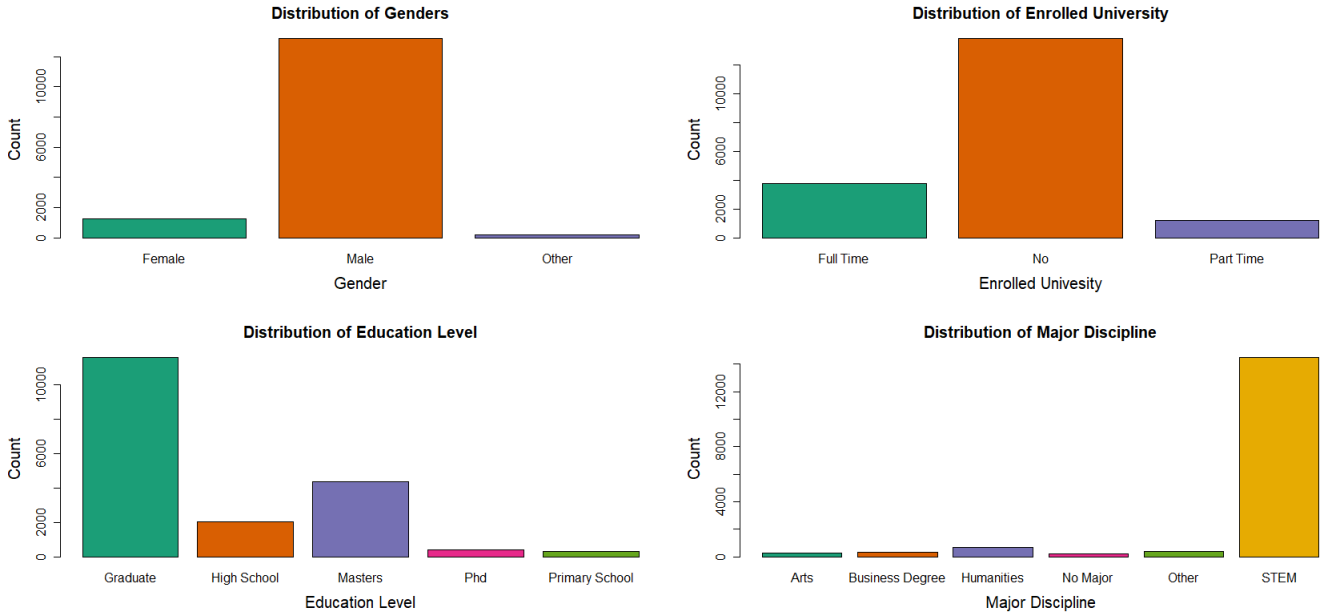


Figure 2: Distribution of Gender, Enrolled University, Education Level and Major Discipline

The fifth categorical variable, **Last New Job** has 6 levels which are *Never*, *1*, *2*, *3*, *4* and *4+*. There are 423 missing values. A majority of the candidates have only been at their previous job for one year. The sixth categorical variable, **Relevant Experience** has two levels which can take on the values *Yes* and *No*. There are no missing values and there are total of 13,792 candidates that do have relevant experience. The seventh categorical variable **Company Size** has 8 levels which are *<10*, *10-49*, *50-99*, *100-500*, *500-999*, *1000-4999*, *5000-9999* and *10000+*. The variable has 5,938 missing values and a majority of candidates come from companies ranging from 50-99 employees. This is the only variable whose missing values outweigh the total of any one level. The eighth categorical variable is **Company Type** which has 6 levels, *Early Startup*, *Funded Startup*, *NGO*, *Other*, *Public Sector* and *Pvt Ltd*. The majority of candidates have previously worked for private companies. These graphs can be found in Figure 3 below.

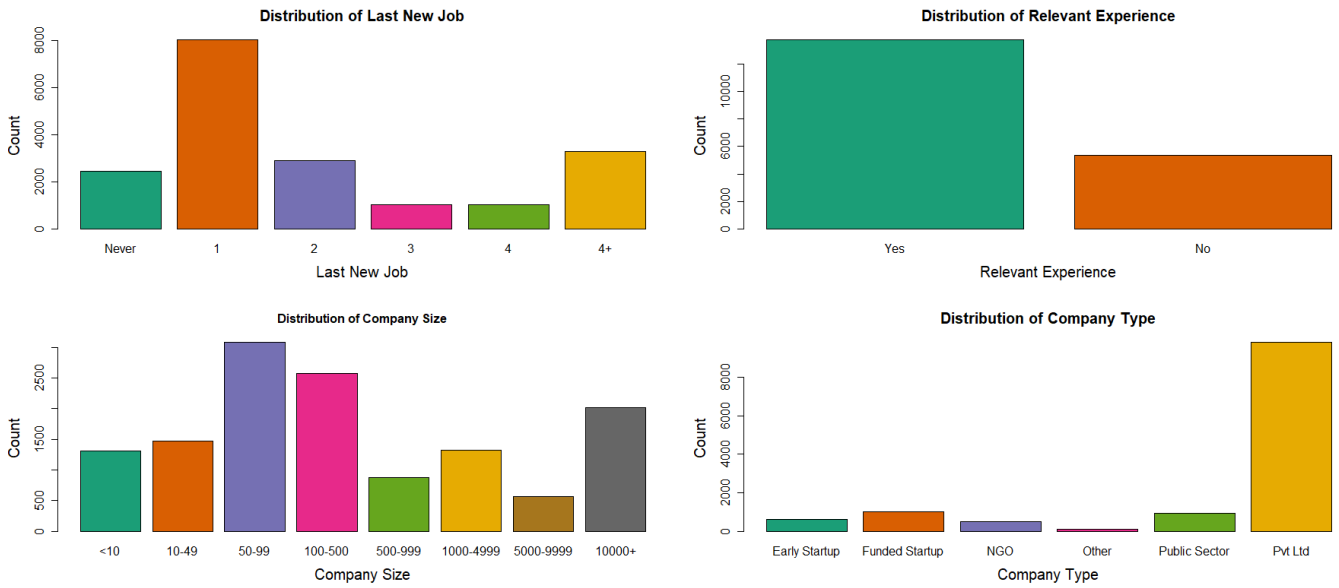


Figure 3: Distribution of Last New Job, Relevant Experience, Company Size and Company Type

The ninth categorical variable **Experience** has 21 levels. These levels range from less than one year of experience to over twenty years of experience. The majority of candidates have over 20 years of experience with a total of 3,286 candidates. There are also a total of 65 missing values for this variable. The graph is displayed in Figure 4 below.

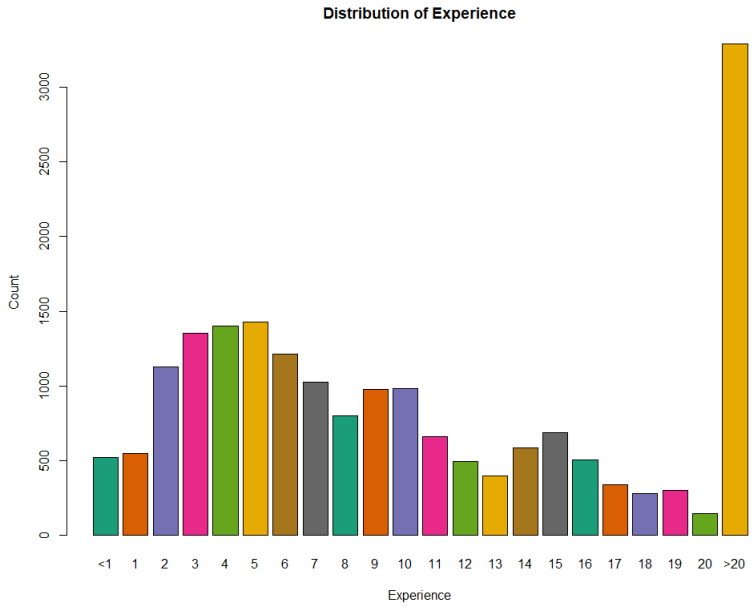


Figure 4: Distribution of Experience

The last categorical variable **City** is the numerical value assigned to the city that the candidate is from. The values for the cities range from 1 to 180. The cities that have the highest number of candidates are 103, 21, 16 and 114 respectively. The graph for this is displayed in Figure 5 below.

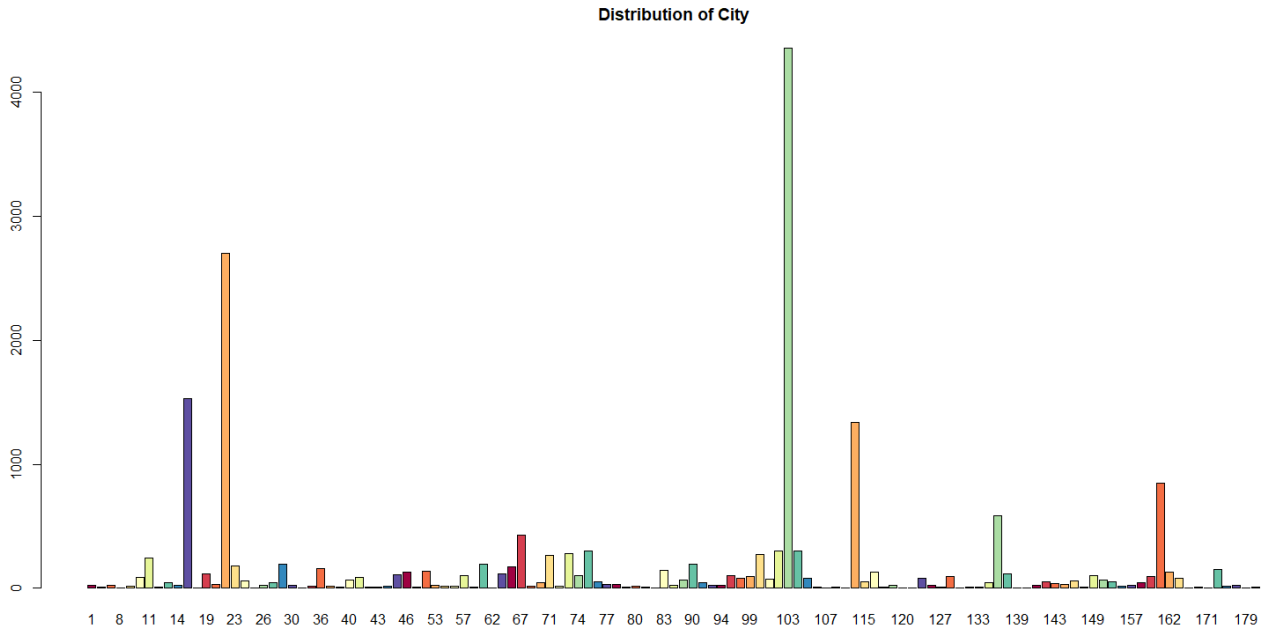


Figure 5: Distribution of City

I will now examine the distributions for the remaining numerical variables **City Development Index** and **Training Hours**. The variable **City Development Index** is displayed in the Figure 6 histogram below. This variable is calculated by taking into account a city's infrastructure, waste, health, education and product. The closer the City Development Index is to 1, the more developed the city is. The distribution for this graph is called bimodal since there are two peaks in the graph and it is also left skewed. The first peak occurs at .6 and the second higher peak occurs at .9. The majority of cities lie in the range above .9.

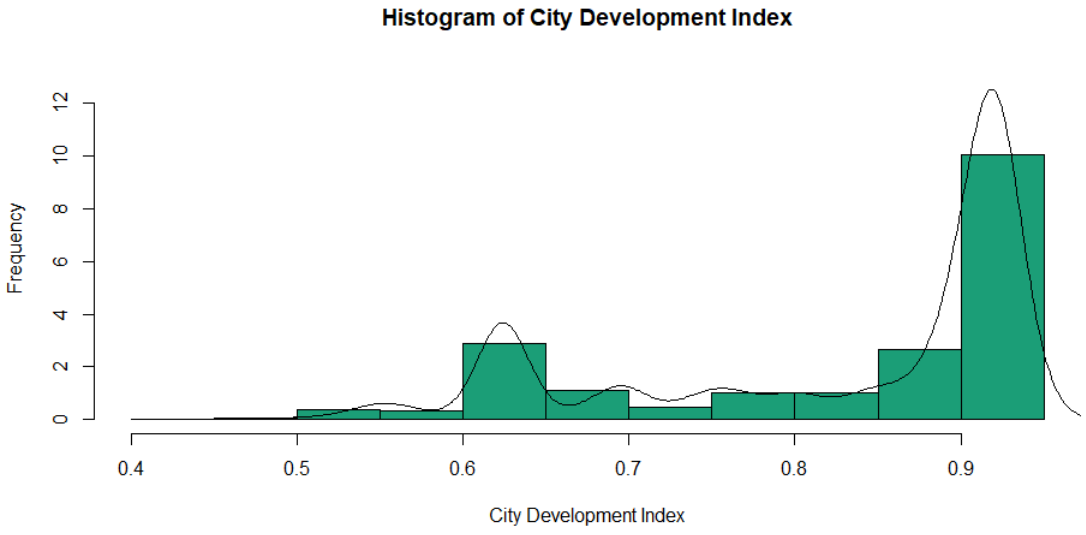


Figure 6: Histogram of City Development Index

The variable **Training Hours** is displayed in the histogram Figure 7 below. This variable displays the total amount of training hours completed by each candidate. The distribution is unimodal which means that the data only has one peak and is right skewed with a majority of the candidates having less than 50 hours of training time.

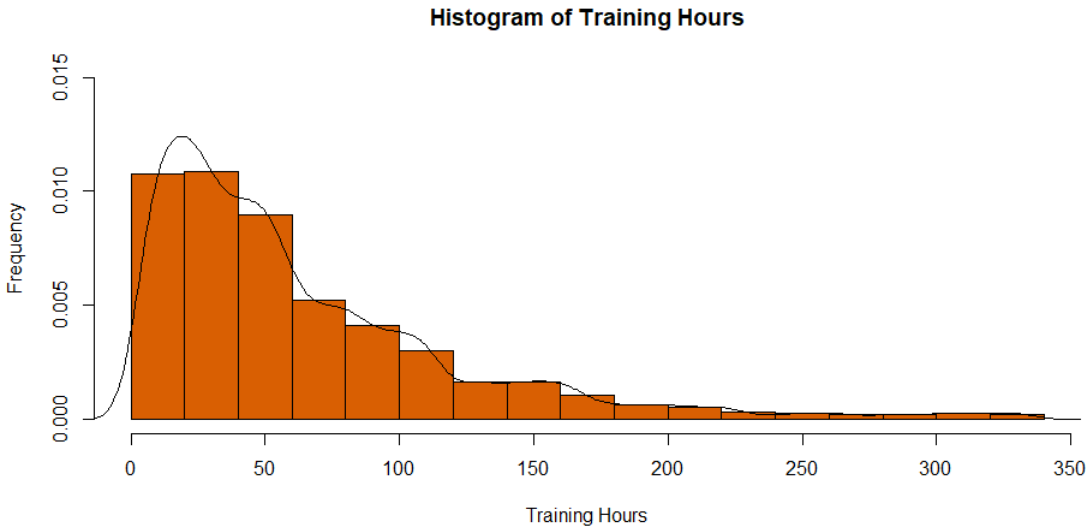


Figure 7: Histogram of Training Hours

Since the purpose of this analysis is determine what factors affect the variable **Target**, I have put together graphs to show how **Target** is affected by each categorical variable above. In Figures 8 and 9, we can see how **Target** is affected by **Gender**, **Enrolled University**, **Education Level** and **Major Discipline**. For each graph the categorical variable's levels are on the horizontal axis and there are two bars for each level. The height of the red bar represents the count of the level given that the candidate did not leave their current job. While, the height of the blue bar represents the count of the level given that the candidates did leave their current job. This will give an overview of how the levels affected the variable **Target**. We see that a majority of candidates that are male have a much higher proportion to stay with their company as compared to the other levels. We also find that the rate at which males left the company was significantly higher than other levels. We also see that not being enrolled in a university has a higher proportion of staying compared to other levels. However, we do not see as drastic of a proportional difference between the enrollees who left and those who are not enrolled and left. In the next graph one interesting take away is the massive difference in proportions between the graduates and masters who leave and stay compared to the other levels. Will the difference in education be a significant predictor in determining employee churn? Lastly, we also see that a majority of applicants that are STEM majors have the highest difference between leaving and staying with their company. While STEM majors massively outweighs the other levels, the proportion of staying is higher than any other level.

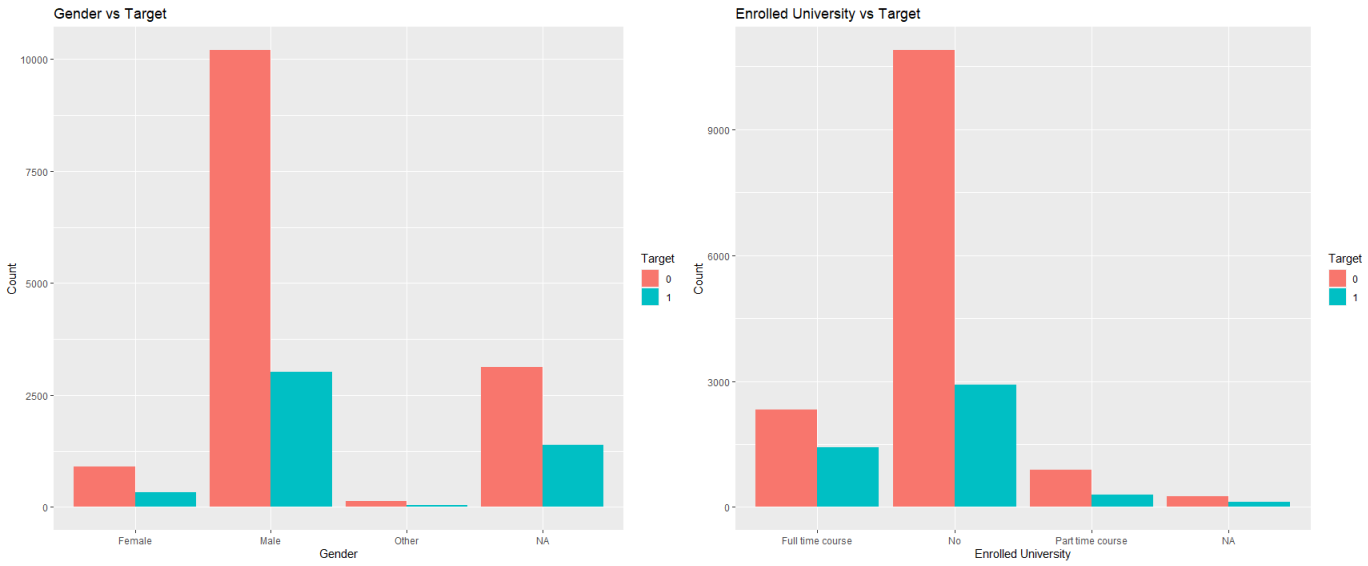


Figure 8: Gender, Enrolled University vs Target

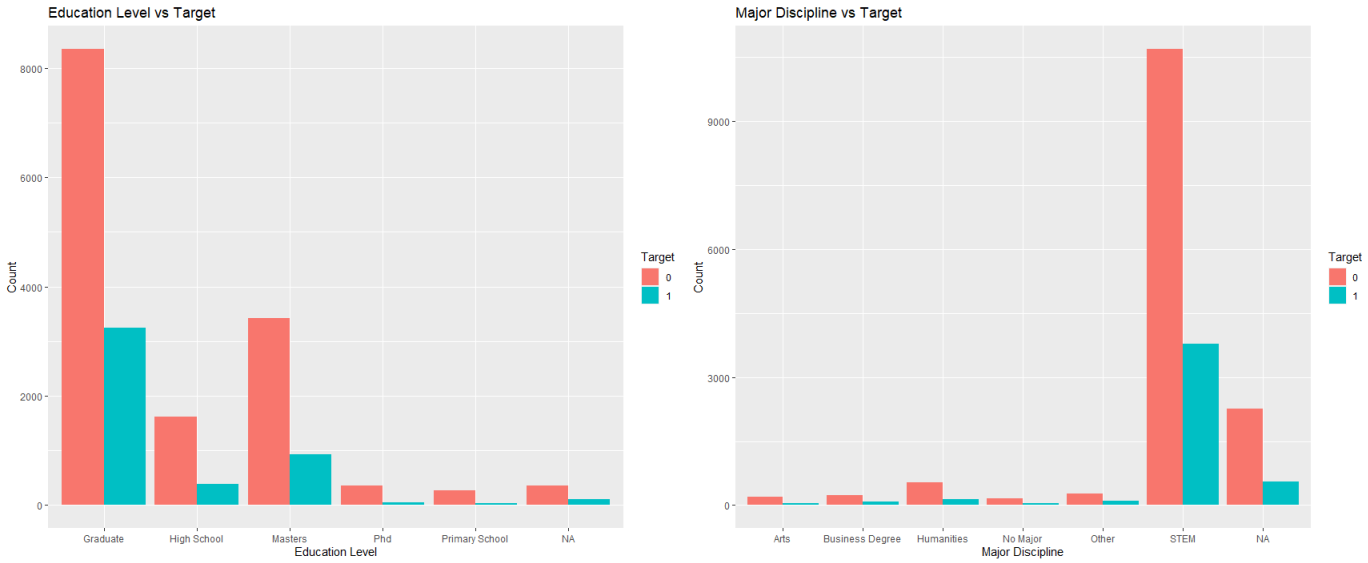


Figure 9: Education Level, Major Discipline vs Target

In the second set of graphs displayed in Figures 10 and 11, we see how **Target** is affected by the variables **Last New Job**, **Relevant Experience**, **Company Size** and **Company Type**. We find that a majority of applicants that have one year of experience were the most likely to stay with their current company. Although, the proportion of those staying compared to those leaving is significantly higher for the candidates who have been at their current company for over four years. In the next graph, we see the effect of **Relevant Experience** on **Target**. Having relevant experience seems to affect staying with a company as opposed to not having experience. The proportion of candidates who have relevant experience stayed more often than left compared to the proportion of candidates who did not have relevant experience. For the next graph, we find that the candidates who did not list their company size tend to have higher rates of leaving than any other level combined. However, the proportion that had the highest rate of staying compared to leaving are the candidates who work at a company with 50-99 employees. Finally, the rate that candidates stayed compared to those that left was far greater for private companies compared to any other type.

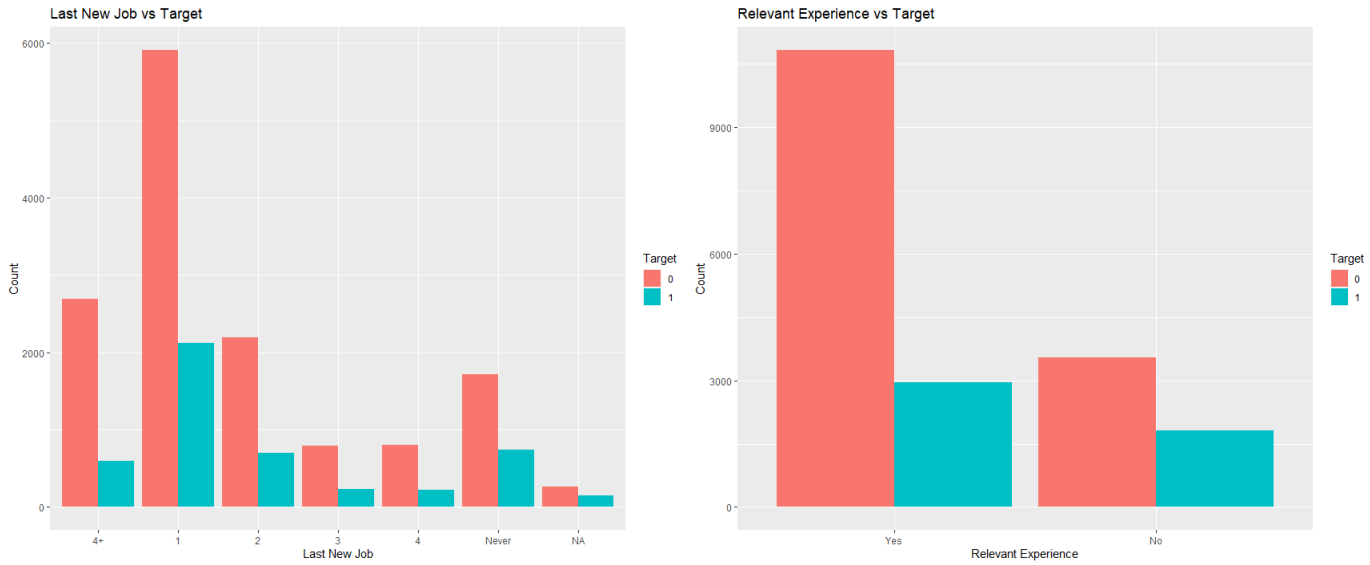


Figure 10: Last New Job, Relevant Experience vs Target

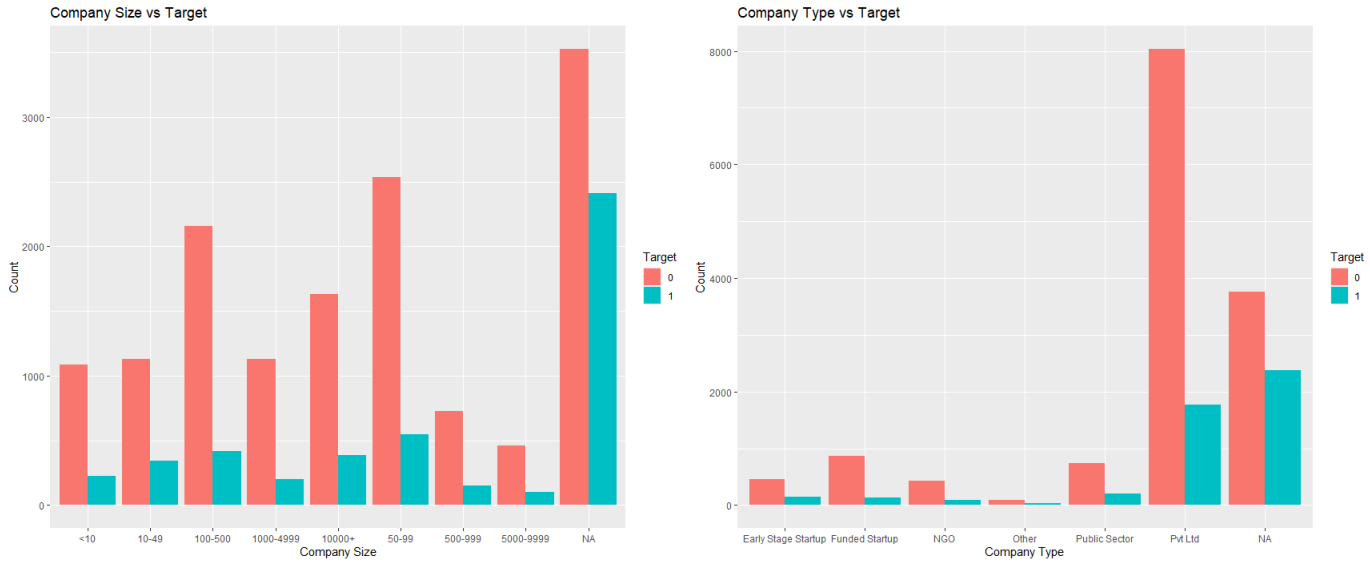


Figure 11: Company Size and Company Type vs Target

The dependent variable, **Target**, can take one of two values, 0 if the candidate did not leave their current company and 1 if the candidate did leave their current company. This variable does not have any missing values and the number of candidates that did not leave is 14,381 and the number of candidates that did leave is 4,777, this is visualized in Figure 12.

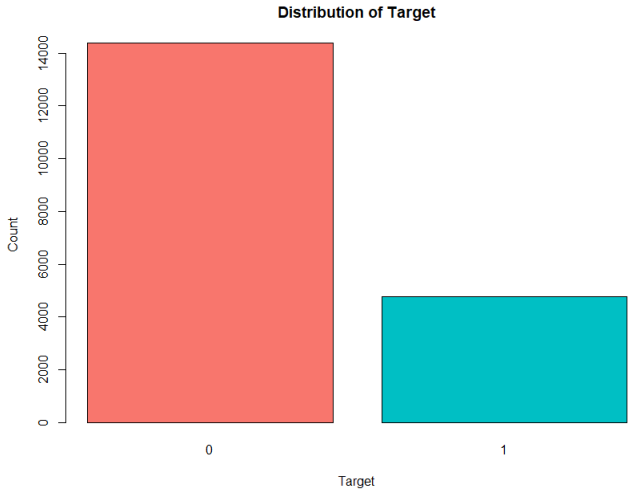


Figure 12: Distribution of Target

In order to perform the analysis and build the model, I will need to clean up the data. One of the main advantages of getting an overview of the data using graphs, is it shows how much missing data there is and any possible errors withing the data. The first step of cleaning data would be to remove any variables that do will not add any significant information to my model. Since the variable **Enrollee Id** is a unique value for each observation, this will not add any significant information and can be dropped. I will next need to convert the **City** variable into a numerical variable. This was done by dropping the "city_" prefix to extract the numerical portion of the value. One disadvantage to using logistic regression is that the variables used to build the model cannot be categorical levels. I will need to convert each of the categorical variables into numeric values. For the ordinal categorical variables such as **Education Level**, **Experience**, **Last New Job** and **Relevant Experience**, they have an order that can be interpreted. When converting factors to numbers, the lowest number in our analysis will be 0, will correspond with the lowest ranked level. Therefore, the variable **Education Level** will be coded as *Primary School* = 0, *High School* = 1, *Graduate* = 2, *Masters* = 3 and *Phd* = 4. Similarly, the variable **Experience** will take on the values 0 to 21 corresponding to the years of experience *<1* to *>20* respectively. The values representing years for **Last New Job** will be coded similarly to how the variable **Experience** was handled. The levels *Never* to *4+* with be replaced with the values 0 to 5. Also, **Relevant Experience** will be coded as 0 for *No* and 1 for *Yes*. The remaining categorical variables will be coded with integers ranging from zero to the number of levels for that categorical variable. The order for these codings does not matter as much and each variable was coded by their levels in alphabetical order. Now that the categorical variables are converted to numerical variables, the missing values will need to be handled.

One of the biggest issues when performing an analysis is how to handle missing data. In this data set there are missing data values for the variables **Gender**, **Enrolled University**, **Education Level**, **Major Discipline**, **Experience**, **Company Size**, **Company Type** and **Last New Job**. In order to handle these missing values, I will use the Multiple Imputation by Chained Equations function in R. This will allow me to impute multiple values at once using the predictive mean matching method. The predictive mean matching method allows me to fill in imputed values by choosing values that closely match other values in the data set according to ? (?). Using MICE will allow me to use a robust and easy method to impute or replace missing data quickly. The way that MICE works is by imputing values observed in the data set, it will choose a data point from the original data set which has a predicted value close to the missing sample (?, ?). However, one disadvantage to this technique is the computation time is slower than other traditional techniques such as using the mean or median. Now that the missing data is cleaned, I will be able to build a logistic regression model.

4 Analysis

The purpose of this analysis is to create a logistic regression model that will be able to identify which factors are the most statistically significant. The reason that I have chosen to use logistic regression is because it is one of the more simpler methods to understand while also providing high accuracy models according to Tuffery (2011). However, in terms of programming one of the main disadvantages is that programming the method can be complex especially when trying to find the best fit model. The logistic regression function can be defined as the equation

$$ln(\frac{P}{1-P}) = \beta_0 + \beta_1 X_1 + ... + \beta_n X_n$$

Where P is the probability that the dependent variable is a 1, the β_i's are the coefficients of the independent variables, X_i where i = {1 to n}. This formula is also called the log odds. We are interested in building a model that will let us determine the coefficients for the equation on the right above. Using this equation, I will be able to predict a candidates probability of leaving by substituting values corresponding to a given candidates information. To accomplish this I will need to make sure that the data first satisfies the assumptions necessary for logistic regression. One major assumption of logistic regression is that the numeric independent variables are not highly correlated with one another according to Tuffery (2011). Any variable with a correlation higher than .9 and less than -.9 is considered highly correlated. In Figure 13 displayed below, we can see that there are variables with a medium correlation.

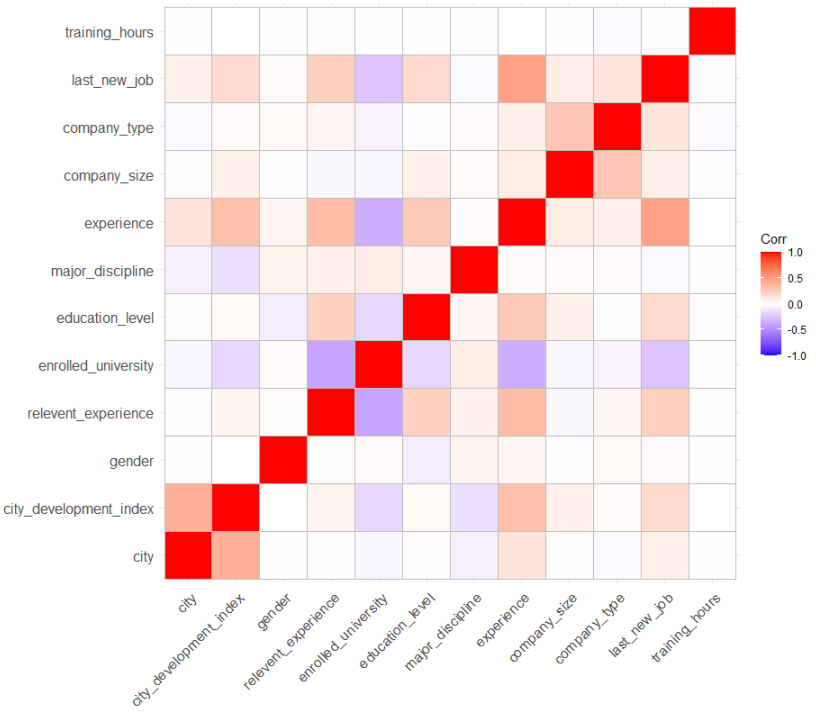


Figure 13: Correlation Plot

The darker colored squares signify stronger correlations between the two intersecting variables. Two variables that are positively correlated are **Last New Job** and **Experience**. We also see that **Enrolled University** is strongly negatively correlated with both **Experience** and **Relevant Experience**. However, since neither of these variables are over the ±.9 threshold and not strongly correlated with other variables, this assumption is met. To validate the logistic regression model, I will need to split the data set into a training and a testing set. The training set will comprise of 75% of the data randomly selected without replacement and the testing set will comprise of the remaining 25%.

First, I will build a logistic regression model with all of the variables included. This will be called Model 1. In the below Model 1 output on the left, we find that most of the independent variables are statistically significant at the except for **City** and **Major Discipline** as seen in Figure 14. A statistically significant variable is one that has a p-value ≤ 0.05. In order to build a better model, I will need to remove the statistically insignificant variables and check the Akaike Information Criterion, AIC, which will tell me if removing a variable leads to an overall better fit. I will do both of these steps by using R's built in step function which will automatically add and remove variables to test if the AIC value lowers which indicates a better model. A lower AIC value means that the model has improved the fit of the data compared to the model before (Tuffery, 2011). Also in Figure 14, the output for the stepwise selection of models is displayed with the final AIC value of 14259 on the right. Even though the variable **Major Discipline** has a p-value of 0.11158, removing that variable did not lead to an overall better fit of the data to the model. In fact, removing that variable increased the AIC value. I will therefore leave that variable in the final model.

```
Call:
glm(formula = target ~ ., family = binomial(link = "logit"),
    data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8012  -0.6957  -0.5285  -0.3709   2.3397

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)             3.9332010  0.2088245  18.835  < 2e-16 ***
city                   -0.0005874  0.0004897  -1.200  0.23026
city_development_index -5.5985165  0.1872041 -29.906  < 2e-16 ***
gender                 -0.2027056  0.0667158  -3.038  0.00238 **
relevent_experience    -0.4868660  0.0503895  -9.662  < 2e-16 ***
enrolled_university     0.1764319  0.0269045   6.558 5.46e-11 ***
education_level         0.1367718  0.0325421   4.203 2.63e-05 ***
major_discipline       -0.0336303  0.0210598  -1.597  0.11029
experience             -0.0206053  0.0041580  -4.956 7.21e-07 ***
company_size            0.0397633  0.0100258   3.966 7.31e-05 ***
company_type           -0.0339667  0.0138789  -2.447  0.01439 *
last_new_job            0.0463535  0.0150063   3.089  0.00201 **
training_hours         -0.0009420  0.0003529  -2.669  0.00761 **
---
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: 14234  on 14355  degrees of freedom
AIC: 14260

Number of Fisher Scoring iterations: 4
```

(a) Model 1 Output

```
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
```

(b) Model 2 Output

Figure 14: Logistic Regression Outputs

Now that I have a final model, I will need to validate how well the data will perform on the testing data set. To test this, I will use the built in function, predict, to generate new values for **Target** using the model. The model will calculate a probability, if the candidate will leave their current job. If the probability is less than .50 then the predicted value for **Target** is 0 meaning that the candidate did not leave their current job and if the probability is equal or greater than .50 then the predicted value for **Target** is 1 meaning that that the candidate did leave their current job. The predicted value of **Target** will then be compared to the actual value of **Target** for the observation. The results for the testing data set can be found in the below confusion matrix. We find that the model has an accuracy of 76.08%, a sensitivity of .936 and a specificity of .2328. According to Tuffery (2011), the sensitivity, α(s), is the probability of correctly detecting an event at the threshold s = Prob(score(x)$\geq$s | x=event) and the specificity, β(s), is the probability of detecting a non-event at the threshold s = Prob(score(x)<s | x=non-event). In this case, I have chosen s to be equal to .50. These two values will be plotted to create the receiver operating characteristic (ROC) curve. Also according to Tuffery (2011), the ROC curve shows sensitivity as a function of the specificity. The main takeaway from this graph is the area under the curve (AUC), which tells how well the score discriminates between the events and non-events. As the AUC approaches 1 then the score discriminates perfectly between populations compared to an AUC approaching 0 which indicates the model is declining in performance as stated by Tuffery (2011). We find that the AUC is equal to .719 which is decent at determining between the values for **Target**. One of the main advantages of displaying the ROC curve, is it displays how well the model performs compared to a random model which is the diagonal line in the graph (Tuffery, 2011). One disadvantage of this method, is that it is not as intuitive as other methods since one has to understand the definitions of sensitivity and specificity and how they are calculated. The output of the confusion matrix

and ROC curve are displayed in Figure 15.

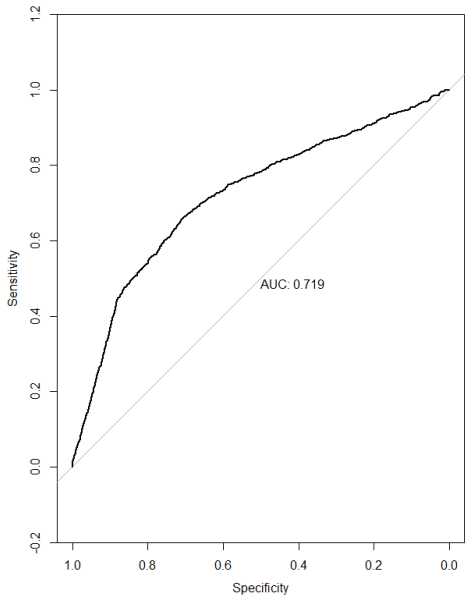
Since the main purpose of this analysis was to determine which factors are significant in determining if a candidate will leave their current company, I have built a model with statistically significant factors. The final model can be written as

$$3.961 - 5.693\beta_1 - .203\beta_2 - .484\beta_3 + .176\beta_4 + .137\beta_5 - .034\beta_6 - .021\beta_7 + .040\beta_8 - .034\beta_9 + .046\beta_{10} - .001\beta_{11}$$

Where β_1 = **City Development Index**, β_2 = **Gender**, β_3 = **Relevant Experience**, β_4 = **Enrolled University**, β_5 = **Education Level**, β_6 = **Major Discipline**, β_7 = **Experience**, β_8 = **Company Size**, β_9 = **Company Type**, β_{10} = **Last New Job** and β_{11} = **Training Hours**. The above equation represents the change in log odds for every change in each of the β_i for i = {1 to 11}.

```
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 15: Confusion Matrix and ROC Curve

5 Data Summary and Implications

For the final model, we find that the statistically significant factors are **City Development**, **Gender**, **Relevant Experience**, **Enrolled University**, **Education Level**, **Major Discipline**, **Experience**, **Company Size**, **Company Type**, **Last New Job** and **Training Hours**. Therefore we reject the null hypothesis and accept the alternative hypothesis proposed at the beginning of the paper. With the AUC of .719 there is room for improvement. In order to improve this score, I could use the approach of Yigit and Shourabizadeh (2017) to create different models and test their accuracy. Another improvement that could be done would be to gather more data such as the employee's satisfaction level with their current employer, their current salary and also what topics were completed in their training hours. This would potentially improve the model and also give the data analytics company more factors that will allow them to determine how to attract more talent to their company.

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

List of Figures