James Anderson
Benjamin Kanarick
Rochan Nehete
Viswa Tej

THE UNIVERSITY OF TEXAS AT AUSTIN

McCOMBS SCHOOL OF BUSINESS

Data Science Programming Final Project

Group 8

Predicting Job Change of Data Scientists:

Description

For our final project, we will explore a dataset compiled to understand which factors lead a person to leave their current job as a data scientist. Using this data, we'll conduct an exploratory data analysis to find patterns and detect trends. This data set contains information on individuals who have successfully passed training courses for employment that were conducted by an undisclosed company (referred to as Company X) interested in Big Data / Data Science. This helps Company X group the valid candidates that would work for them and cater the training courses to match the specific needs and qualifications of this group, ultimately leading to more polished and educated employees for the company.

On a macro-scale, we chose this data set and question at hand due to the benefit it can provide major companies and corporations in an economy. From these predictions, we can give Company X (and many others too) a sense of understanding pertaining to the factors that lead an individual to leave a current job as well as identify trends similar trends within individuals that lead them to gravitate towards one company over another. As stated, it is crucial if a company knows (to some extent) the factors, trends, and profile of a candidate who is likely to come work for them vs. who would not. If a company knows exactly who to target and who to not, one could assume a plethora of time, resources, and money could be saved and properly allocated.

Exploratory Analysis of the Data

For our data, the set is split into a training set and a test set which consists of 19,158 and 2129 observations, respectively. There are 14 columns in the training set and 13 columns in the test set, for the test set does not have the "Target" column of 0 or 1. This target column is what we are solving for, a 1 indicates that given the observations of an individual they are willing to work for Company X. On the other hand, a 0 indicates that an individual would not switch jobs/work even though they filled out the initial training courses provided by Company X. As for each observation (individual) in both the training and test set, the predictors are indicated in Exhibit 1 and include factors such as relevant experience, gender, and current company size.

After the data was cleaned and all null/missing values were accounted for, we plotted a correlation heatmap for all numeric attributes, as seen in Exhibit 2. Furthermore, we took these numeric values and made a pair plot in order to see the distribution of each observation in regards to if said observation had a “Target” value of 0 or 1. This pair plot can be found in exhibit 3 of the appendix. As displayed by Exhibits 2 and 3, these plots cannot determine the significant correlation between any of the numeric attributes, so we took a further look at the rest of the predictors in the set.

In order to display the presence of the non-numeric predictors, we plotted bar plots for each, and they are displayed in Exhibits 4-8. In each barplot, the overwhelming majority of each category favours “Target” = 0. This indicates that, regardless of the attribute, more individuals are not willing to leave their current role/work for Company X compared to those who are. An interesting trend we noted can be found in exhibit 5, the influence of education level on target. As an individual's education level increases, the % of “Target = 1” decreases, indicating that more decorated academics prefer to stay at their current jobs. We inferred from this that individuals with higher degrees make better candidates for a role, and therefore are more likely to have a position they fancy and would not want to leave.

In exhibit 9 we can see that the peak for density of both the target variables reach their respective peaks within the 0-50 training hour range. This shows that the majority of the employees already tend to make up their mind within the first 50 hours of their training.

In exhibit 10, we see the effect of city development on how likely an individual is to switch jobs. The graph indicates that the more developed a city is, the less likely one is to move jobs or be classified as “target” = 1.

Solution and Insights

Before we explain our modelling approaches, it is important to note how we scaled and encoded the data in order to be used in our classification models. We have used the MinMaxScaler() function on our measure columns in order to have a normal distribution as well as reduce the influence of outliers. As for the dimensions, we have implemented one_hot_encoding() for the purpose of transforming the dimensions into a boolean value that is understandable by our classification models. We also removed the columns enrollee_id, city and company size as they were not showing any relevance towards the data in our EDA.

Also prior to modelling, we upsampled using the SMOTE() function to assure that there are an equal amount of “target” = 0 and “target” = 1 when creating our training and test sets. In the EDA section, it is noted that there is much more “target” = 0 observations (roughly an 80/20 split), and for our models, we wanted to assure that this imbalance was fixed. SMOTE would select examples that are close in the feature space, draw a line between the examples in the feature space and draw a new sample at a point along that line.

Models

We split our data into a 25-75 test-train split which we will use for the following models:

- Logistic Regression
- KNN
- Random Forest
- Multinomial Naive Bayes

Logistic Regression

We used logistic regression initially to classify the test set we created. We set the max iterations to 1000 and fitting the data, we obtain an accuracy of 74.4% with a precision of 73.7% and a recall of 76.6%. Logistic regression gives us a false positive of 13.78%. These results can be found in exhibit 11 of the appendix

K-nearest Neighbors

For K-nearest Neighbors, we try to fit our model for different values of k ranging from 0 to 40. Upon plotting the error rate against the tried K value, we get the least error rate for K = 3. This graph can be found in exhibit 12 of the appendix. Hence upon selecting K=3, we get an accuracy of 77.5% with a precision of 74%, both numbers slightly higher than that of logistic regression. Our false positives are up to 15.1%, higher than that of Logistic Regression.

Random Forest Classifier

For our Random Forest Classifier, we built decision trees on different samples of our training set and averaged them out across all the different trees, leaving us a tree with 82.87% accuracy. Out of all the models we ran, we found this model to be the most accurate and practical in regards to our problem at hand. Our Random Forest Classifier had a small false positive rate of only 9.06%. The total matrix of results from our Random Forest Classifier can be seen in exhibit 13 of the appendix.

Multinomial Naive Bayes

Finally, we chose the Multinomial Naive Bayes model (MNB) because it is also another useful way to analyze and predict categorical data. Unfortunately, the MNB model only performed at an accuracy level of 65.15%, with a false positive of 13.36%. Even though the MNB model may not be the best fit to solve and explain our problem at hand, it is a great reference to gauge how much better some of our other models (such as RF classifier) performed. The complete summary output for the MNB model is displayed in exhibit 14 of the appendix.

Conclusion

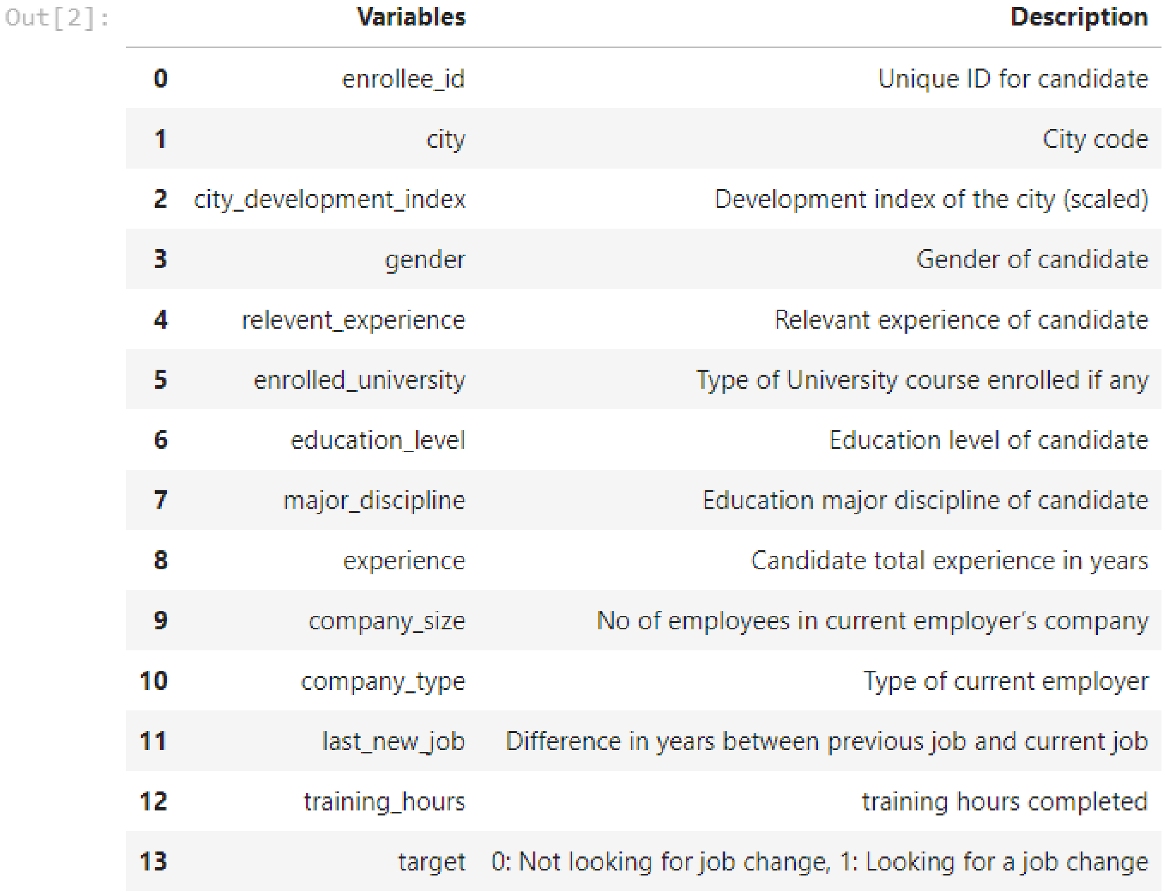
In summation, we explored a dataset compiled to understand which factors can lead a person to depart their current job as a data scientist. We then conducted an exploratory data analysis to conduct trends in our set and then modelled the data using various methods such as KNN and RF Classifier.

In our results, we found that the most important attributes that affected an individual's decision were the city development index, training hours, and the last new job held. On the contrary, we noticed that attributes such as experience and gender had much less impact in regards to the decision-making process.

As for the specifics of our modelling, we came to the conclusion that the Random Forest Classifier was the most beneficial model, with an accuracy of 82.87%.

Appendix

Exhibit 1: Names of all predictors in the training and test set. Note: the ‘target’ response variable is not indicated in the test set.



Out[2]:

	Variables	Description
0	enrollee_id	Unique ID for candidate
1	city	City code
2	city_development_index	Development index of the city (scaled)
3	gender	Gender of candidate
4	relevent_experience	Relevant experience of candidate
5	enrolled_university	Type of University course enrolled if any
6	education_level	Education level of candidate
7	major_discipline	Education major discipline of candidate
8	experience	Candidate total experience in years
9	company_size	No of employees in current employer's company
10	company_type	Type of current employer
11	last_new_job	Difference in years between previous job and current job
12	training_hours	training hours completed
13	target	0: Not looking for job change, 1: Looking for a job change

Exhibit 2: Correlations between all numeric attributes

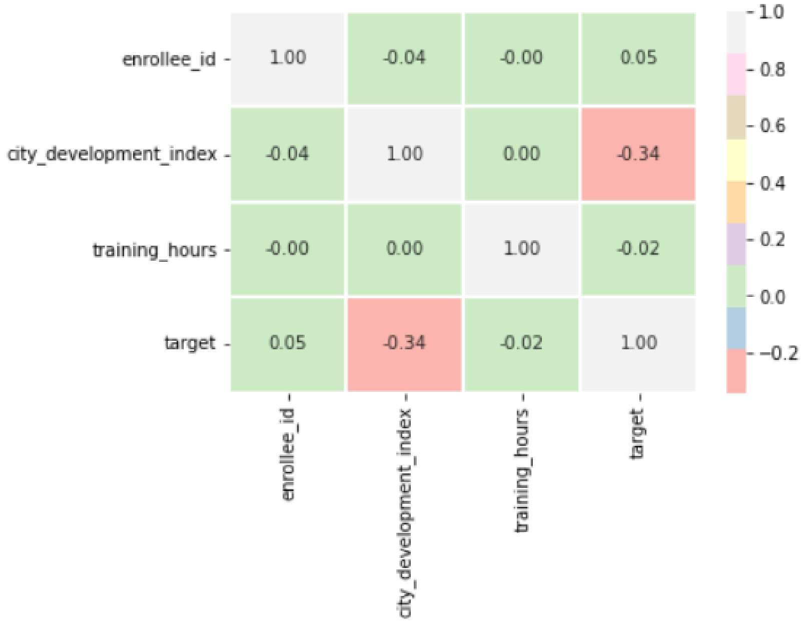


Exhibit 3: Pairplot of all numeric variables in the data set

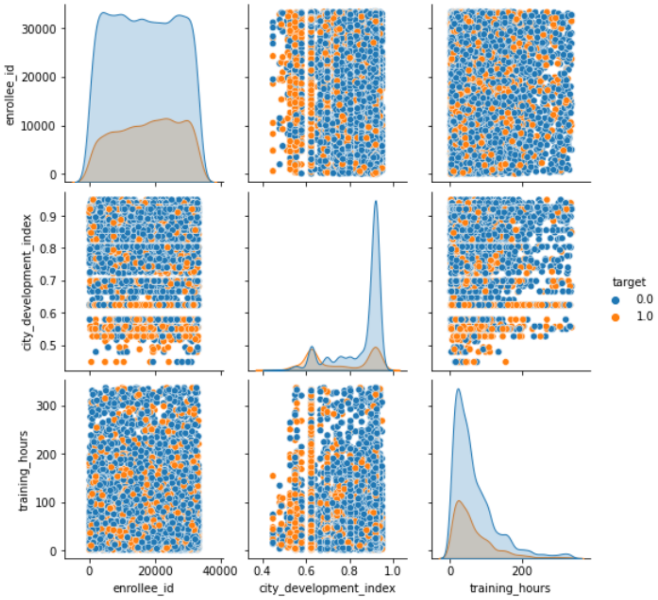


Exhibit 4: Influence of gender on target

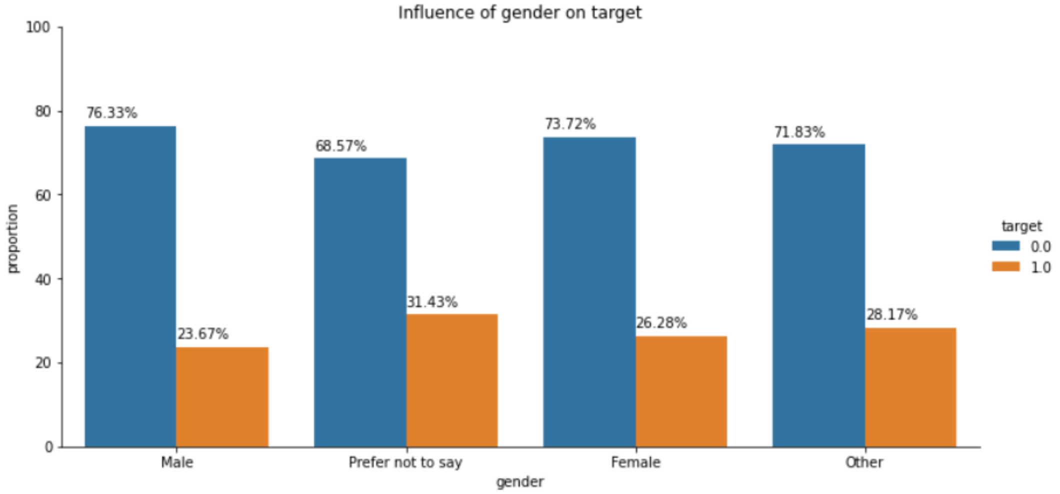


Exhibit 5: Influence of education level on target

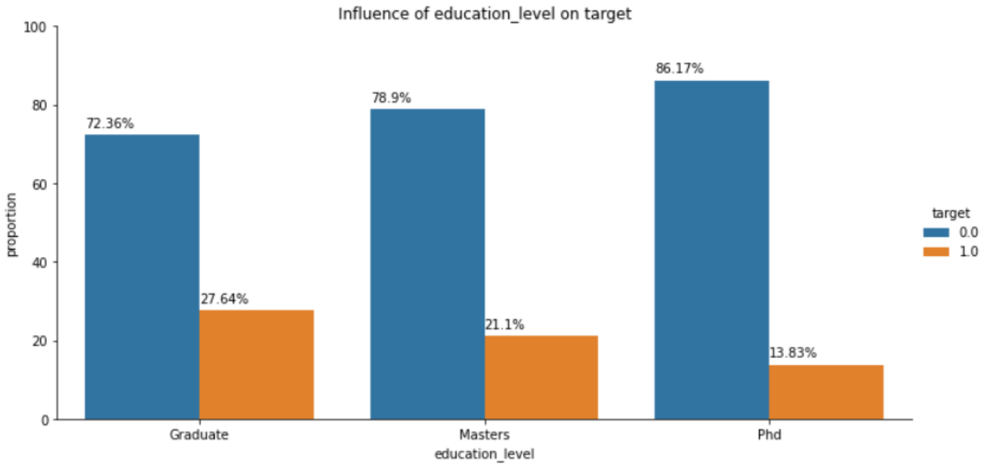


Exhibit 6: Influence of major discipline on target

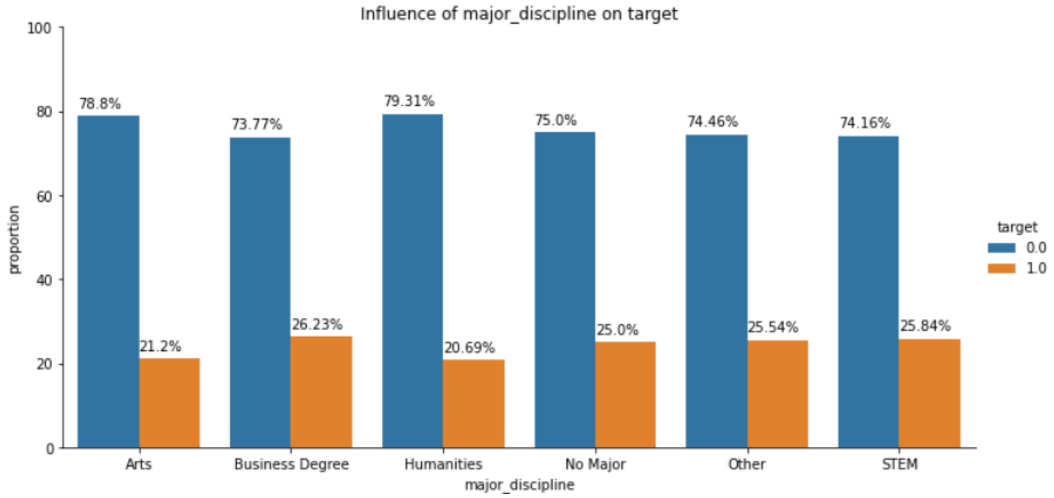


Exhibit 7: Influence of last new job on target

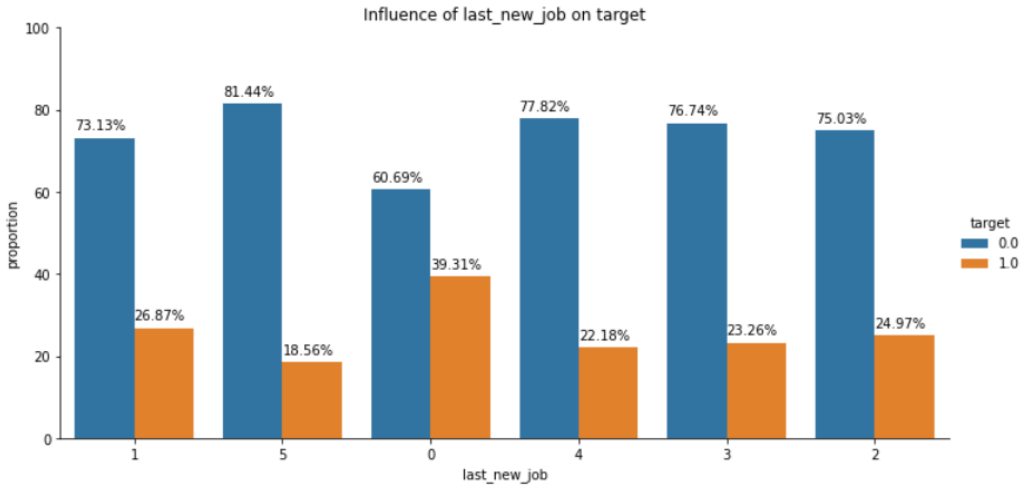


Exhibit 8: Influence of company size on target

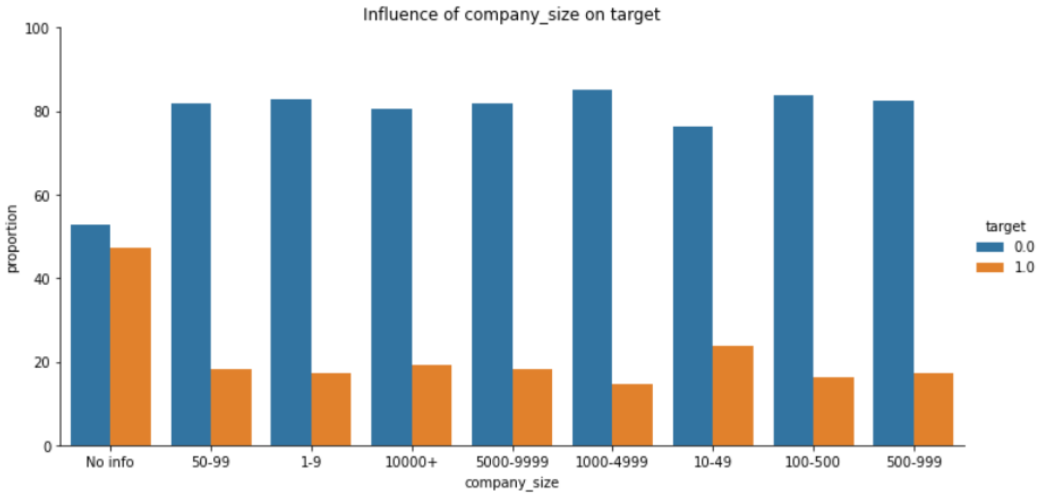


Exhibit 9: Influence of training hours on target

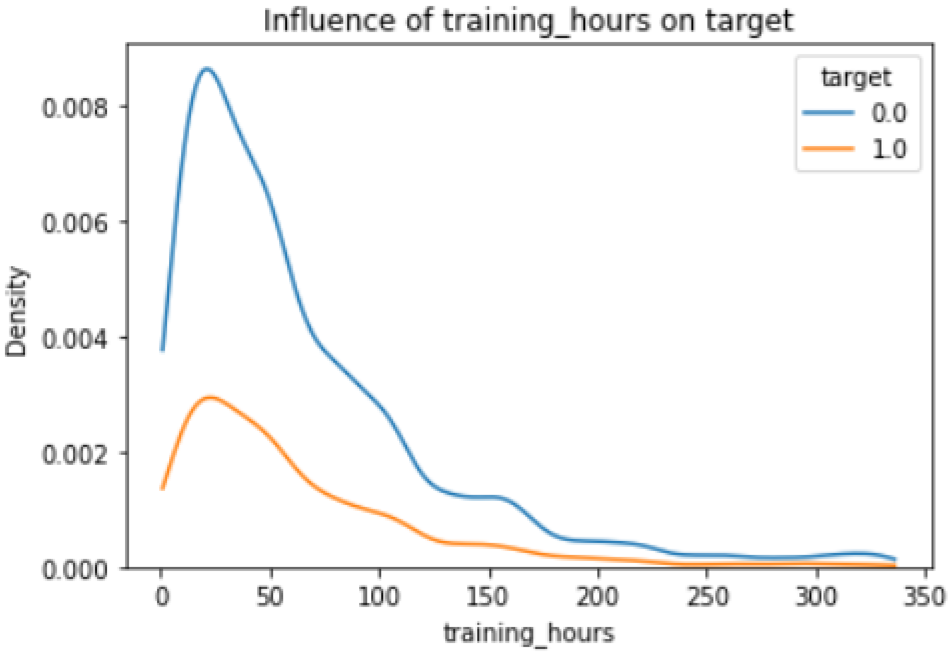


Exhibit 10: Influence of city development on target

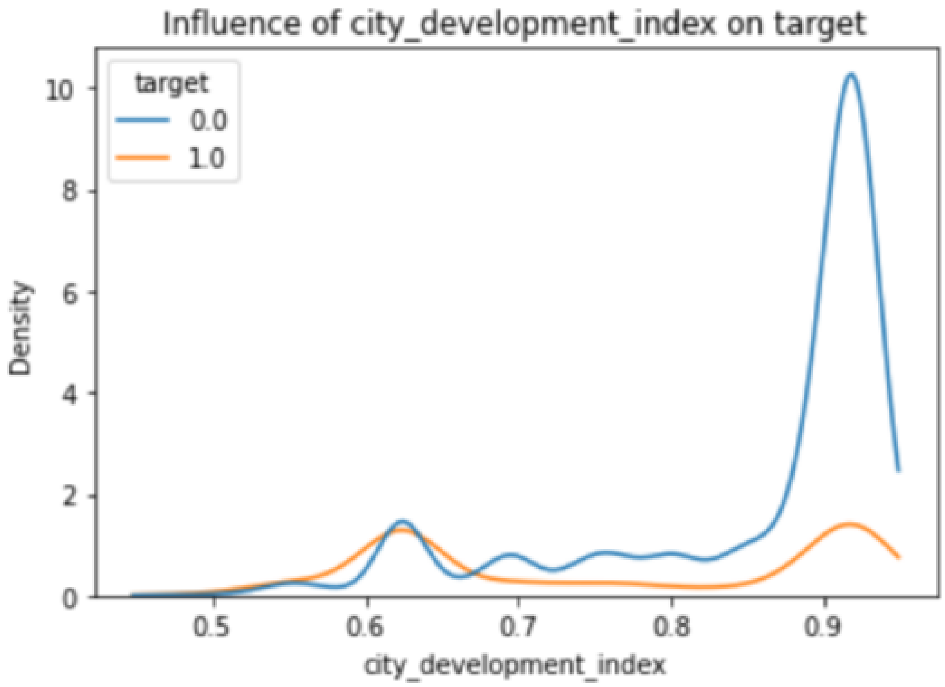
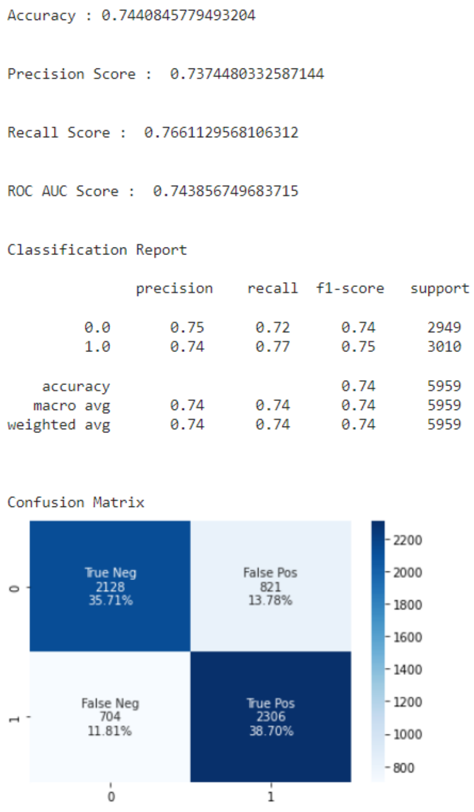


Exhibit 11: Logistic regression results



```
Accuracy : 0.7440845779493204

Precision Score :  0.7374480332587144

Recall Score :  0.7661129568106312

ROC AUC Score :  0.743856749683715

Classification Report

              precision    recall  f1-score   support

         0.0       0.75      0.72      0.74      2949
         1.0       0.74      0.77      0.75      3010

    accuracy                           0.74      5959
   macro avg       0.74      0.74      0.74      5959
weighted avg       0.74      0.74      0.74      5959


Confusion Matrix
```

Exhibit 12: KNN selection of K

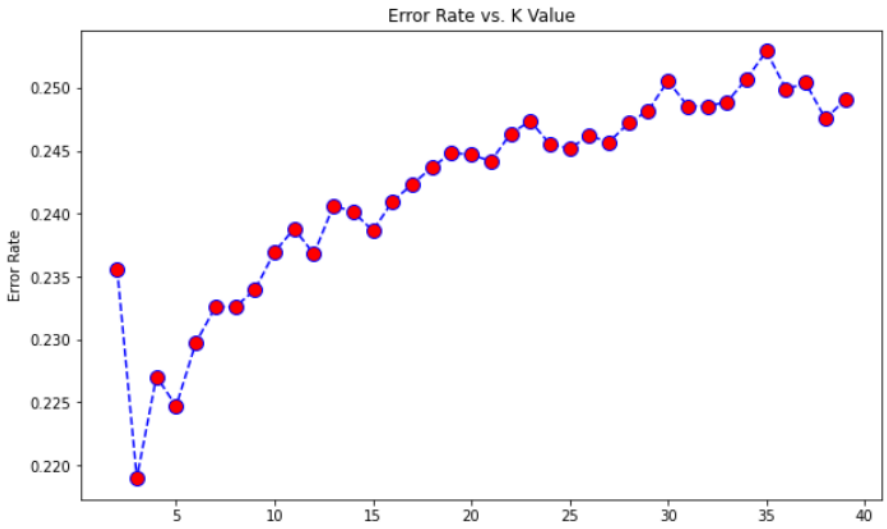
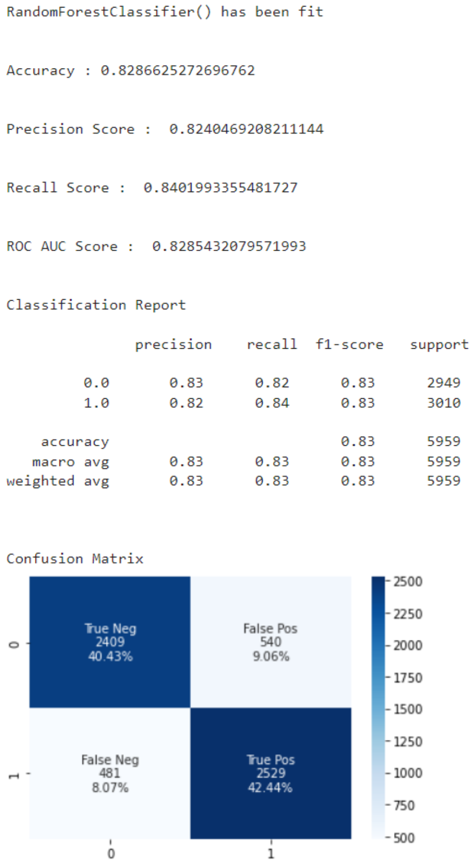
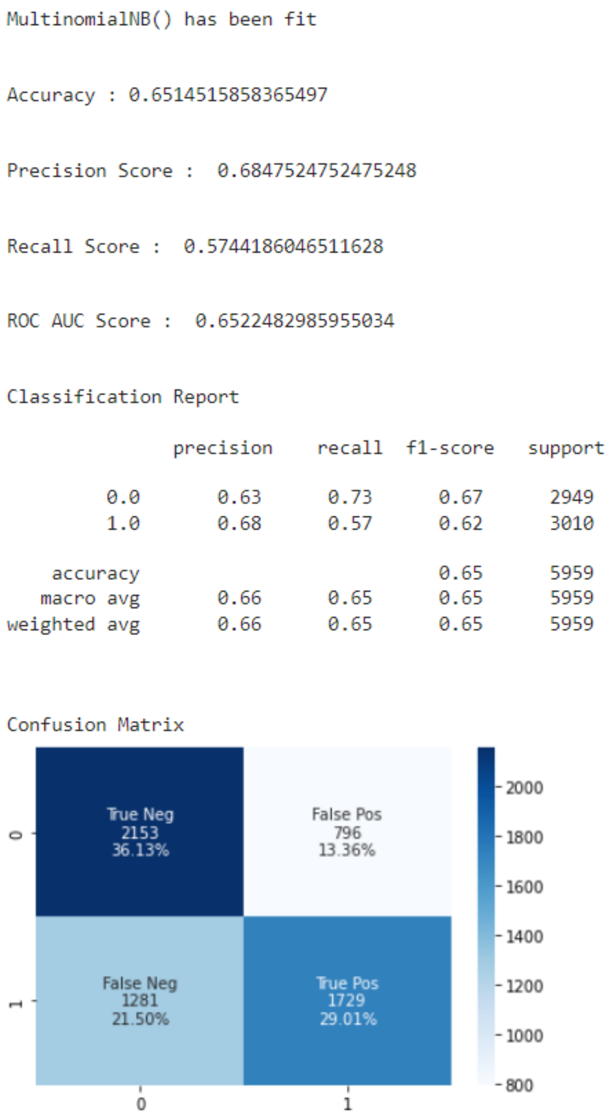


Exhibit 13: Results of Random Forest Classifier



```
RandomForestClassifier() has been fit


Accuracy : 0.8286625272696762


Precision Score :  0.8240469208211144


Recall Score :  0.8401993355481727


ROC AUC Score :  0.8285432079571993


Classification Report

              precision    recall  f1-score   support

         0.0       0.83      0.82      0.83      2949
         1.0       0.82      0.84      0.83      3010

    accuracy                           0.83      5959
   macro avg       0.83      0.83      0.83      5959
weighted avg       0.83      0.83      0.83      5959


Confusion Matrix
```

Exhibit 14: Results of Multinomial Naive Bayes Model



```
MultinomialNB() has been fit


Accuracy : 0.6514515858365497


Precision Score :  0.6847524752475248


Recall Score :  0.5744186046511628


ROC AUC Score :  0.6522482985955034


Classification Report
              precision    recall  f1-score   support

         0.0       0.63      0.73      0.67      2949
         1.0       0.68      0.57      0.62      3010

    accuracy                           0.65      5959
   macro avg       0.66      0.65      0.65      5959
weighted avg       0.66      0.65      0.65      5959


Confusion Matrix
```

References

HR Analytics: Job Change of Data Scientists (2020. Retrieved from https://www.kaggle.com/datasets/arashnic/hr-analytics-job-change-of-data-scientists?select=sample_submission.csv

Mobius. License: CC0 Public Domain