FRAUD DETECTION MODEL FOR ONLINE PAYMENT SERVICE

JIBOKU OLUWATUNMISE OLANREWAJU

THE PROBLEM

Blossom Bank is a multinational financial services group, that offers a variety of financial products, headquartered in London,UK. Blossom Bank has recently detected some fraudulent transactions through their data. In a bid to actualize future fraudulent transactions, there would be a need for a Machine Learning model to be established.

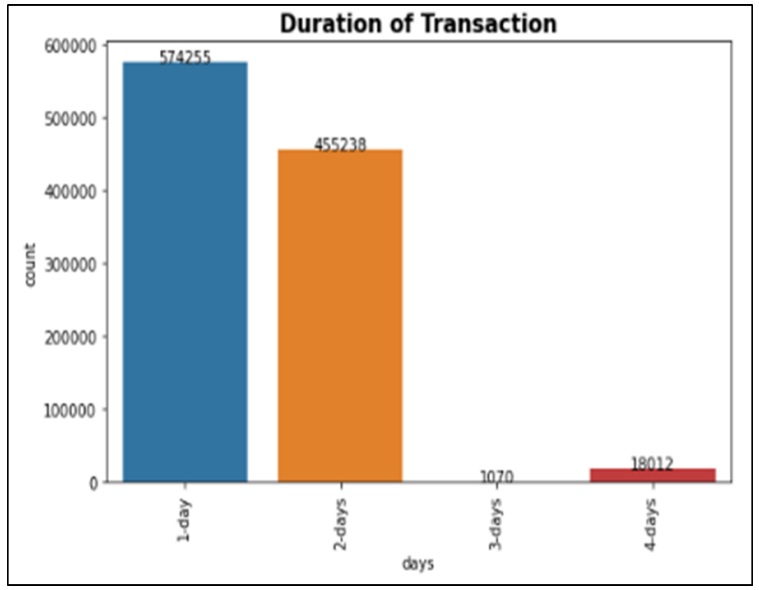
THE SOLUTION

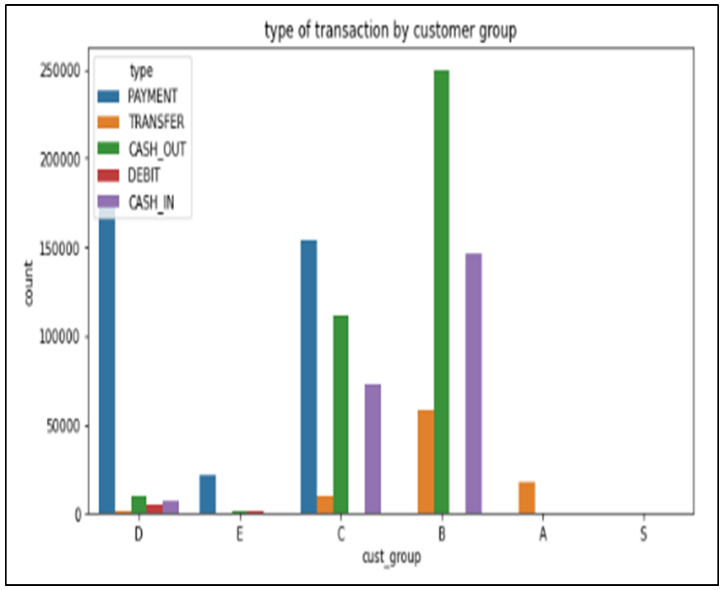
The aim of this project is to use different machine learning algorithms to predict which customers will be the most likely to subscribe to the bank's new product –bank term deposit. Towards this aim, the objectives were:

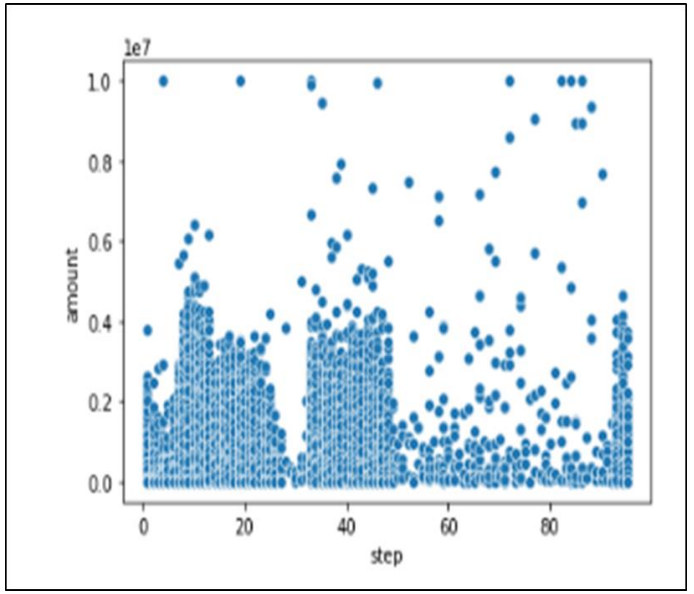
1. Exploratory Data Analysis
2. Feature Engineering
3. One-Hot Encoding
4. Model Selection
5. Model Training
6. Model Accuracy, Precision and Recall
7. Cross-Validation

EXPLORATORY DATA ANALYSIS

- Before modelling, it is important to explore and visualize the raw data to ensure that I am familiar with its contents so that I can derive as much insights as possible from it.
- For this project, the first thing I did was to look at the data by columns so I can understand the kind of data I am working with in terms of data types, data size, data shape, etc. Following this, I conducted some univariate, bivariate and multivariate analysis to see what the relationships between columns are and how useful this might be to make sense of the important features that would come later.



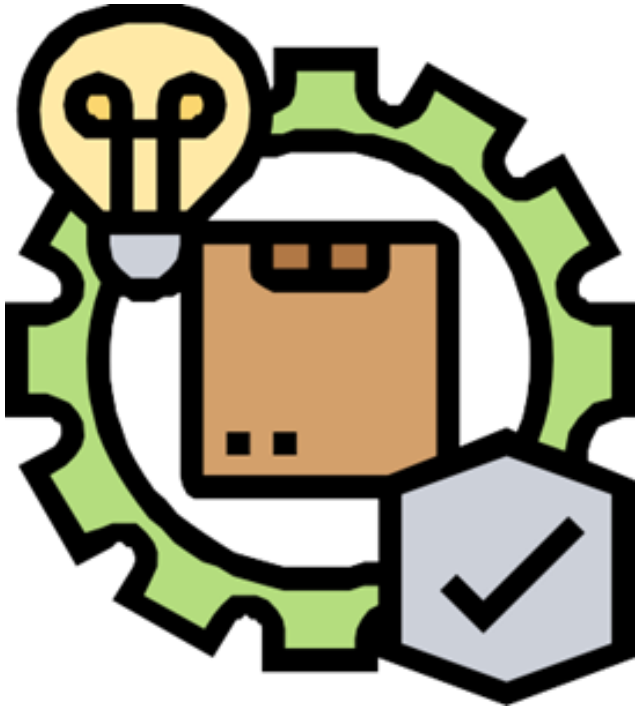




THE MODEL

PREPARING:

To prepare the data for modelling, I feature engineered the column y so that its contents are integers and not strings. This is important as most models only work with numerical values. Added to this, I also one-hot encoded other categorical columns by putting the corresponding columns in a list and then using the dummy feature on pandas to convert the contents in those columns to integers (0 or 1).



SELECTING & TRAINING:

Here, I created the code that will train and test four models from which only one model will be chosen based on the score. I use a list and a loop for the list to test each model in the list with the train-test code. In this code, I set y to be our target variable which is coincidentally named ‘y’ in the data and I set x to be all other columns as they are the independent variables which y is dependent on. Our test size is set to 40%.

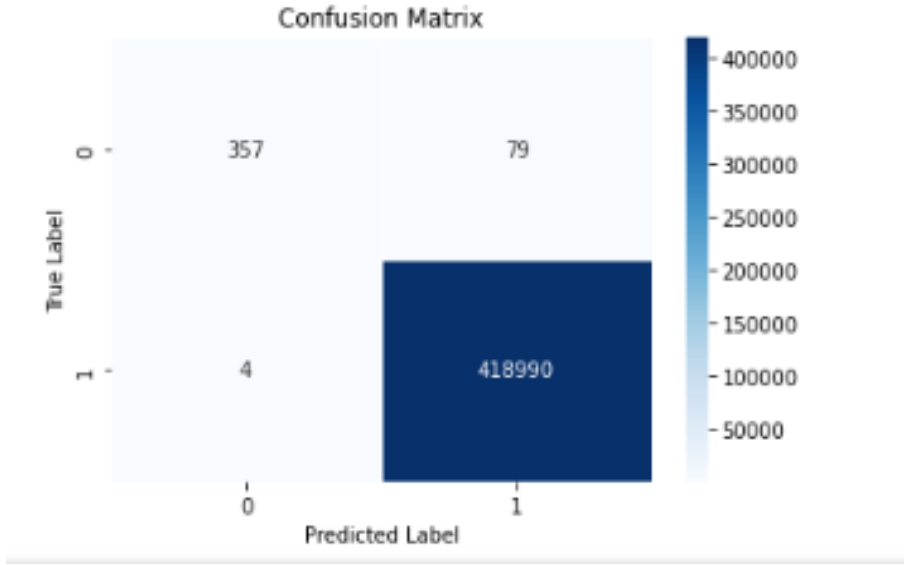
With the highest accuracy score, the RandomForestClassifier (RF) was chosen as the desired model.

MODEL INTERPRETATION

- In a bid to establish which model would be the most effective in detecting the fraudulent transactions, we consider their accuracy, precision and recall.
- In terms of accuracy, the RandomForest Classifier is at 99.98% which means that it would have 418,990 of the sample prediction correct. It was the model with the highest accuracy.
- In terms of precision, the RandomForest Classifier was also the model with the highest score with 99%.
- In terms of recall, the RandomForest Classifier was again the model with the highest score with 82%.
- Because of the consistency in being the model with the highest accuracy, precision and recall scores, the RandomForest Classifier is the most suitable model for dedecting fraudulent transactions.

For RandomForestClassifier, Accuracy score is 0.9998021123906253

	precision	recall	f1-score	support
Fraud	0.99	0.82	0.90	436
Not Fraud	1.00	1.00	1.00	418994
accuracy			1.00	419430
macro avg	0.99	0.91	0.95	419430
weighted avg	1.00	1.00	1.00	419430



CROSS VALID EVALUATION

- The cross-validation shows how well my model can generalize to new data by testing multiple trainings and tests. Using 10 splits and a function called ‘trainer_mcv’ which details the scoring mechanism for each model to loop through. With a score of 97% the results showed that Random Forest remains the most accurate model and that it can generalize to new data

CONCLUSION

1

Because the RandomForest Classifier has the highest accuracy, it should be established within the bank through the engineering team. The next steps would be to deploy the model into the bank's algothrithm.

2

In a bid for conveniency the engineering team could automate the model. The dedection would be accurate because through the cross evaluation conducted it has been realized that the model would be able to accurately dedect fraudulent transactions when given new datasets.

3

The bank would heavily benefit from integrating this model because the risk of customer's accounts on fraudsters would be reduced thereby giving them more publicity as a safe heaven.

THANK YOU

JIBOKU OLUWATUNMISE OLANREWAJU


oluwatunmisejiboku@gmail.com


https://www.linkedin.com/in/oluwatunmise-jiboku-4979a0212//


https://github.com/Jiboku-Oluwatunmise/


+234 805 2954 338