ONLINE PAYMENTS FRAUD DETECTION

BY NWAAMAKA IDUWE

PROJECT DEFINITION

Also known as BB PLC, Blossom Bank is a multinational financial services group, that offers retail and investment banking, pension management, asset management and payments services, headquartered in London. Blossom Bank has recently detected some fraudulent transactions on its network and is looking to solve this problem using a machine learning algorithm.

Towards this, the aim of this project is to help Blossom bank identify relationships between the given data, find the best machine learning algorithm to help them detect fraudulent transactions on their network.

Once the model is completed, the end benefits will be; better security for customer fund, increased trust in the bank by the customers, confidence to expand services knowing well that fraud will not be a problem. This will in turn help profitability.

The project was carried out in the following order:

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

Data Cleaning & Exploratory data analysis (EDA)

Data Inspection:

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, missing data, etc.

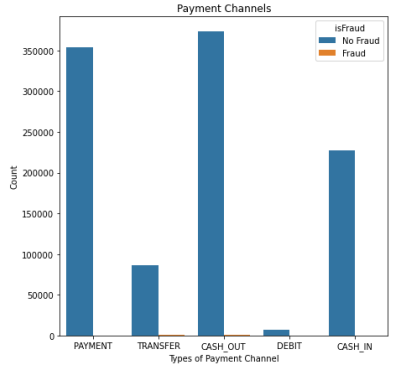
Data Cleaning:

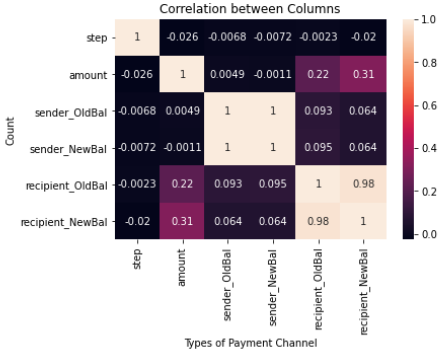
To make the data easy to understand, I changed the column names within. I also changed the contents in the target column from 0 , 1 to ‘No Fraud’ , ’Fraud’ respectively.

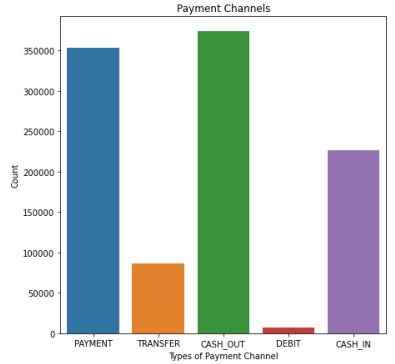
Exploratory Data Analysis:

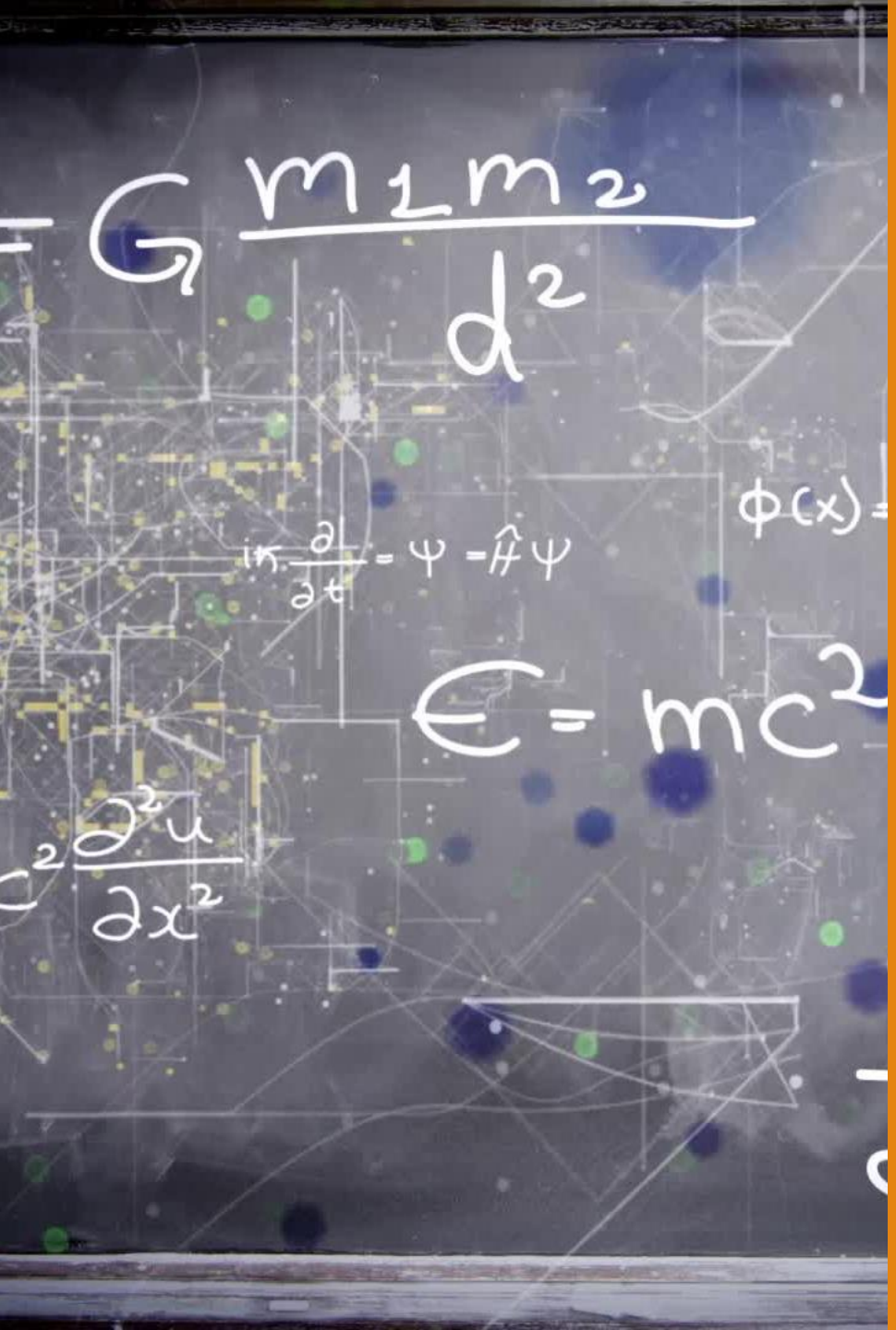
Before modelling, it is important to explore and visualize the raw data to ensure that I am familiar with its contents so that I can derive as much insights as possible from it.

For this project, 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.









Preparing, Selecting , Training and Testing the Model

PREPARING:

To prepare the data for modelling, I first deleted the 'sender', 'recipient' and 'class_amt' columns as they are irrelevant to the analysis. After this, I one-hot encoded the only remaining column with object dt 'type' so that its contents are integers and not strings. This is important as most models only work with numerical values. Using the dummy feature on panda, I encoded its contents to 0's and 1's.

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 used 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 the target variable which is named 'isFraud' in the data and I set X to be all other columns as they are the independent variables which y is dependent on. I set the test size to 45%.

I also created a code to return a confusion matrix for each model along with their accuracy, precision and recall.

By comparing the results using all three metrics –accuracy, precision and recall, the Random Forest Classifier was the best performing model.

Result Interpretation

The result gives us the accuracy, precision, recall and confusion matrix of each model.

Accuracy: this tells us how often our model is correct and we can see that the most accurate model is the Random Forest Classifier at 99.97% having 471,276 of the sample prediction correct. However, accuracy on its own is not a good enough means of evaluation.

Precision: this tells how many of the positive prediction really are positive i.e. how many detected items from the data are truly relevant. Again, the Random Forest has the highest precision score at 97%. Interestingly, Logistic Regression which had a fantastic accuracy score of 99.88% has a disappointing precision score of 46%. This is an example of why accuracy as a single metric is poor.

Recall: this tells us how well the model is able to predict positives i.e. how many of the relevant items from the data are detected by the model. It looks at both true positives and false negatives. This time, the Decision Tree Classifier had the highest recall score of 79% with the Random Forest Classifier close by with recall score of 78%. However, a 1% difference does not qualify Decision Tree as the better model because considering all three metrics, Random Forest is far ahead with an 11% difference in precision score and a slightly higher accuracy score.

Cross-validation Evaluation Using K-fold

In addition to these, I conducted a final test called cross validation using a tool called K-fold from the sklearn metrics library. The cross-validation test tells me how well my model can generalize to new data by testing multiple trainings and tests. I used 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 well.

Cross-Validation Result Interpretation

Like in the first set of results, the Random Forest Classifier has the highest accuracy, precision and second highest recall score, falling behind Decision Tree by 1%. These scores remain constant in both evaluations and this is a good sign of the reliability of our model. As mentioned before, a 1% difference does not qualify Decision Tree as the better model because considering all three metrics, Random Forest is far ahead with an 11% difference in precision score and a slightly higher accuracy score.

Recall would be the most important metric for choosing the right model because as our problem definition hints, we need to be as accurate as possible to ensure our customer funds are highly secured. This means our model must consider all relevant factors and this is what Recall shows us.

In terms of confusion matrix, both False negatives and True positives are important to consider as we aim for 100% accuracy wherever possible.

LIMITATIONS AND RECOMMENDATIONS

Limitations:

The main limitation to this project was the lack of relationship between the columns in the data. This made it slightly difficult to conduct Bivariate EDA on the data.

Recommendations:

If possible, Blossom bank can provide more data on the transactions under consideration to allow for a better analysis of the relationship between the data. This can also lead to new discoveries that will better help detect fraud in the data.

Further Steps For Blossom Bank

As described, the best performing model is the Random Forest Classifier. Towards solving the problem at hand, Blossom bank should deploy this model to the production team. The next steps would require using the model to find out the most important columns and features from the data the IT security team should focus on when detecting fraud.

THANK YOU