**Capstone Project**

**Fraud Detection in Financial Institute**

# Process overview

The following diagram shows the overall end-to-end process for defining, designing and delivering the Capstone project.



# 

Table of Contents

[Process overview 1](#_Toc84327491)

[1. Problem statement 3](#_Toc84327492)

[2. Industry 3](#_Toc84327493)

[3. Stakeholders 3](#_Toc84327494)

[4. Business question 3](#_Toc84327495)

[5. Data question 3](#_Toc84327496)

[6. Data 4](#_Toc84327497)

[6.1 Fields of dataset 4](#_Toc84327498)

[6.2 How was this data generated? 4](#_Toc84327499)

[7. Hardware Specification 5](#_Toc84327500)

[8. Data science process 5](#_Toc84327501)

[8.1 Data analysis 5](#_Toc84327502)

[8.1.1 Distribution of fraudulent transactions 5](#_Toc84327503)

[8.1.2 EDA Findings 5](#_Toc84327504)

[8.2 Modelling 6](#_Toc84327505)

[8.2.1 Feature Engineering 6](#_Toc84327506)

[8.2.2 Features used in modelling 6](#_Toc84327507)

[8.2.3 Models used 6](#_Toc84327508)

[8.2.4 Hyperparameter Tuning 6](#_Toc84327509)

[8.2.5 Training duration 6](#_Toc84327510)

[8.2.6 Model Performance Metrics 6](#_Toc84327511)

[8.3 Outcomes 7](#_Toc84327512)

[8.4 Implementation 8](#_Toc84327513)

[9. Data answer 8](#_Toc84327514)

[10. Business answer 8](#_Toc84327515)

[11. Response to stakeholders 8](#_Toc84327516)

[12. End-to-end solution 8](#_Toc84327517)

[13. References 9](#_Toc84327518)

# Problem statement

Fraud has quickly become a multi-billion-dollar problem for the banking and finance industry. A report by The Association of Certified Fraud Examiners (ACFE, 2018) suggest that banks are losing up to 5% of their revenue to fraud alone. The Guardian (2021) reported that in UK alone, £754m were stolen from bank customers during the first half of 2021 – a 30% rise on the same period in 2020.

It is essential to be able to identify fraudulent transactions and be able to block them in real time as these activities would not only the impact the assets of consumers negatively, but also those of the organizations and their brand reputation.

# Industry

* Banking and Finance Industry

# Stakeholders

* Management of Financial Institute
* Government Authority

# Business question

* Banks are losing money due to frauds and their reputation is on the line, can fraud be detected and prevented to reduce loss in revenue

# Data question

* How can we identify fraudulent transactions using the dataset?
* Using various machine learning algorithms to find the one with the highest accuracy for fraud prediction, and more importantly can fraudulent transaction be predicted?

# Data

* Source of data: [Kaggle](https://www.kaggle.com/ealaxi/paysim1)
* 6,362,620 entries, 11 columns
* No null or empty values
* No duplicated records

## Fields of dataset

|  |  |
| --- | --- |
| **Field Name** | **Field Description** |
| step | Maps a unit of time in the real world. In this case 1 step is 1 hour of time. Total steps 744 (30 days simulation). |
| type | Type of transaction - CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER |
| amount | Amount of the transaction in local currency |
| nameOrig  (name\_origin) | Customer who started the transaction |
| oldbalanceOrg  (old\_balance\_origin) | Initial balance before the transaction |
| newbalanceOrig  (new\_balance\_origin) | New balance after the transaction |
| nameDest  (name\_destination) | Customer who is the recipient of the transaction |
| oldbalanceDest  (old\_balance\_destination) | Initial balance recipient before the transaction. Note that there is not information for customers that start with M (Merchants) |
| newbalanceDest  (new\_balance\_destination) | New balance recipient after the transaction. Note that there is not information for customers that start with M (Merchants) |
| isFraud | This is the transactions made by the fraudulent agents inside the simulation. In this specific dataset the fraudulent behavior of the agents aims to profit by taking control or customers accounts and try to empty the funds by transferring to another account and then cashing out of the system. |
| isFlaggedFraud | The business model aims to control massive transfers from one account to another and flags illegal attempts. An illegal attempt in this dataset is an attempt to transfer more than 200,000 in a single transaction. |

## How was this data generated?

Data is generated by running PaySim several times using random seeds for 744 steps, representing each hour of one month of real time, which matches the original logs. Each run took around 45 minutes on an i7 intel processor with 16GB of RAM. The final result of a run contains approximately 24 million of financial records divided into the 5 types of categories: CASH-IN, CASH-OUT, DEBIT, PAYMENT and TRANSFER. This dataset only contains one quarter of the original dataset.

# Hardware Specification

* Windows OS (64 bit)
* 16 GB RAM
* Processor: AMD Ryzen 7 5800H (8cores / 16 threads)

# Data science process

## Data analysis

* The dataset is without duplicated records or null values

### Distribution of fraudulent transactions

* 6,354,407 valid transactions
* 8213 fraudulent transactions
* Dataset is highly imbalanced

### EDA Findings

* Fraudulent transactions are only found in Transfer and Cash-Out transactions. For this project, we will be focusing on these two types of transactions.
* isFlaggedFraud field is inaccurate in marking fraudulent transactions that is of transfer type and with more than $200,000. Over 2700 fraudulent transactions are not marked.
* Checks on whether there are any same accounts used for Transfer (Destination) and Cash-Out (Origin) transactions were conducted but no records are found.
* Step field has a fair distribution against fraudulent transaction
* 85% of the transactions have erroneous account balance in origin account, the amount in the transactions with the old\_balance\_origin and new\_balance\_origin does not tally
* 74.5% of the transactions have erroneous account balance in destination account, the amount in the transactions with the old\_balance\_destination and new\_balance\_destination does not tally
* 93.6% of the transactions have erroneous account balance in either origin or destination

## Modelling

### Feature Engineering

* After finding out on the erroneous account balance in both origin and destination accounts, two new features are added (balance\_origin\_error, balance\_destination\_error)
* balance\_origin\_error = amount + new\_balance\_origin - old\_balance\_origin
* balance\_destination\_error = amount + old\_balance\_destination - new\_balance\_destination

### Features used in modelling

* step, amount, old\_balance\_origin, new\_balance\_origin,

old\_balance\_destination, new\_balance\_destination,

balance\_origin\_error, balance\_destination\_error, type\_CASH\_OUT,

type\_TRANSFER

### Models used

* Decision Tree Classifier
* Random Forest Classifier
* Extra Trees Classifier
* Extreme Gradient Boosting Classifier
* Gaussian NB

### Hyperparameter Tuning

* An user defined function was created: hyperparameter\_tuning
* With the use of pipeline, first the data was scaled (StandardScaler), then fit into the respective Machine Learning Algorithms
* Using GridSearchCV to find the best hyperparameters

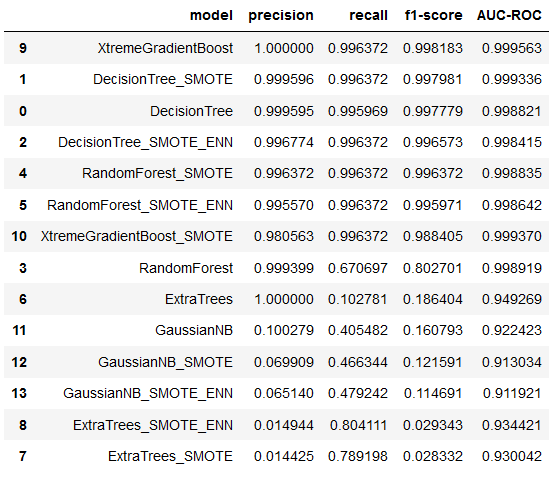
### Training duration

* Training duration varied between models, the range of timing span across 15 mins to 2 hours

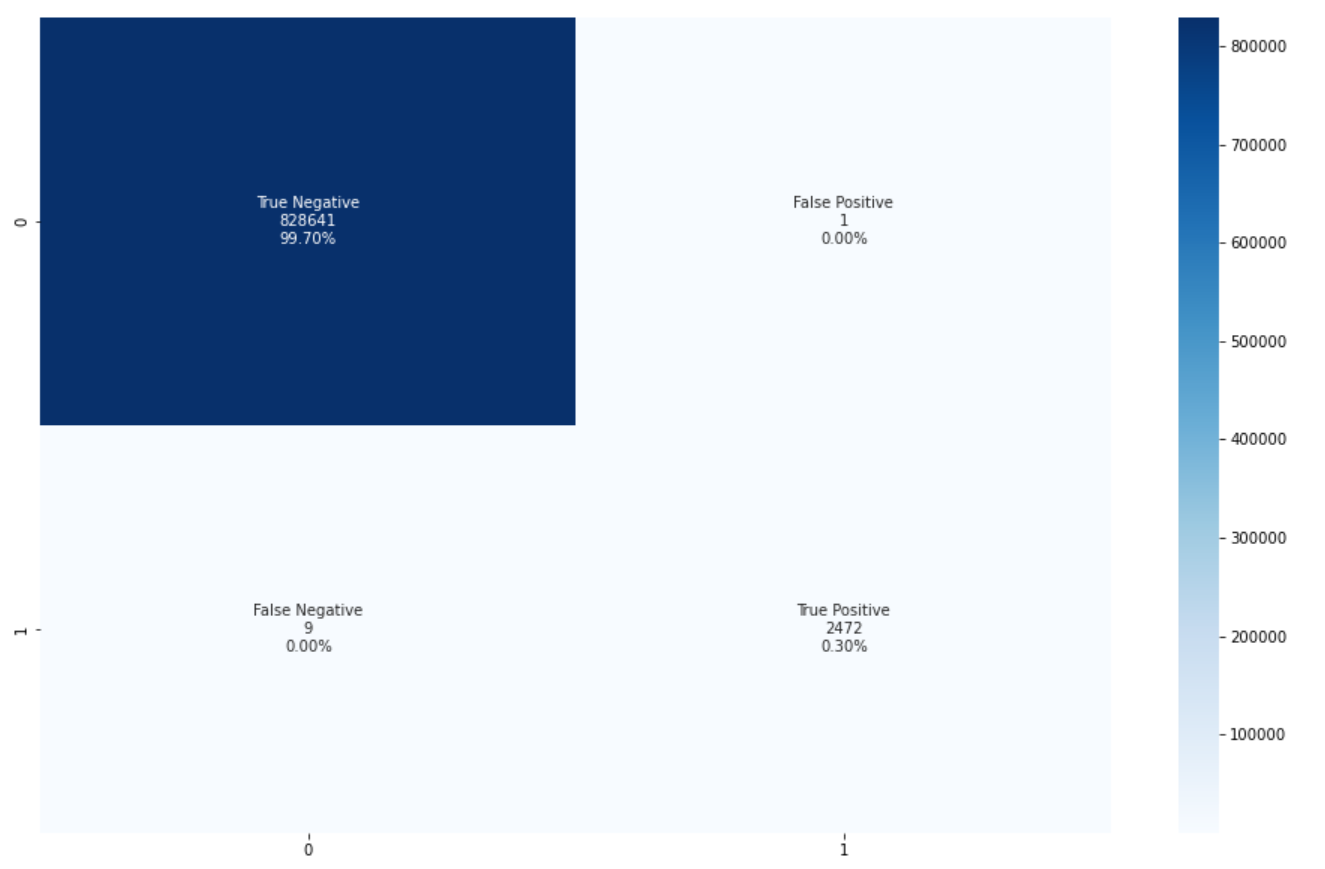
### Model Performance Metrics

* Recall
* Precision
* F1-score
* AUC-ROC
* Confusion Matrix

## 8.3 Outcomes



DecisionTree classifier trained with SMOTE resampled data performed the best in predicting fraudulent transactions.



The model predicted 9 transactions to be valid however they are fraudulent transactions. This is the key misclassification figure that we are interested in as we aimed to reduce the amount of loss to a financial institute through frauds. And this model performed the best in terms of lower number of misclassifications as well.

## 8.4 Implementation

The model might not be suitable for deploying into a real-world scenario at the moment. As the dataset used was synthetically generated, it will likely have different fields and features compared to the real-world data. In addition, there are errors in the account balances of the dataset used. However, we could apply the methodologies used in this project for applications with real-world banking transactions.

# Data answer

* Using 5 different models, we are able to identify the best model for prediction fraudulent transaction.
* We are able to predict fraudulent transactions over 90% of the time with the dataset
* Machine Learning algorithms are able to accurately predict fraudulent transaction for over 90% of the time

# Business answer

* Fraud can be detected using Machine Learning algorithms and therefore loss in revenue can be reduced.

# Response to stakeholders

* In order to identify patterns and complex techniques used by criminals, it is essential to use machine learning algorithms to aid us.

# End-to-end solution

* The real-world transactional data should be passed to the model for continuous training as criminal techniques are evolving by day.
* Model to be deployed to it should be able to predict the validity of the transaction in real time.
* Once predicted, the transaction should be blocked and an email or notification should be sent the banking staff for further investigation.

# References

* Where are the data and code used in the project? (show a simplified list of main items: notebooks, datasets, exported models)
* Resources used
  + Pandas
  + Numpy
  + Matplotlib
  + Seaborn
  + Sklearn
  + Imblearn
  + Xgboost
  + Collections
  + Dill
* https://www.theguardian.com/money/2021/sep/22/fraud-in-uk-at-level-poses-national-security-threat-bank-customers-covid