Data Analysis Techniques and Tools

Business Analytics

Presented by Pharaoh Kipkirui Chirchir

18 Jan 2024

Definition of the Business Problem

- The primary objective of this data analytics techniques is to identify and prevent potentially fraudulent transactions at ABSA Bank.
- This will help in Identifying and preventing potential financial losses, reputation damage, trust and regulatory consequences associated with fraudulent activities.
- This will lead to retention of high profitable customers over time.
- Fraud detection using machine learning is a necessity for ABSA Bank to put in place a proactive monitoring and prevention mechanisms.
- This will help ABSA Bank to reduce time consuming manual reviews as well as denial of legitimate transactions.
- We will analyze credit card data, downloaded from Kaggle.

Tools for data Analysis

Python: Utilize popular libraries such as scikit-learn, TensorFlow, and PyTorch for machine learning and data analysis.

R: It can be used for statistical analysis and visualization.

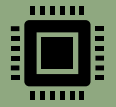
SQL: Employ SQL for database for data mining and joining of different datasets.

Data Visualization Tools: Use tools like Tableau and Power BI to create interactive visualizations for better insights and communication of the results.

Excel – This will be used to perform simple analysis of the dataset

Identifying the data required

Absa bank will need to provide customer transaction data. The dataset will need to include day to day transactions made by the customers, amounts of transactions made, time when transactions are made, among other data variables.

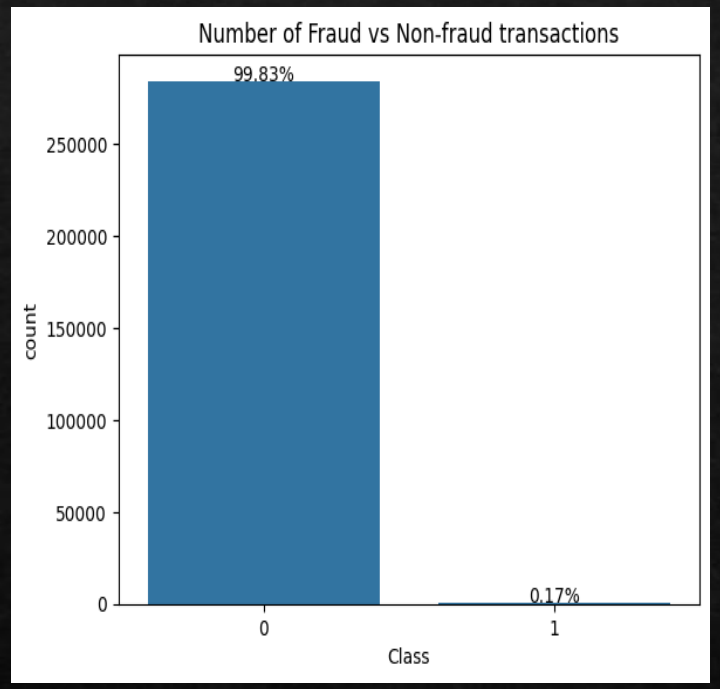


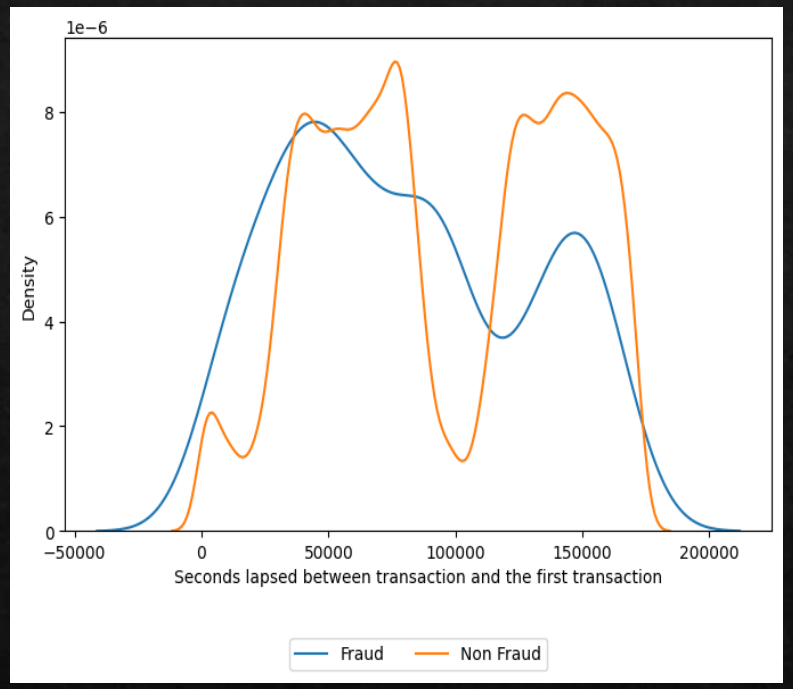
The data provided should also contains the fraudulent transactions and non-fraudulent transactions

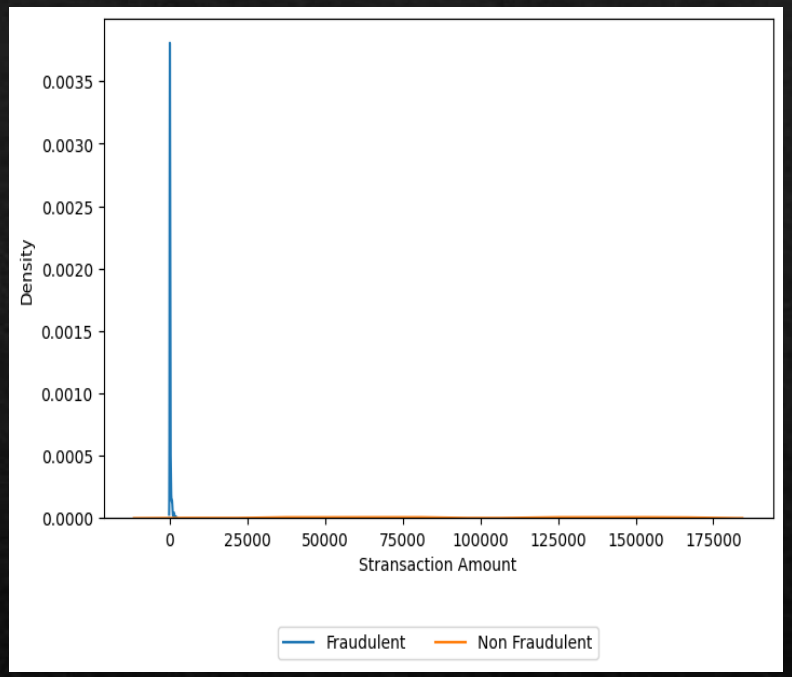
Customers data is highly unbalanced, and we will require measure the accuracy using the Area Under the Precision-Recall Curve using different models under Jupyter libraries.

For the real-life scenario, we will analyze the Credit Fraud data obtained from Kaggle to establish the best model we will use to detect Fraud which can be employed by ABSA Bank

Real life Scenario: Credit card Fraud detection



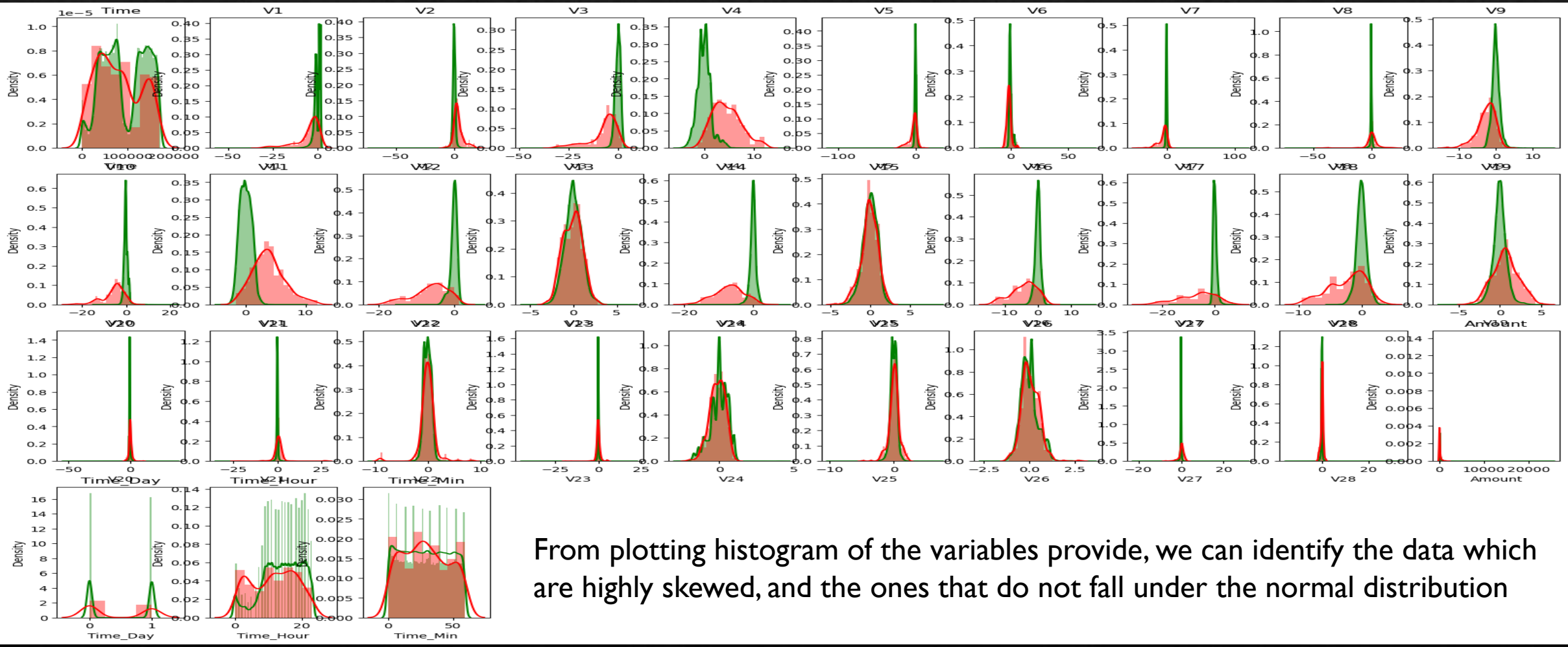




- The dataset is highly imbalanced with 284315 non fraud cases and 492 fraud cases
- Distribution of time column is not a clear indicator of fraud transactions in the dataset
- Fraud transactions are mostly denser on the lower range amount, whereas the non-fraud transactions are spread out through low and high range amounts

Real life Scenario: Credit card Fraud detection

- Exploratory Data Analysis: Histogram showing the distribution of datasets



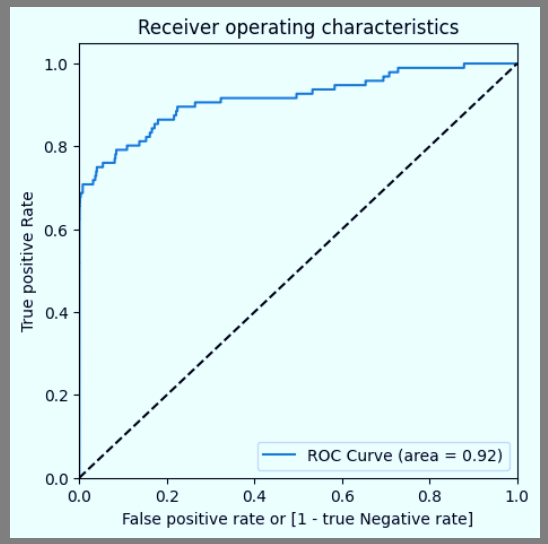
From plotting histogram of the variables provide, we can identify the data which are highly skewed, and the ones that do not fall under the normal distribution

Data Analysis techniques

- Anomaly Detection: Statistical methods and machine learning algorithms are used to identify unusual patterns or outliers in the transaction datasets.
- Predictive Modeling: Develop models that predict the likelihood of a transaction being fraudulent based on historical patterns and relevant features.
- Regression Analysis: This is modelling relationships between dependent and more than one independent variables. This includes the linear regression and logistic regression.
- Descriptive Statistics: This is the summary of the basic features of the data sets to understand the shape of data, null values, measures such as mean, mode median among others
- Machine Learning: This are used for detecting patterns in the datasets example fraud patterns and used in decision making.

Fraud detection using different models

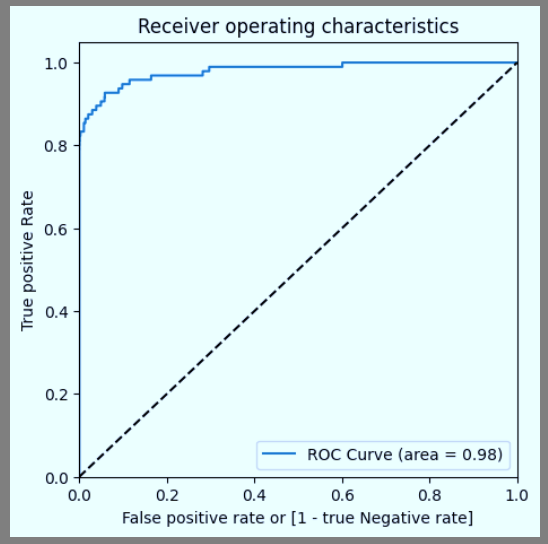
Logistic regression



-------------Classification Report-------------

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56866
1	0.67	0.62	0.65	96
accuracy			1.00	56962
macro avg	0.84	0.81	0.82	56962
weighted avg	1.00	1.00	1.00	56962

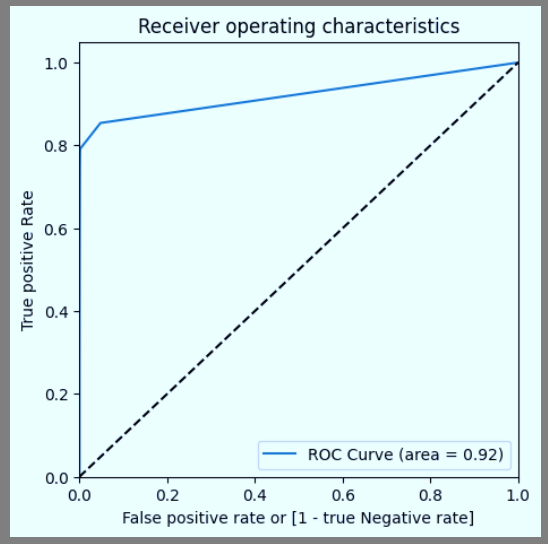
XG Boost



-------------Classification Report-------------

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56866
1	0.91	0.74	0.82	96
accuracy			1.00	56962
macro avg	0.95	0.87	0.91	56962
weighted avg	1.00	1.00	1.00	56962

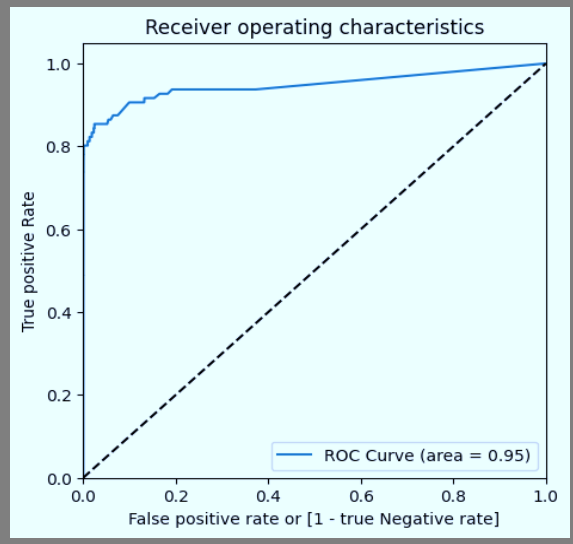
Decision tree



-------------Classification Report-------------

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56866
1	0.65	0.58	0.62	96
accuracy			1.00	56962
macro avg	0.83	0.79	0.81	56962
weighted avg	1.00	1.00	1.00	56962

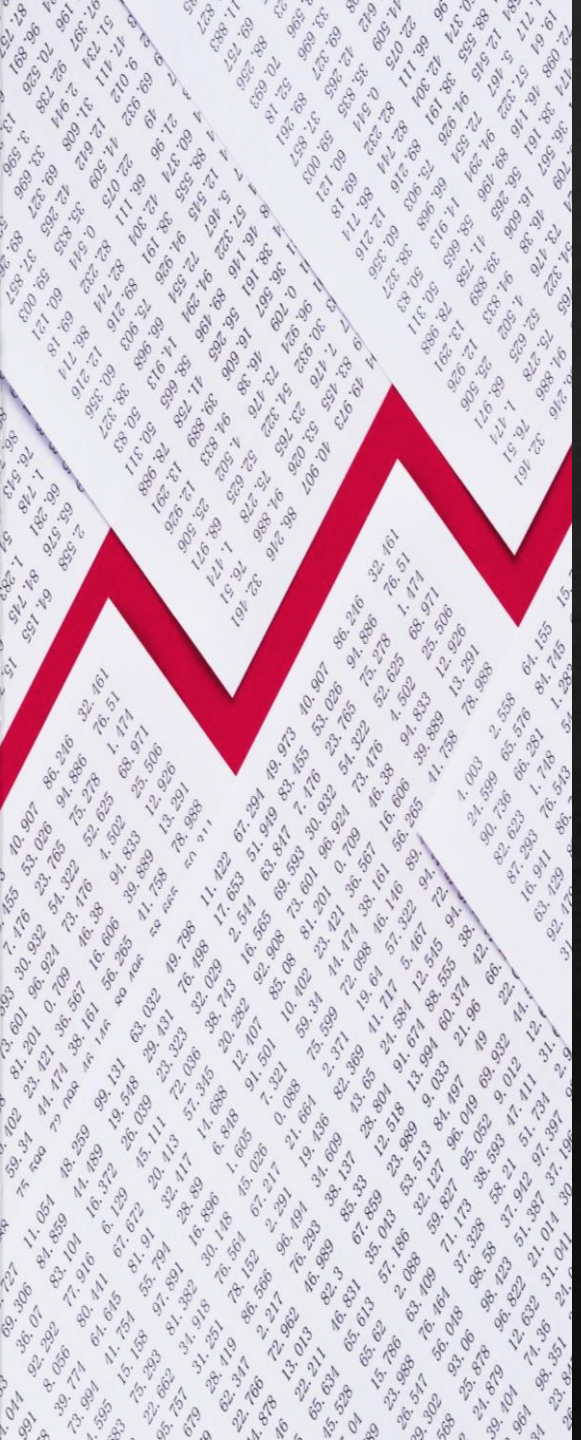
Random Forest



-------------Classification Report-------------

	precision	recall	f1-score	support
0	1.00	1.00	1.00	56866
1	0.74	0.62	0.68	96
accuracy			1.00	56962
macro avg	0.87	0.81	0.84	56962
weighted avg	1.00	1.00	1.00	56962

- From the models above, based on ROC, XG boost is performing well with a macro avg of 91% followed by Random Forest. It has also a higher f1 score of 82%



Summary from the analysis of the Credit Card Fraud data

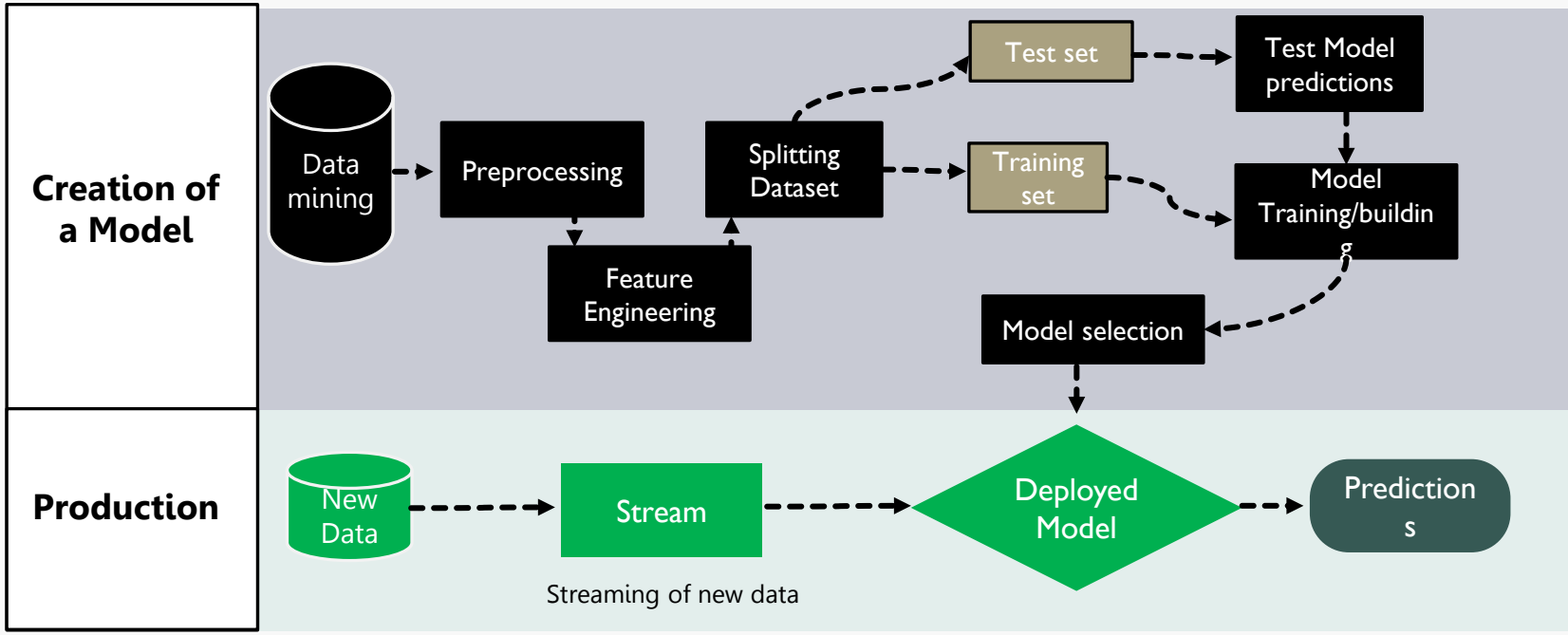
- Decision can be made based on accuracy metrics and the best model can be selected when it comes to fraud detection.
- Descriptive statistics was used to understand the basic structure of the data set and the variables that required to be normalized and standardized, so that it gives a uniform and consistent results.
- From the credit card system XG Boost is performing better among other models and can be selected to detect Fraud.
- The data was imbalanced, and it is important to always balance the dataset, which for our scenario we used the under-sampling technique to balance the dataset to get better results
- Machine learning was employed using supervised learning technique, to perform predictive modelling in detecting fraudulent transactions

Steps involved in developing Fraud detection model

Two step model used in detection of Fraud.

Step 1:
Creation of a model

Step 2:
Production of a model



Steps involved in developing Fraud detection model

Model Creation

- Data Mining – Gather relevant data from different sources, including customer information and transaction records.
- Preprocessing- Perform Exploratory data analysis to understand data, including handling of missing values, outliers and scaling numerical features
- Feature Engineering – Create and generate features and new variables for better understanding of the data.
- Splitting of Data – Split datasets into training and testing sets
- Test Model Prediction – Experiment with different machine learning algorithms to identify the best performing in fraud detection.
- Building/Training of model – Use the dataset to teach a model to make predictions based on patterns it learn during the test process.
- Selection of the best Model – this involves identifying and using the best model with a high accuracy based on the score metrics chosen

Model Creation

Production

- **Deployed Model –** The best performing model is chosen and is used to make prediction of a transaction on a likelihood to be a fraudulent transaction
- **Data Streaming –** This is the real time processing and transmission of customers credit card data.
- **Predictions –** This involves using the trained model to make prediction based on transaction data being fraud

Improving ABSA Fraud Prevention Techniques

Analyze the results of the model to identify weaknesses of the model and the fraud types that are not being detected effectively

Constantly review false positives and false negative in the model to understand the reasons behind the errors and access the impact to the institution

Evaluate models performance metrics using f1 score and ROC curve.

Explore additional feature engineering modifications to existing features in the model

Conduct training to employees to increase fraud awareness and new trends of fraud as discovered by the model.

Conduct regular audits and testing of the fraud prevention measures to ensure their effectiveness.

End.

Thank you.