CREDIT CARD FRAUD DETECTION - MACHINE LEARNING

GROUP 33

Contents

Table of Figures

Abstract

These days, there is a rapid growth in electronic transaction and many people use online payment methods, one of the most common payment methods being credit card. It is approximated to be more than half of the transaction types. With emerging technologies, there is an increase in fraudulent activity using credit cards. Credit card fraud is a significant issue affecting financial institutions and cardholders globally. In this project, we present an advanced Credit Card Fraud Detection System that make use of machine learning and data analysis techniques to accurately identify and predict if a transaction is legitimate or fraud.

The project begins by using a publicly available dataset containing two days of credit card transactions. The dataset is highly imbalanced, with mostly legitimate transactions and a small fraction of fraudulent ones. This imbalance presents a challenge in building models that can effectively distinguish between the two classes while minimizing false positives. Modelling previous credit card transactions using information on those that turned out to be fraudulent or not is part of the Credit Card Fraud Detection Problem. The goal of this project is to minimize inaccurate fraud categories while detecting 100% of the fraudulent transactions.

Introduction

Credit cards are now the most preferred way for customers to transact either offline or online(Saurabh Bagchi, 2023). One downside that has been witnessed over the past few years of this increasing digital phenomenon is the rise of fraud on the credit card. “Fraud is an uncommon, well considered, imperceptibly concealed, time evolving and often carefully organized crime which appears in many types and forms (Van Vlasselaer, Baesens, 2013). In the UK, there were 32.3 million people with credit cards or charge cards in 2016, which roughly translates to 6 in every ten adults. The numbers have only grown since then from 2016 to now.

It is important for credit card companies to be able to recognize if the credit card transactions are fraudulent or normal so that they will not charge customers for items which they did not purchase. The main aim of this project is to detect fraudulent transactions with credit card details. Credit card fraud refers to using stolen credit card or someone else’s credit card without their consent for financial transactions.

(Hand, Manilla and Smyth, 2001)As a first step, the source data must be identified that could be of potential interest. Some basic exploratory analysis can be considered here. This will be followed by a data cleaning step to get rid of all inconsistencies. In the analytical step, an analytical model will be estimated on the preprocessed data. Once the model has been built, it will be interpreted and evaluated by the fraud experts (Van Vlasselaer, Baesens, 2013).

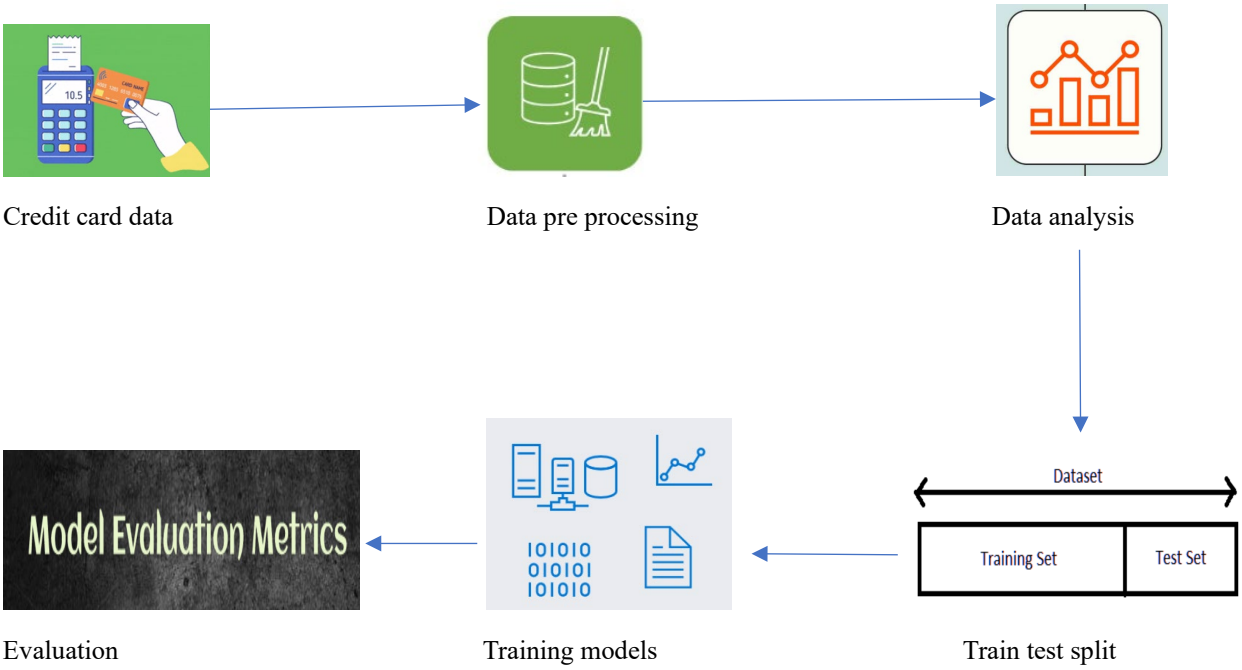

Table 1:Workflow

Dataset

This Kaggle dataset contains credit card transactions by card holders (https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data). Due to privacy reasons, the features are masked and presented after Principal Component analysis. This consists of data from transactions over a period fo 2 days. The goal is to distinguish between legitimate and fraudulent transactions. The key features are time in seconds elapsed between each transaction and the first one in the dataset. We also have transaction amount in Amount column. The columns 'Class' represents whether a transaction is fraudulent or not. The value of 0 indicates legitimate transaction and the value of 1 indicates a fraudulent transaction. V1,V2,etc are features obtained through PCA and kept anonymous for privacy of cardholders.

There are a total of 2,84,807 transactions out of which only 492 are fradulent. This makes this dataset highly imbalanced and skewed.

Key Challenges

Credit card fraud detection using machine learning poses challenges in terms of requirement on very high accuracy and the need for low false positives. In most datasets, including the one discussed here, the number of legitimate transactions are significantl much higher than the fraudulent transactions. Due to this class imbalance, trained models will have a bias towards majority class which is the legitimate class. This makes it difficult to identify the fraudulent transactions effectively.

While working with transaction data, it is essential to protect the privacy of card hardholders. Anonymization techniques are often engaged to make sure the personal and sensitive details of the cardholders are not exposed.

Exploratory Data Analysis(EDA)

In exploratory data analysis, the dataset is analysed to understand the main statistical characteristics with visual and statistical methods. EDA is basically the first look at the data. During this process, we look for understanding patterns within the data and the relationships betweeen the features. This will also look outliers within the dataset. For data preprocessing null values should be eliminated. There are no null values in this dataset. The bar chart is plotted to find out the count for each classes. There are more than 250000 legit transactions and a very less number of fraudulent transactions out of the total transactions. It is very clear from this chart that the dataset is highly imbalanced.

Analysis on Class

The fraud and legit transactions are stored separately and the amount column is statistically analysed. It was found that the mean value of amount of all the legit transactions is around 88 dollars. 25% of the legit transaction amount is less than 5 dollars. Similarly, the mean transaction amount value for fraud transactions is about 122 dollars which is quite high when compared to legit transactions.

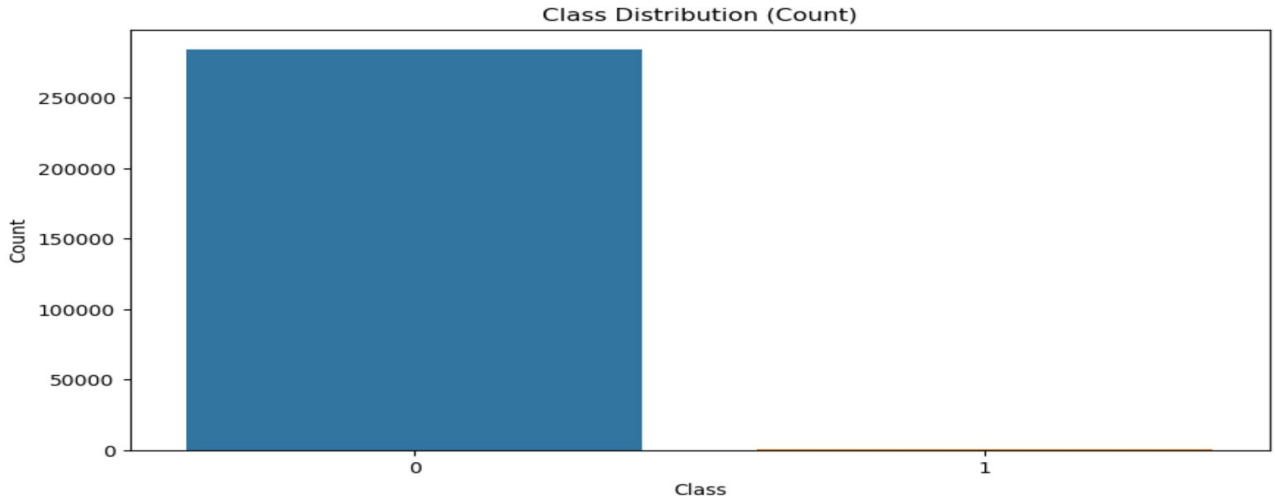

Table 2 : Class count

Analysis on Amount

The number of transactions with respect to amount was very small for fraudulent cases when compared to the legit cases.

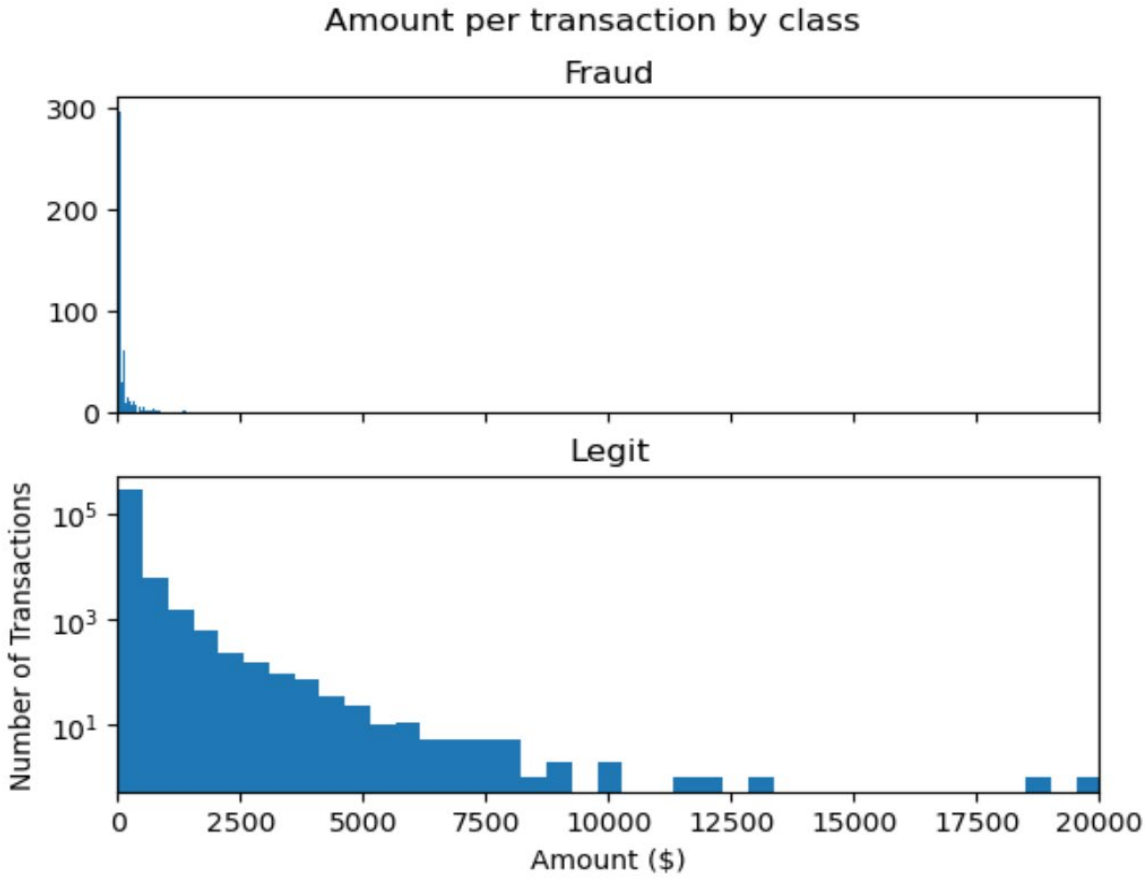

Table 3: Distribution of Amount

Analysis on Time

The below plot is done to analyse the different transactions for fraud and normal in terms of Time. This will show us if fraudulent transactions occur more often than legitimate transactions during certain timeframes.

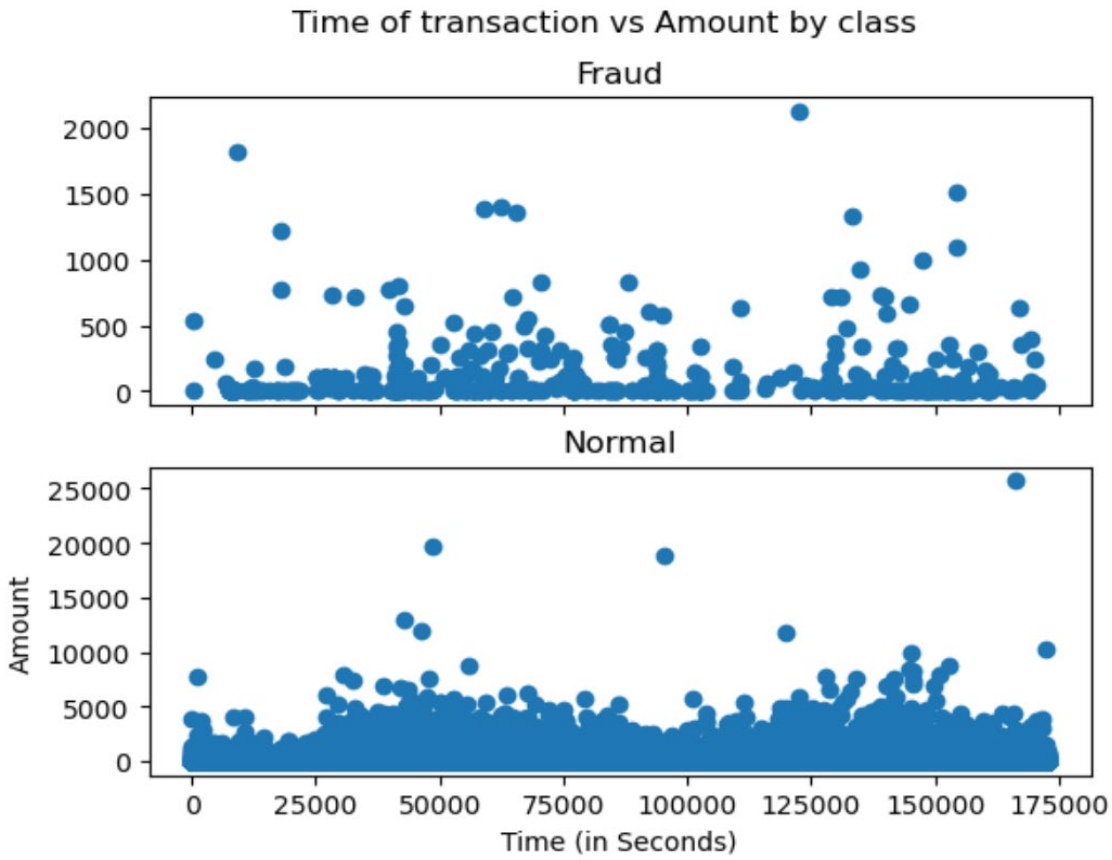

Table 4:Time vs Amount

Scatter plot on the features

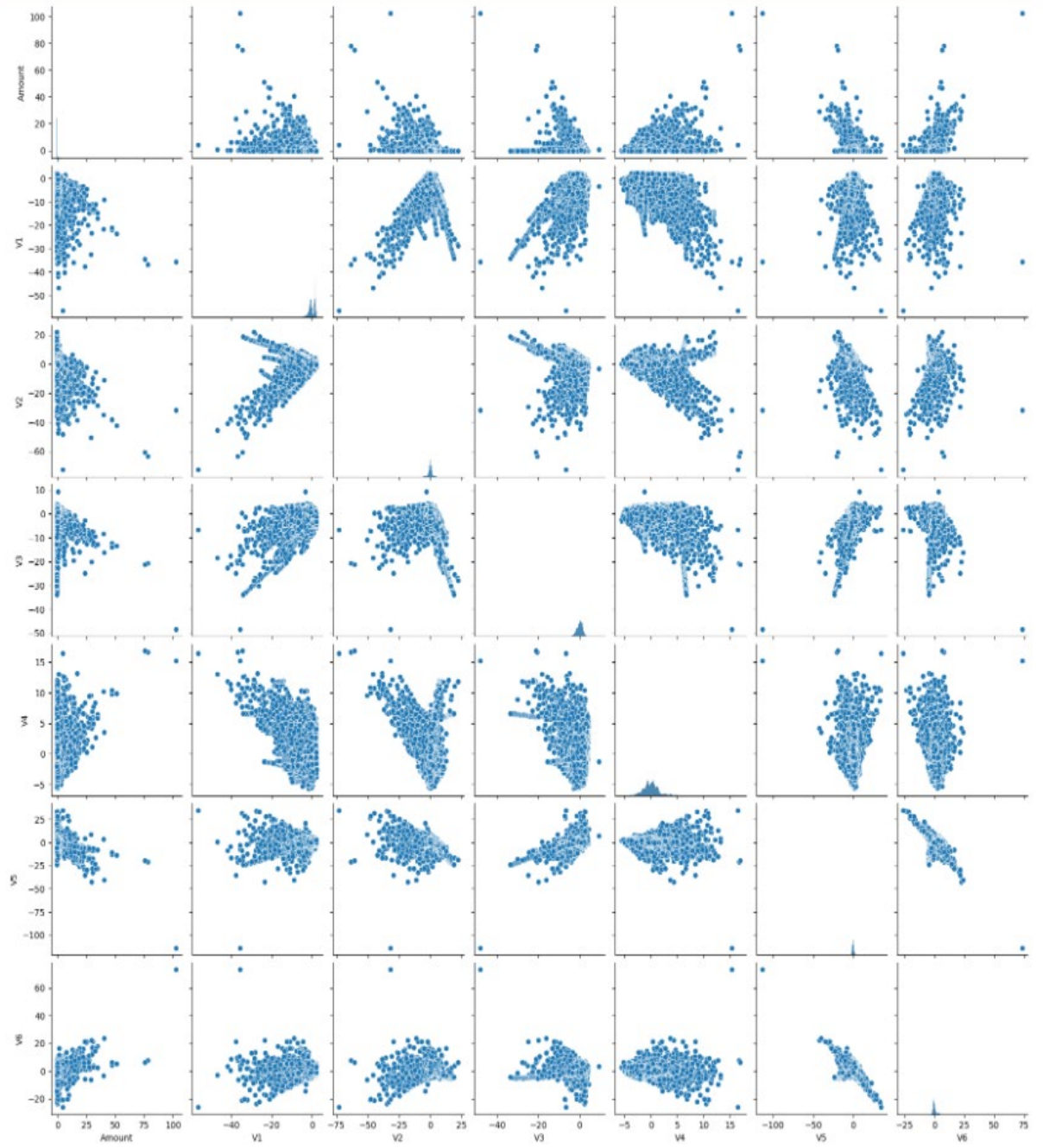

Table 5:Scatter plot

Correlation Matrix

The matrix is showing how all the features are with respect to the Class variable. Heatmap will help understand the data by visualizing it in a better way.

The values around 1.0 are indicated by blue and show a strong positive correlation between the variables. The negative correlations are represented by a deep red.

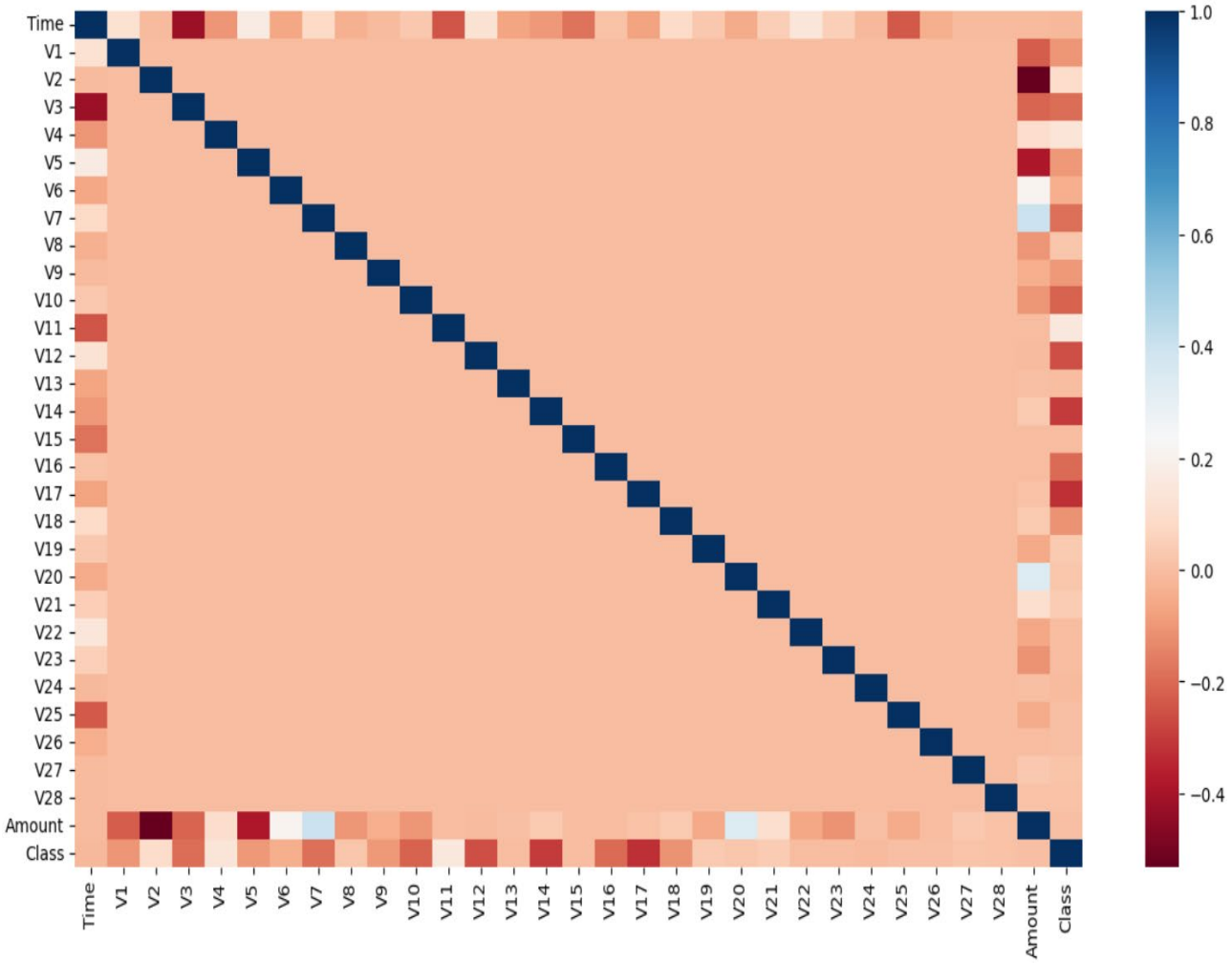

Table 6:Heat Map

Data preprocessing

Data pre processing is a pre requisite for modelling. One of the first steps in data preprocesign s to remove the missing data. Since in our dataset, there are no null values,

this step is skipped. Pre processing is one challenging step as the dataset is very unbalanced. When predicting a transaction as fraud or legit, the classifier typically favors the majority class. It will be biased to label every transaction as a legitimate transaction. Classifiers learn better from a more balanced distribution (Bart Baesens,2013)

Undersampling is used here to build a sample dataset from the original dataset which will then contain a similar distribution of legit and fraudulent transactions. 492 legit samples are chosen from the dataset and concatenated with the 492 fraudulent transactions. This gives a new dataset which has a balanced data. This puts only the relevant data in the model.

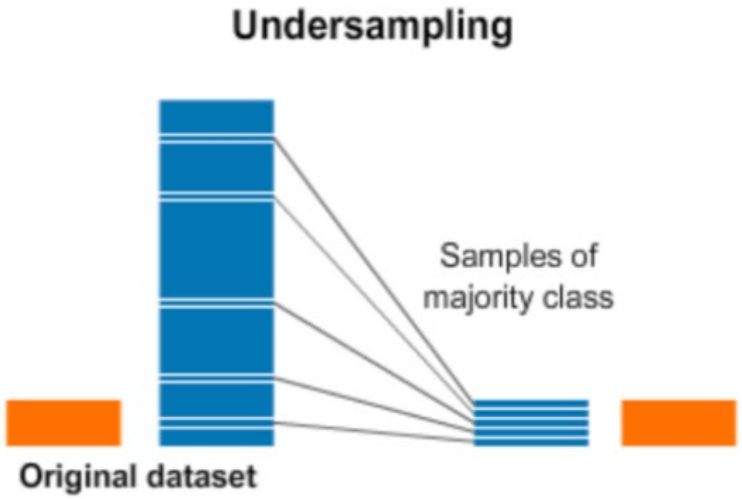

Table 7:Undersampling

The mean values of all the columns are compared with the original dataset and the new balanced dataset. It is almost similar which infers that the nature of the new dataset has not changed and we have obtained a good sample.

Train Test Split

A machine learning's performance is highly dependent on the dataset that it is being trained on. The dataset is split into training and testing data. 80% of data is taken as Training data and 20% is taken as Testing data. The features are present in X. This is obtained after dropping the column X. The labels are present in Y which has the corresponding values of Class variable which denotes 0 for legit and 1 for fraud transactions. X and Y is splitted into training and testing data. All the features of training data and all the corresponding labels are stored in X_train and Y_train respectively. Similarly the features of testing data

will be stored in X_test and the corresponding labels are stored in Y_test. This training data will be fed to the machine learning model.The test data is used to evaluate the accuracy of the model.

Training Model

The training data is used to train the model. This model will then predict if a particular transaction is legit or fraudulent. Generally for binary classfication problems, Logistic Regression model is used.

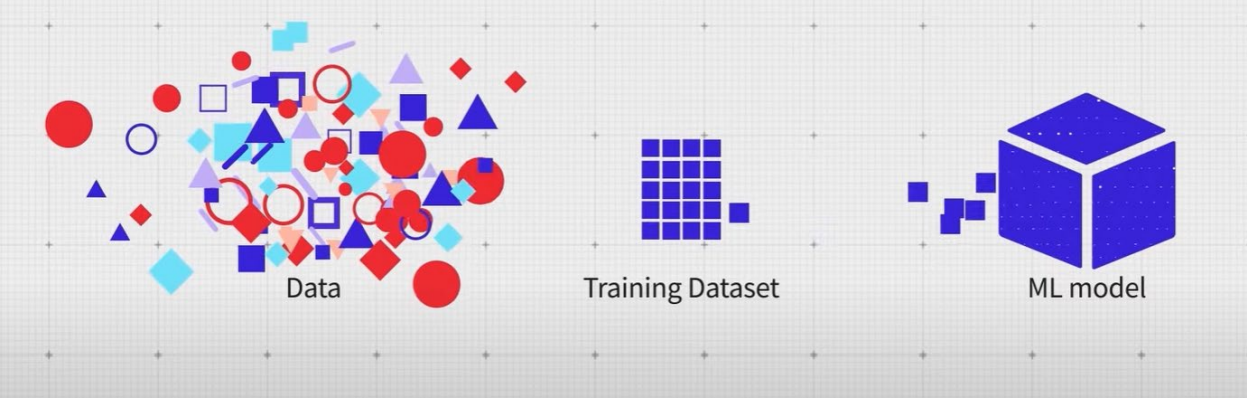

Table 8:Training the dataset

Evaluation

The final step is to evaluate the model. Once training is done, we use the testing data to evaluate how well the model does. This is the data set that includes the datapoints that the model has never seen. The perfromance of the model on the test data shows us how much we were able to create a generalized model that can represent the real world.

Table 9:Data split

Supervised - Logistic Regression

(Laura, 2017)Supervised learning algorithms learn from a training set of labeled examples (exemplars) to generalize to the set of all possible inputs. A regression analysis is used for modeling relationships between variables. Logistic regression predicts whether something is true or false instead of predicting something continuous. It makes it possible to infer or predict a variable. The goal of logistic regression is to estimate the probability of occurrence. The value range for the prediction should be between 0 and 1. Accuracy on training data is 92.38% from Logistics regression model and accuracy on testing data is 91.37%

Accuracy and performance

The training accuracy indicates how well the model performed on the data it was trained on. We got a training accuracy of 92.38% from our model which suggests that the model is performing well on the training dataset. It is correctly classifying approximately 92.38% of the training data points.

However, high training accuracy can also meanss it is a sign of overfitting. The model might have learned to memorize the training data but may not generalize well to new, unseen data.

The testing accuracy is also known as the validation accuracy which represents the model's performance on a testing dataset that it has not seen during training. We got a testing accuracy of 91.37% on the testing data. This suggests that the model correctly classifies about 91.37% of the testing data which is unseen.

The testing accuracy is slightly lower than the training accuracy, which is expected but generally acceptable. A small difference between training and testing accuracy indicates good model generalization.

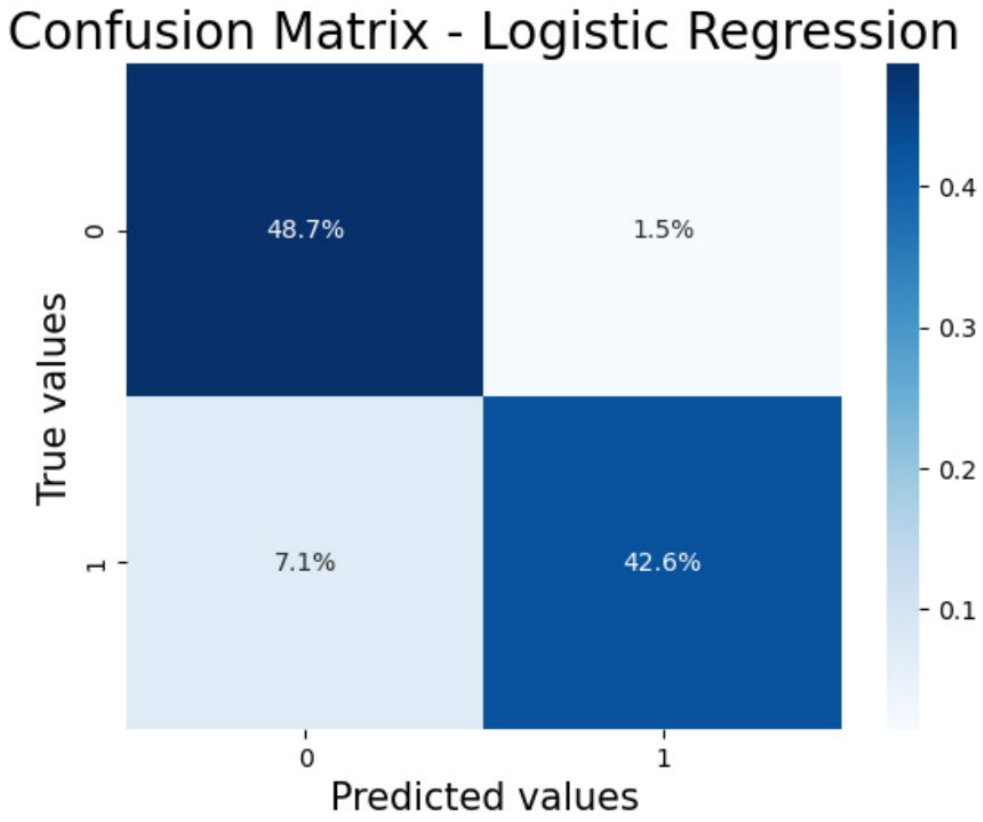

Table 10:Confusion matrix

It can be helpful to examine the confusion matrix, which provides information about true positives, true negatives, false positives, and false negatives. This allows you to assess the model's performance in terms of precision, recall, F1-score, and specificity.

True positives are 48.7% where the model predicted the transaction as fraudulent when it was indeed fraudulent. The model recognised 42.6% true negatives, which means it correctly predicted 42.6% non fraudelent transactions. Percentage of false positives and false negatives are less which indicates the transactions flagged as fraudulent but in real it was a legitimate and vice versa. False negative of 1.5% represents a missed fraudulent transaction. The false positives indicates legitimate transactions detected as fraud.

In the context of credit card fraud detection, our key goal is to minimize false negatives, which are missed fraudulent transactions and at the same time keeping false positives at an acceptable level. From the evaluation report, we got he below values.

	precision	recall	f1-score	support
0	0.87	0.97	0.92	99
1	0.97	0.86	0.91	98

The recall % means that the model correctly identified approximately 97% of all legitimate transactions and 86% of fraudulent transactions in the dataset. The F1-Score indicates a good balance between precision and recall. The support value shows that there are 99 and 98 instances of legitimate and fraudulent transactions respectively in the dataset. This overall results in a strong performance.

The R square error metric is 0.65 and this justifies the performance of the model. This indicates the target variable is well explained by the combination of the independent variables as a single unit.

Logistic Regression Curve

The curve shows how much the classifier can distinguish between the true-positive rate versus the false-positive rate, ie, the number of times the classifier hit the prediction against the number of times the classifier missed the prediction.

This model represents the realtionship between features and the probability of a transaction being fraud.

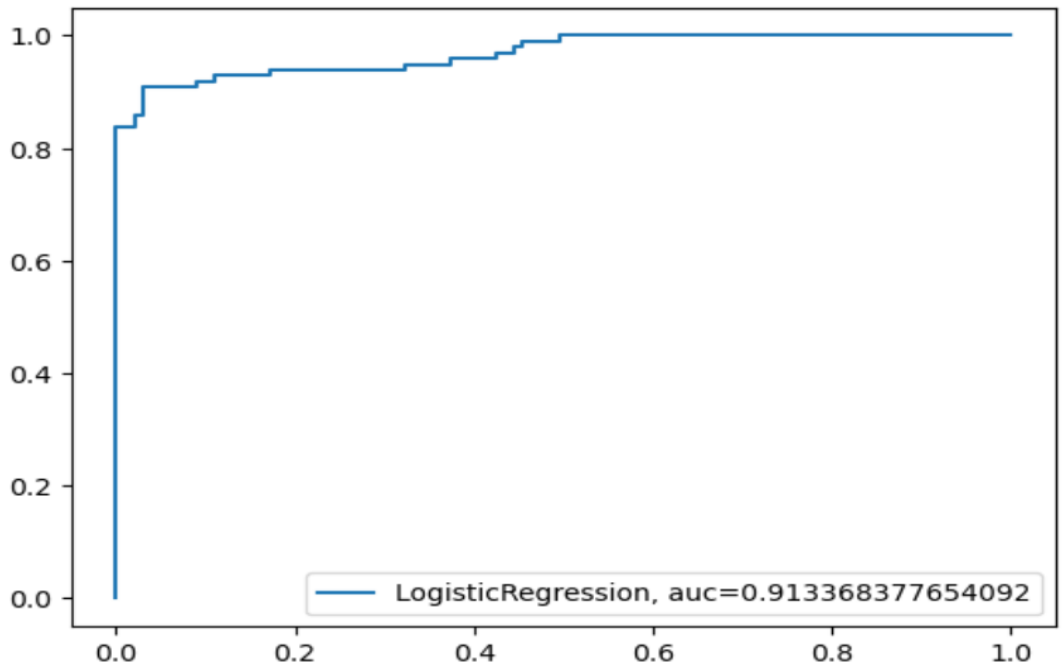

Table 11:Logistic regression model

Unsupervised – K Means

The goal of K means is to group similar data points together into a predefined number of clusters. The closer data points are more similar and farther ones are less similar. K means clustering is not typically used for credit card fraud detection models. Fraud detetction is a supervised problem, where the model is trained on labels and features to make the predictions.

The Silhouette Coefficient is obtained as 0.56. A higher value of this coefficient indicates that the data points are well clustered. The clusters are well defined and they are reasonably separated. The Calinski_harabasz Coefficient is 4951.58. This is assessing the quality of clusters in a dataset. The high score suggests that the chosen number of clusters and the data structure are a good match.

The Completeness score is 0.0205 which is low and suggets the clustering alogorithm can be further improved. Homogeneity score assess if the cluster formed are composed of data points which truly belong to that cetagory. But a low score of 0.0475 implies the algorithm will find it difficult to identify the different classes in the dataset.

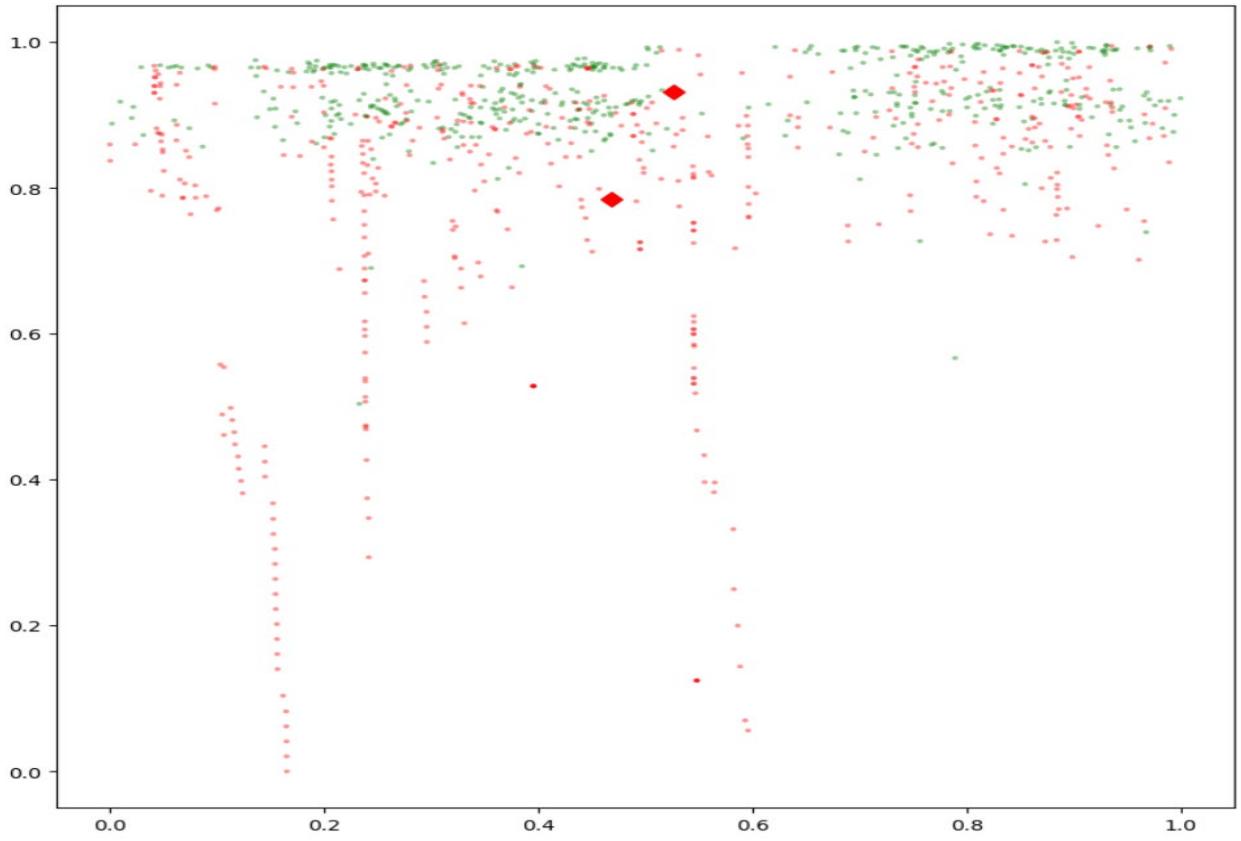

Table 12:K Means clustering

Conclusion

From our analysis, we can conclude that even though credit card poses a significant threat for financial institutions, with proper utilization of machine learning we can address this issue. The logistic regression model has been found effective in identifying the potential fraudulent transactions. The accuracy of model was very high which is mostly due to proper data preprocessing and balancing. Oversampling or under sampling techniques are crucial to eliminate class imbalance issues. With innovative technologies, credit card fraud detection methods can be made more robust.

References

https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud/data

Bart Baesens (2013) Fraud Analytics

Hand, Manilla, Smyth((2001) Priniciples os Data Mining

Laura Igual , Santi Seguí (2017) Introduction to Data Science

DATAtab, https://www.youtube.com/@datatab

https://www.digitalocean.com/community/tutorials/r-squared-in-r-programming