**Project Report**

**Find Default (Prediction of Credit Card fraud)**

**Abstract-**

This Project is focused on credit card fraud detection in real world scenarios. Nowadays credit card frauds are drastically increasing in number as compared to earlier times. Criminals are using fake identity and various technologies to trap the users and get the money out of them. Therefore, it is very essential to find a solution to these types of frauds. In this proposed project we designed a model to detect the fraud activity in credit card transactions. This system can provide most of the important features required to detect illegal and illicit transactions. As technology changes constantly, it is becoming difficult to track the behavior and pattern of criminal transactions. To come up with the solution one can make use of technologies with the increase of machine learning, artificial intelligence and other relevant fields of information technology; it becomes feasible to automate this process and to save some of the intensive amounts of labor that is put into detecting credit card fraud. Initially, we will collect the credit card usage data-set by users and classify it as trained and testing dataset using a random forest algorithm and decision trees. Using this feasible algorithm, we can analyze the larger data-set and user provided current data-set. Then augment the accuracy of the result data. Proceeded with the application of processing of some of the attributes provided which can find affected fraud detection in viewing the graphical model of data visualization. The performance of the techniques is gauged based on accuracy, sensitivity, and specificity, precision. The result is indicated concerning the best accuracy for Random Forest are respectively.

Keywords— Random forest algorithm, Criminal transactions, Credit card

**1. INTRODUCTION**

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency compared to

the usual way. Another way is the detection and

prevention measures put in place by the government

and banks. Credit card fraud stand as a major

problem in the world nancial institutions resulting

in annual loss of billions of data. Over the years, the

government and banks implemented some steps to

detect credit card fraud, but perpetrators are still

devising various means to avoid these detections

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant f

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency co

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency co

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency co

PREDICTIVE ANALYTICS ON CREDIT CARD FRAUD

DETECTION USING CLASSIFICATION MODELS

Oluwadamilola Oluwole

Computer Science Department

Ausn Peay State University

Clarksville Tennessee, USA

ooluwole@my.apsu

1. INTRODUCTION

With online shopping growing by the day, credit

cards are used to purchase goods and services either

physically or virtually. In the physical based

approach, the cardholder presents his card physically

to a merchant for payment. For fraudulent

transaction to be carried out by this approach, an

aacker steals the card and uses it for transaction. In

the virtual approach, an aacker can use the secure

code, card number, expiration date etc., to perform

fraud on the account. One way this fraud can be

detected, is to analyze the spending paern on a card

and nd out if there is an inconsistency co

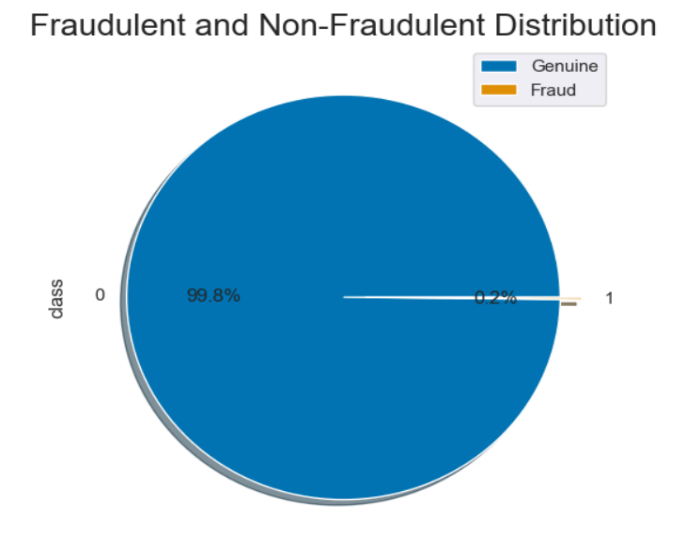
With online shopping growing by the day, credit cards are used to purchase goods and services either physically or virtually. In the physical based approach, the cardholder presents his card physically to a merchant for payment.

For fraudulent transaction to be carried out by this approach, an attacker steals the card and uses it for transaction. In the virtual approach, an attacker can use the secure code, card number, expiration date etc., to perform fraud on the account. One way this fraud can be detected, is to analyze the spending pattern on a card and find out if there is an inconsistency compared to the usual way. Another way is the detection and prevention measures put in place by the government and banks.

Credit card fraud stand as a major problem in the world financial institutions resulting in annual loss of billions of data. Over the years, the government and banks implemented some steps to detect credit card fraud, but perpetrators are still devising various means to avoid these detections. This project aims to perform predictive analytics experiments on a credit card dataset using three classification algorithms. The results gotten in these experiments can help to know the best model to use when trying to predict if a transaction is a normal one or a fraudulent one. The classification algorithms to be used in these experiments will be Logistic regression, decision trees and naïve bayes classifiers.

2. DATA

Datasets are very important when working on classification problems. One major challenge that could be faced by scientists and researchers is the task of gathering data. Some datasets are made public while some are not especially when it relates to finance domain. There are 284,807 records in the dataset out of which only 492 are fraudulent. Because of the fewer number of transaction for a particular class, the data can be said to be unbalanced. The unbalanced class distribution can be visualized in a bar diagram given below



The dataset contains the numerical values on features which have been transformed by a dimension reduction technique known as Principal Component Analysis (PCA). However, due to the confidentiality issue, the original features were not disclosed. The dataset was made up of 30 features out of which 28 features have been generated by principal component analysis. In this project, we will make use of the logistic regression model, decision trees and naïve bayes in the predictive analysis of credit card fraud.

For this project, the dataset will be split into 80% training set and 20% test set. This is to allow for a more accurate result. The models were fit on the training dataset and the test set were used to test the performance of the trained model.

As earlier mentioned, the dataset is highly unbalanced

Which showed higher number of normal transactions than fraudulent. In a situation as this, if make use of this dataset as it is to train a model, the model will be biased to the normal transactions which will eventually result in the poor performance of the model when tested with a test data. In order to resolve this problem, there will be a need to resample the dataset. There are many techniques that can be used to achieve this, but for the sake of this project, random oversampling technique was used. Some of these resampling techniques are discussed below.

1. **Random Undersampling**

This involves randomly selecting samples from the majority class and removing those samples from the training dataset. One drawback to this approach is that there could be vital information loss.

1. **Random Oversampling**

This involves randomly selecting examples from the minority class and adding them to the training dataset. It involves replicating the minority class until the two classes have equal frequency. This method does not lead to information loss as compared with the random undersampling method but there could be overfitting of data. For this project, the random oversampling was used before training the dataset.

3. METHODS

Predictive analytics entails the analysis of data, statistical models and machine learning algorithms with the ability to show likelihood of outcomes in the future based on a historic dataset. For this project, some algorithms are used for the analysis. These algorithms are: Logistic regression, naïve bayes classifiers, and decision trees.

**a. Logistic Regression**

Logistic regression is a model mainly used for classification purpose when the target variable is categorical. It is a statistical model that involves the use of a logistic function in modeling a binary response variable. There are some types of logistic regression but for the purpose of this project, the binary logistic regression will be used which will involve two classes. In this model, the prediction is expressed as the chances of obtaining outcome belonging to each class. Logistic regression makes use of an equation where input values(x) are linearly combined with weights or coefficient values which are learned during the training to predict an outcome usually called the y value. This value will be modelled as a binary value (0 or 1). An example of a logistic regression is shown below,

y = e^(c + m\*x) / (1 + e^(c + m\*x))

y is the predicted output, c is an intercept, m is the coefficient of x. Basically, logistic regression uses 0.5 as its threshold value. It classifies probability values into two classes; where probability values below 0.5 belongs to class 0, and probability values above 0.5 belongs to class 1.

**b. Decision Trees**

Decision trees are one of the tools used in predictive analysis. It is a non-parametric method that can be used for both regression and classification. It is a model used to forecast the value of a target variable by learning from the data features. The algorithm used in decision tree classifier is called the ID3 algorithm.

This algorithm tends to construct decision trees using some approaches such as top-down or greedy approach. It starts by selecting a best attribute and assigning it as the decision attribute for the node. It then creates a new descendant otf the node. It sorts the training example to the appropriate descendant node leaf of the node. It continues by sortiung the training examples until it gets to an appropriate node leaf. If these are classified, it stops and if otherwise, it iterates over the new leaf nodes. One important thing to note is how the best attribute is chosen. For the ID3 algorithm, the attribute with the most information gain could be the best attribute. Before we define information gain, it is necessary to define a measure which involved the use of information theory which is called the entropy. An entropy can be said to measure the level of impurity in each sample.

It is mathematically shown below,

Diagram, text

Description automatically generated



S is the training sample

P\_+ is the proportion of positive samples

P\_- is the proportion of negative samples.

After the entropy has been computed, information gain could be defined as a measure of the effectiveness of an attribute in the training dataset. Information gain can be shown mathematically below,



This could also be written as



**c. Naïve Bayes**

. This is a classifier that has its probabilistic model based on the Bayes’s theorem. For classifications, it is necessary to use a variable X to predict Y.

This can be done using the bayes’ theorem shown below,

Text

Description automatically generated

This could further be explained as,

Text

Description automatically generated with medium confidence

P(Y|X) cannot be gotten directly; however we can obtain P(X|Y) and P(Y) from the training data.

4. IMPLEMENTATION

For this project, the sklearn library in python was used for all the classifiers. The pandas library was used to read the credit card dataset, the matplotlib library was used to plot the charts and the curve. The data was first loaded, and features were selected and stored in a variable ‘X’. After the extraction of attributes and corresponding labels, they were then split into training and test data. The classifiers (logistic regression, decision trees and naïve bayes) were imported from the sklearn library and were fit one after the other on the attributes. For the logistic regression, some hyperparameters were used for fine tuning. For this project, the classifiers were first fit on the unsampled training dataset to get the performance. They were later fit on the oversampled training datasets. After fitting the models on the training dataset, the classifiers were used to predict the labels of the test attributes. The predicted classes were evaluated on metrics as discussed below,

a. Confusion Matrix

This is a table of four combinations of predicted and actual values. It is mainly used in predictive analysis and for computing other performance metrics suchas accuracy, recall, precision, etc. It describes the overall performance of a model when used on some dataset. For this project which involves binary classification, a 2x2 confusion matrix as shown below.

Table

Description automatically generated

Figure : Confusion Matrix

**b. Recall**

Recall, also called sensitivity, is the fraction of true positives to the actual positivecases. recall consists of the number of true positives found from the true positive cases. The formular is given below,

Recall = TP / TP + FN

**c. Precision**

Precision is a fraction of true positives to the sum of true positive and false positive. This metric gives us the number of cases that are true positives. The formular for it is shown below,

Precision = TP / TP + FP

**d. F1-score**

This is the harmonic mean of the recall and the precision. It has values within 0 and 1 range. The mathematical formular for calculating f-score is shown below,

A picture containing graphical user interface

Description automatically generated

**e. Area Under Receiver Operating Curve**

This is the area under the ROC curve. AUC is one of the most accurate metrics in predictive analytics especially on imbalanced data. It gives us an insight of how a model will perform when used at different probability thresholds.

The threshold is 0.5 by default for classification. The AUC score can be used to plot a curve showing the performance of models. This can be done by computing the True Positive Rate (TPR) also known as Sensitivity and False Positive Rate (FPR) also known as Specificity and plotting them to get the ROC curve which is a curve showing the performance of each model by plotting True positive rate (TPR).

The mathematical formular for calculating the TPR and FPR is shown below,



Text

Description automatically generated with medium confidence

Sensitivity and specificity are inversely proportional. This means that a decrease in the threshold, causes the sensitivity to increase and causes the specificity to decrease. On the other hand, an increase in the threshold, causes the specificity to increase, thereby decreasing the sensitivity. For this project, of all the metrics explained above, the AUC score was used to evaluate the performances of the models.

5. RESULTS

The results of the experiment carried out on both the unsampled data and the sampled data are shown below,

5.1 Logistic Regression

Undersampling:



Oversampling:



**5.2 Decision Tree**

**Undersampling:**

****

**Oversampling:**



# **5.3 Naïve Bayes (NB)**

**Undersampling:**

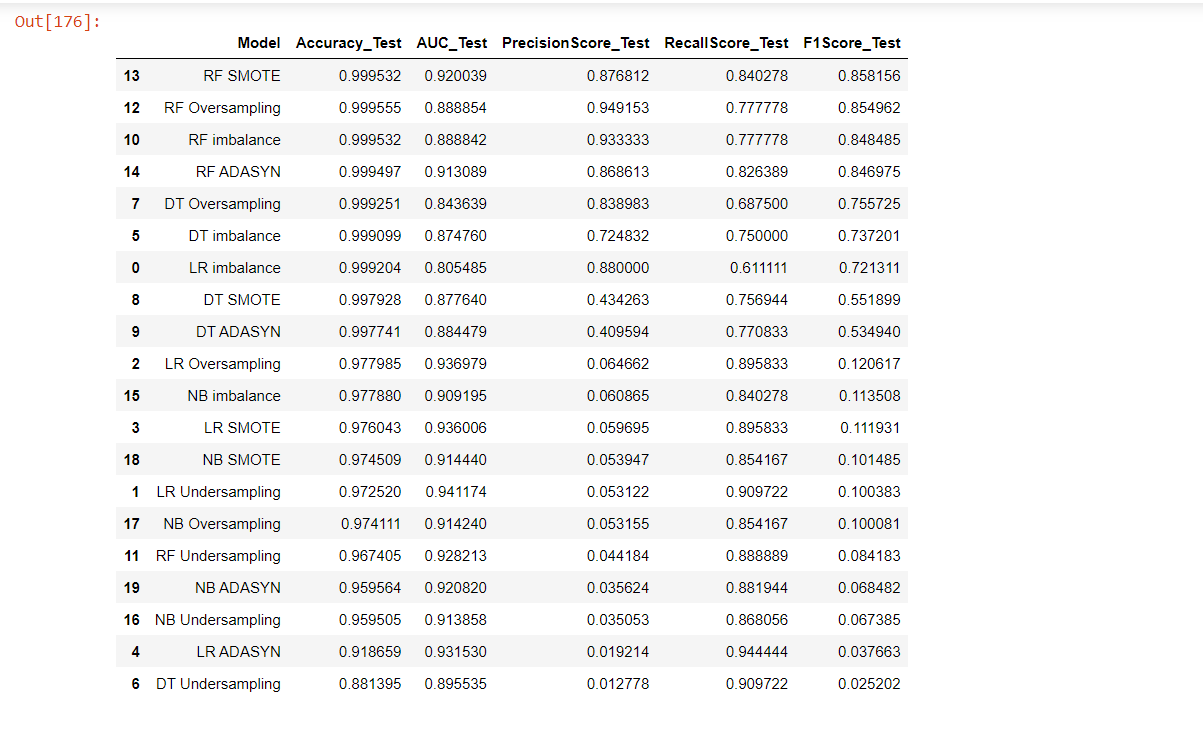
****

**Oversampling:**

****

**5.4 Result Summary**

Below is a summary of the results obtained from the experiment,



6. CONCLUSION

In this project, credit card transactions were modelled on different algorithms to help detect fraudulent transactions. The models were fit on both the unsampled data and the sampled data. This project was aimed at getting the algorithm that will perform well whether the class is evenly balanced or highly unbalanced. For the logistic regression model, it performed very well on sampled data but performed poorly on unsampled data.

This model will only be best to use when we have the balanced training dataset. The naïve bayes model performed well on both the unsampled and the sampled data from the experiment carried out. This shows us that the naïve bayes classifier will do well whether it is a balanced dataset or an imbalanced dataset. I will recommend that for future work, researchers should perform experiments on other classifiers using the same set of data.