TABLE OF CONTENTS

Definitions, Acronyms and Abbreviations 2

1. Introduction..3
1.1. Overview 3
1.2. Scope 3
1.3. Objective...3

2. Literature Survey 4

3. Methodology 4
3.1. Proposed Approach...4
3.2. High Level System Architecture 6

4. Environment Requirements...7
4.1. Hardware Requirements..7
4.2. Software Requirements...7
4.3. Data Requirements 7

5. Results...8

6. Conclusion 9

7. Future Works..9

8. References 10

Definitions, Acronyms and Abbreviations

- **Decision Tree** - A decision tree is a decision support tool that uses a tree-like graph or model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm that only contains conditional control statements.

- **Random Forest** - Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean prediction of the individual trees.

- **Artificial Neural Network** - An artificial neural network is an interconnected group of nodes, inspired by a simplification of neurons in a brain. Here, each circular node represents an artificial neuron and an arrow represents a connection from the output of one artificial neuron to the input of another.

- **SVM** - Support-vector machines or SVMs are supervised learning models with associated learning algorithms that analyse data used for classification and regression analysis.

- **Accuracy -** The measure of correct predictions made by the model – that is, the ratio of fraud transactions classified as fraud and non-fraud classified as non-fraud to the total transactions in the test data.

- **Recall -** Sensitivity, or True Positive Rate, or Recall, is the ratio of correctly identified fraud cases to total fraud cases.

- **Precision -** Precision is the ratio of correctly predicted fraud cases to total predicted fraud cases.

1. Introduction

1.1. *Overview*

Nowadays the usage of credit card has dramatically increased. As credit card becomes the most popular mode of payment for both online as well as regular purchase, cases of fraud associated with it are also rising. We have tried to create a model for credit card transactions and how it can be used for fraud detection. We train the model with the normal behaviour of a cardholder. If an incoming credit card transaction is not accepted by the trained model with sufficiently high probability, it is consider to be fraudulent. At the same time , we try to ensure that the correct transactions are not rejected. We have used different models and compared them. At the end, we present the experimental results along with the output of the different models.

1.2. *Scope*

It is important that credit card companies are able to recognise fraudulent credit card transactions so that customers are not charged for items that they did not purchase. Keeping this in mind, we want the scope of our project to be able to build a model that correctly identifies a fraudulent transaction, and hence helping in protecting the innocent customers and the banks and the agencies.

1.3. *Objective*

- Create an application to detect fraud Credit Card transactions
- Implement different classification models like Decision Tree, Artificial Neural Networks, Random Forest and SVM
- Provide security to the customers at the time of transactions.
- Compare different trained models for accuracy, recall and precision

2. Literature Survey

The review paper [1] provides the results of an investigation regarding the difficulties of credit card fraud detections and gives details of various types of credit card fraud detection techniques. The research paper [2] gives the calculation of decision tree and random forest and the phases involved in the algorithm. The research paper [3] provides information on how to use SVM for credit card fraud detection. The research paper [4] provides information on how to use ANN for credit card fraud detection.

3. Methodology

3.1. Proposed Approach

What is the proposed approach?

For the purpose of this project, we have used different classification models/algorithms which are as follows:

1. Decision Trees
2. Random Forest
3. Artificial Neural Network
4. Support Vector Machines(SVM)

The dataset has 31 features most of which are the results of PCA. We remove the time attribute. Class 0 represents normal transaction, and class 1 represents fraud. We split the dataset into two sets. 70% of data in training set to train the model and remaining 30% of data in testing set to test the model. After making the classifier, we fit it onto the training set. Once the model is trained, we run it on the test data. We do this for different models and compare their result.

As the dataset has 492 frauds out of 284,807 transactions, we can see that the dataset is highly unbalanced and Class 1 is under-represented. Therefore the trained model is inefficient and it is trained to predict only Class 0 because it does not have sufficient training data. We get a high accuracy because of the skewness of the dataset, therefore it is not advisable to test the accuracy so we calculate recall.

To remove the skewness problem, we use under-sampling so as to create a balanced dataset. For this, we scale the amount feature and remove the old amount column. Now we keep all of our under-represented data(Class 1) while add the same number of features of Class 0 to create a new dataset comprising of an equal representation from both classes. Now that we have a balanced dataset, we split it into training and test set, train the model and test it for the new dataset. Finally, we test the model against the original skewed dataset. We create confusion matrix for the same from which we calculate accuracy, precision and recall.

Why is it better than the current approach to solve the problem?

1. The detection of the fraud use of the card is found much faster than the existing system.
2. In case of original system even the original card holder is also checked for fraud detection.
3. But in this system no need to check the original userWe can find the most accurate detection using this system.
4. This reduces the tedious work of an employee.

3.2. *High Level System Architecture*

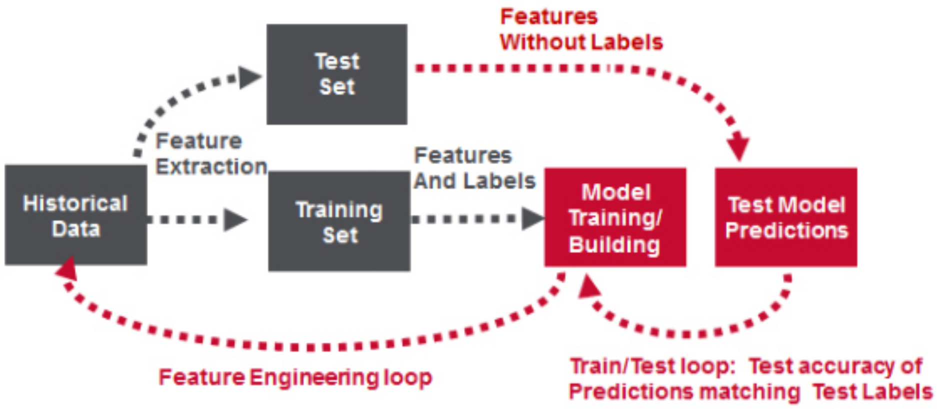


Image 1. Workflow of the model

The supervised machine learning workflow which we have used has the following steps:

- Feature engineering to transform historical data into feature and label inputs for a machine learning algorithm.
- Split the data into two parts, one for building the model(training set) and one for testing the model(testing set).
- Build the model with the training features and labels.
- Test the model with the test features to get predictions. Compare the test predictions to the test labels.
- Calculate accuracy, precision and recall.
- Repeat the steps with different classification models and compare them.

4. Environment Requirements

4.1. *Hardware Requirements*

• SYSTEM	: intel i5
• HARD DISK	: 128GB
• RAM	: 8GB

4.2. *Software Requirements*

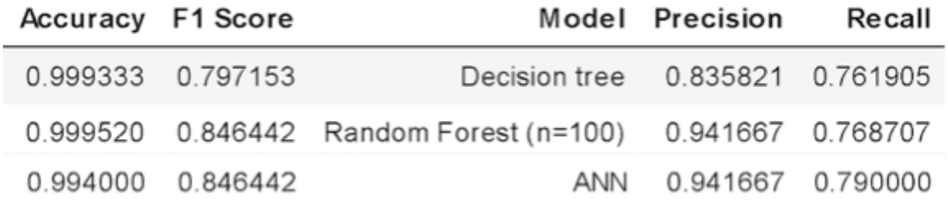
- Jupyter Notebook
- Pandas library
- Numpy library
- Matplotlib library
- Seaborn library
- Sklearn
- Ubuntu or Windows Operating System

4.3. *Data Requirements*

credit.csv (https://www.kaggle.com/mlg-ulb/creditcardfraud#creditcard.csv)

This datasets contains transactions made by credit cards in September 2013 by European cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.It contains only numerical input variables which are the result of a PCA transformation. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and ‘Amount. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

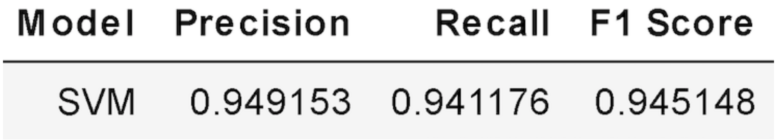
5. Results



Accuracy	F1 Score	Model	Precision	Recall
0.999333	0.797153	Decision tree	0.835821	0.761905
0.999520	0.846442	Random Forest (n=100)	0.941667	0.768707
0.994000	0.846442	ANN	0.941667	0.790000

Table 1. Recall,Precision, & Accuracy obtained when model trained from Original data

We see in Table 1. that the accuracy is very high. This is because the dataset is highly skewed and unbalanced. Therefore, it is not appropriate to compare on basis of accuracy so we compare recall. We can see that the recall value is best for ANN, so we can say that it has performed better than the other models



Model	Precision	Recall	F1 Score
SVM	0.949153	0.941176	0.945148

Table 2. Recall,Precision, & Accuracy obtained when model trained from Under-sampled data

After we sample the dataset and train the model again with under-sampled data we see in Table 2. that the recall value is much better as compared to the original skewed data.

6. Conclusion

In this model we have analysed and detected the fraud in online credit-card transactions in real time. It classifies the transactions according to the spending habits of the customer which helps in detecting whether the current transaction is genuine or not. When we trained the model from original dataset, the recall value for ANN was maximum among all, so we can say that it has performed better than the other two models. But it was not appropriate to train the model on the entire dataset as the dataset was highly unbalanced, so we used under-sampling to create a more balanced dataset and then trained the model on this new dataset.

We can see from the tables, that the recall, precision and accuracy are more appropriate when we use the model trained using under-sampled data as compared to the values that we got from the original skewed data.

7. Future Works

We can further improve the model by having more training data and also by either using a better algorithm or by modifying our parameters. Speed of the software can be enhanced by implementation of algorithms of less complexity.

Also here we are only detecting which transactions are fraud and which are not. We can try to improve it to even prevent it. Further enhancement can be

done by making this system secure with the use of certificates for both merchant and customer and as technology changes new checks can be added to understand the pattern of fraudulent transactions and to alert the respective card holders and bankers when fraud activity is identified. The dataset available on day to day processing may become outdated, it is necessary to have updated data for effective fraud behaviour identification.

8. References

[1] Samenah Sorournejad, Zahra Zojaji, Reza Ebrahimi and Amir Hassan Monadjemi, “Survey of Credit Card Fraud Detection Techniques”

[2] Lakshmi S V S S, Selvani Deepthi Kavilla, “Machine Learning For Credit Card Fraud Detection”

[3] Sitaram Patel and Sunita Gond, “Supervised Machine(SVM) Learning For Credit Card Fraud Detection”

[4] Chandrahas Mishra, Dharmendra Lal Gupta and Raghuraj Singh, “Credit Card Fraud Identification Using Artificial Neural Network”