**CREDIT CARD FRAUD DETECTION**

**Prapti Patra, Srijal Vedansh, Vishisht Ved, Anup Singh**

[**prapti565@gmail.com**](mailto:prapti565@gmail.com)**,** [**srijal.vedansh7@gmail.com**](mailto:srijal.vedansh7@gmail.com)**,** [**vishishtved8@gmail.com**](mailto:vishishtved8@gmail.com)**,** [**anupsingh841208@gmail.com**](mailto:anupsingh841208@gmail.com)

**School of Computer Engineering, Kalinga Institute of Industrial Technology, Bhubaneswar, India.**

**Abstract**

One of the most frequent problems that we are facing today is Credit Card Fraud Detection and the most definite reason behind it is the phenomenal increase in online transactions. Currently, we’re often facing such fraud cases due to unauthorized purposes of money transactions in our everyday life. Hence, to detect such fraudulent activities, we can use the credit card fraud detection system. In this paper, we propose our approach to detect such frauds.

Our project mainly aims to focus on three Machine Learning Algorithms. The algorithms that we have used are Logistic Regression, Decision Tree Classifier and Random Forest Classifier. The derived results are based on accuracy, precision, recall and F1- score. We have used all three algorithms for both Undersampling and Oversampling cases. The algorithm which provides the best accuracy, precision, recall and F1-score is considered to be the best algorithm to detect the fraud.

**Keywords**

Credit card, Fraud detection, Credit card fraud, Logistic regression, Decision tree.

1. **Introduction**

Fraud detection in credit cards involves monitoring the activity of card holders in order to estimate and prevent unauthorized transactions and objectionable behavior. In the present world, credit card fraud is a growing concern especially in the corporate and finance industries. Our population is highly dependent on the internet today and that’s one of the main reasons for online fraudulent transactions. Although, offline transactions go through similar fraud cases as well. We have data mining techniques to detect these frauds but the result is not very accurate. Hence, we need some promising methods to minimize such credit card frauds. We can do that with the help of efficient machine learning algorithms. [1]

As shown below in Figure 1.1, due to the increasing growth of internet users, the finance company issues credit cards to individuals. When coming to the issuance of the card, the condition involved is that the cardholder will pay back the original amount they borrowed along with the additional charges they agreed to pay.

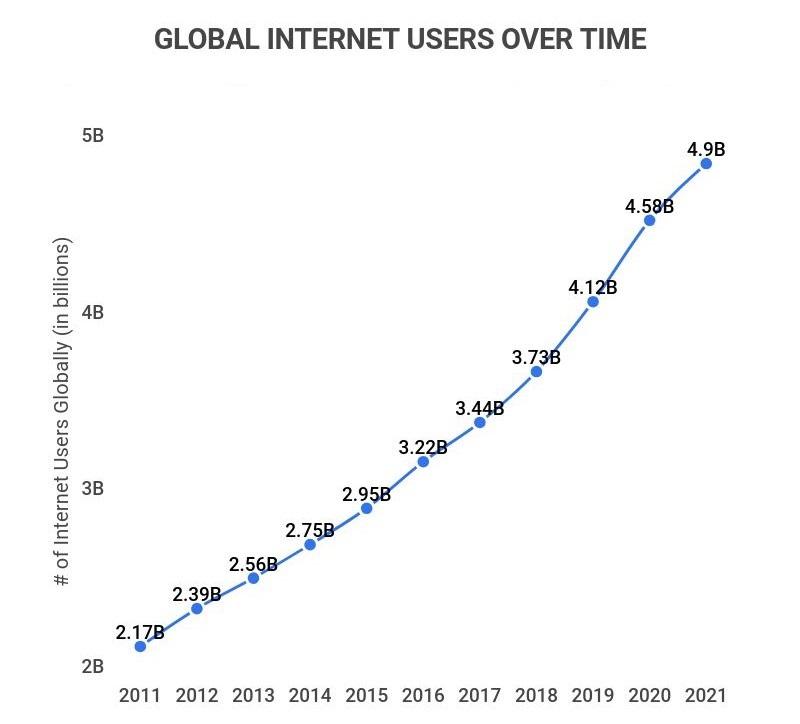


Figure 1.1: Growth of Internet Users

Machine learning algorithms are designed in order to analyze all the authorized transactions and report the suspicious ones. Professionals investigate the reports and contact the cardholders to verify whether the transaction is legitimate or fraudulent. [2]

We have used three algorithms.

***Logistic regression:***

Logistic regression is a statistical method used in machine learning for binary

classification problems, where the goal is to predict whether an input belongs to one of two categories. The logistic regression algorithm uses a logistic function to calculate the probability of the input belonging to each category.

***Decision Tree Classifier:***

A decision tree classifier is a machine learning algorithm used for classification problems. It is a type of supervised learning algorithm that uses a tree-like model to classify data. The tree is made up of nodes and branches, where the nodes represent a test on an attribute and the branches represent the outcome of the test. At each node, a decision is made based on the value of a feature or attribute, and the decision leads to the next node in the tree until a classification decision is made at the final node.

***Random Forest Classifier:***

Random forest is a popular ensemble learning method in machine learning. It is a supervised learning algorithm that builds multiple decision trees and combines their predictions to produce a final prediction.During the training phase, the random forest algorithm builds a forest of decision trees by repeatedly selecting a random subset of the data and features, and then growing a decision tree using that subset. The algorithm then combines the predictions of all the decision trees in the forest to make a final prediction. [3]

The figure 1.2 below shows a rough architecture of the fraud detection system.

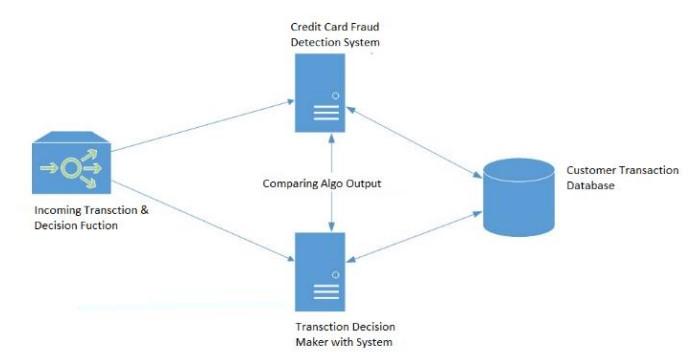


Figure 1.2: Rough architecture diagram for Fraud Detection [1]

1. **Main Contribution**

Our contribution in the paper are as follows:

* Objective- Our objective is to detect fraud credit card transactions using machine learning techniques.
* Method- This paper makes the use of three ML techniques namely Logistic Regression, Decision Tree Classifier and Random Forest Classifier.
* Result- Logistic Regression provided the best accuracy on undersampling data with 95.78%. Whereas on oversampling data, Random Forest Classifier provides the best accuracy with 99.99%.

1. **Literature Review**

Researchers have introduced several new techniques for credit card fraud detection, involving both Machine learning and deep learning algorithms and neural networks. Below listed are some related works in this regard. [4]

* In 2019, Yashvi Jain, NamrataTiwari, Shripriya Dubey, Sarika jain have researched a few fraud detection techniques like SVM, ANN, Bayesian networks, KNN, Fuzzy logic system. The authors inferred that the KNN, decision trees, and SVM algorithms had medium-level accuracy, while Fuzzy Logic and Logistic Regression had the lowest accuracy among all the algorithms. On the other hand, Neural Networks, naive bayes, fuzzy systems, and KNN algorithms had a high detection rate. Logistic Regression, SVM, and decision trees offered high detection rates at the medium level. However, there were two algorithms, namely ANN and Naive Bayesian Networks, which performed better at all parameters, but they were expensive to train. A significant drawback of all the algorithms was that they did not produce the same results in all types of environments. They provided better results with one type of dataset and poor results with another type. For instance, KNN and SVM algorithms performed well with small datasets, whereas logistic regression and fuzzy logic systems showed good accuracy with raw and unsampled data.
* In 2019, Heta Naika and Prashasti Kanikar performed research on four algorithms namely Naive Bayes, Adaboost, Logistic Regression and J48. Naive Bayes is a classification algorithm that utilizes Bayes' theorem to calculate the probability of an event occurring. On the other hand, Logistic regression is similar to linear regression, but it is typically used for classification tasks. Linear regression is commonly used for prediction or forecasting values. J48 is an algorithm used for creating a decision tree and solving classification problems. It is an extension of the ID3 algorithm and is widely used and analyzed in machine learning, which works primarily with categorical and constant variables. Adaboost is designed for binary classification, mainly used to improve the performance of decision trees. It is often used in fraud detection, such as classifying transactions as fraudulent or non-fraudulent. Researchers have found that both Adaboost and Logistic Regression have the same accuracy; however, the Adaboost algorithm is more suitable for detecting credit card fraud due to its faster processing time.
* In 2019, research has been done by Sahayasakila V, D.Kavya Monisha, Aishwarya and Sikhakolli Venkatavisalakshiswshai Yasaswi. The authors have introduced two significant algorithmic techniques - the Whale Optimization Techniques (WOA) and Synthetic Minority Oversampling Techniques (SMOTE). The primary objective of these techniques is to enhance the convergence speed and resolve the issue of data imbalance. SMOTE technique addresses the problem of class imbalance by generating synthetic transactions that are re-sampled to ensure data accuracy. The WOA technique is then applied to optimize the synthesized transactions. This algorithmic approach improves the reliability, efficiency, and convergence speed of the system.
* In 2018 Navanushu Khare and Saad Yunus Sait, in their study, the authors investigated decision trees, random forest, SVM, and logistic regression on a highly skewed dataset. They evaluated the performance based on metrics such as accuracy, sensitivity, specificity, and precision. The results showed that the accuracy for Logistic Regression was 97.7%, for Decision Trees was 95.5%, for Random Forest was 98.6%, and for SVM classifier was 97.5%. The authors concluded that the Random Forest algorithm performed the best and had the highest accuracy among the other algorithms for detecting fraud. They also found that the SVM algorithm had a data imbalance problem and did not produce better results for detecting credit card fraud.
* In a related domain, Yu and Wang proposed an Outlier data-mining approach to detect fraudulent data-points in a dataset. Their method considers fraud as an isolated point in the vector space, which can either appear independently or be part of a small group of clustered data-points. According to their findings, the technique achieves an accuracy of 89.4% when the outlier threshold is set to 12.

1. **Proposed Model**

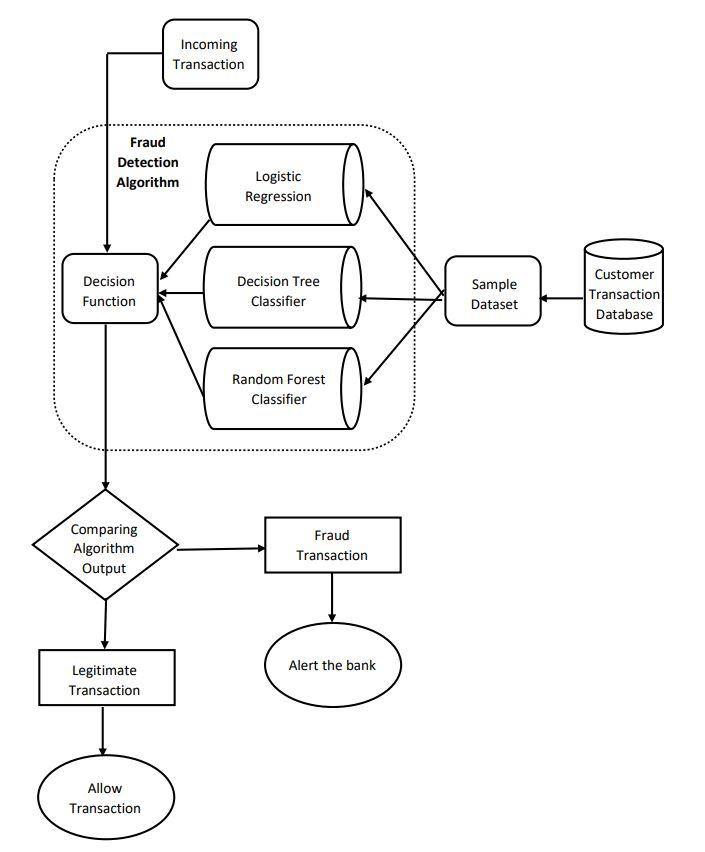


Figure 4.1: Methodology of Fraud Detection

Figure 4.1 is a flowchart showing the methodology of our fraud detection system. We collect the sample dataset from the customer transaction database, and we train it with the three models that we are using namely Logistic Regression, Decision Tree Classifier and Random Forest Classifier. When an user performs a transaction, it is taken into the decision function using the fraud detection algorithm. The output is compared and analyzed. If found Legitimate, transaction is approved. If found fraud, the respective bank is alerted for verification.

1. **Diagram and Explanation**

The figure 5.1 below shows the basic steps for project planning and execution.

* First we collect input from the dataset and split it into training and testing sets.
* Next, we prepare the data. We choose a model and train our dataset.
* We deploy our model, and evaluate its performance by testing it.
* Finally, we use our model on testing data to make predictions accurately. [5]

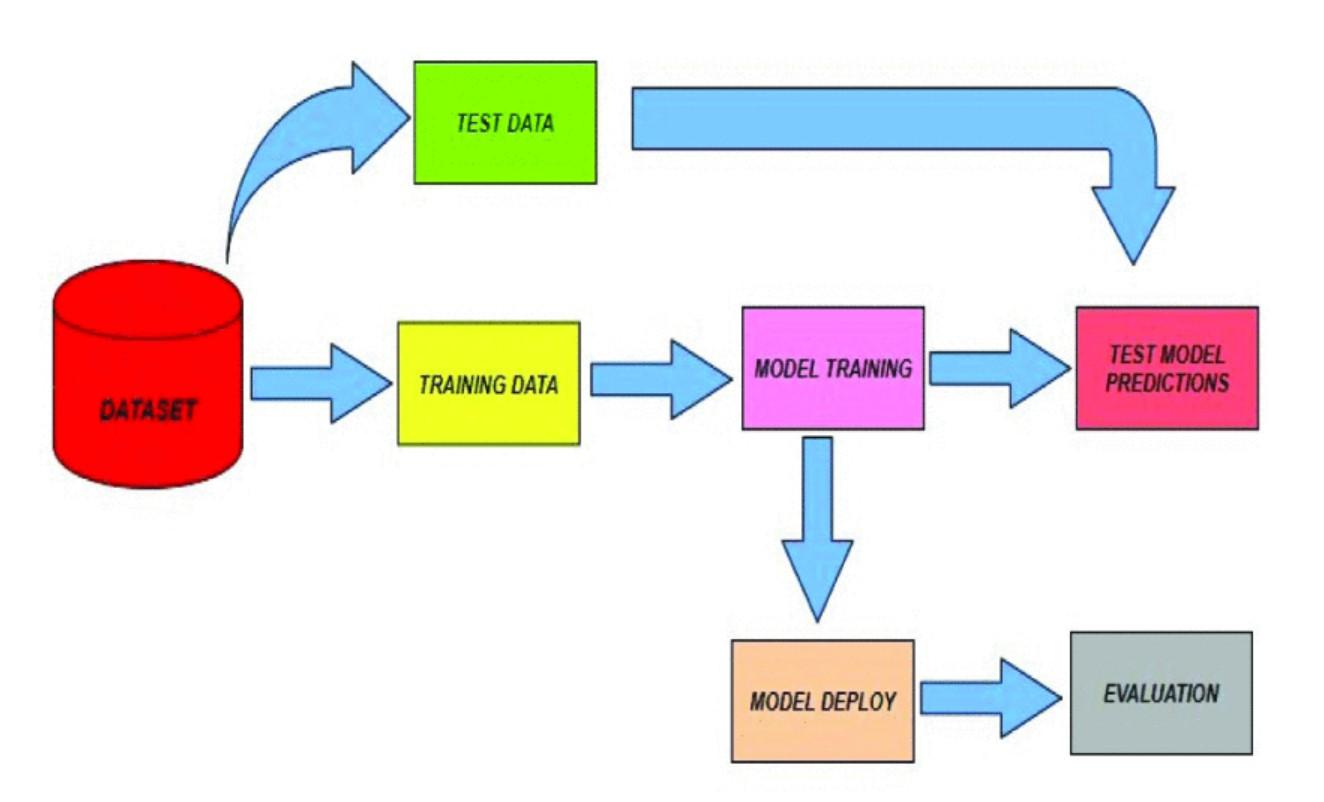


Figure 5.1: Steps of project planning

1. **Result Analysis**

We have obtained our dataset from Kaggle [6], a widely used website for downloading datasets.

A full cross-validation has been performed to validate the performance of the algorithms.

True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

UNDERSAMPLING: For undersampling, After performing a complete mean of the following data it is observed that **Logistic Regression** provides the best results on the data.

Accuracy- 95.789 Precision- 100

Recall- 92.156 F1-score- 95.918

Table 1: Results for Undersampling data

| **Alogirthm used** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- | --- |
| I. Logistic Regression | 95.789 | 100 | 92.156 | 95.918 |
| II. Decision Tree Classifier | 90.000 | 93.684 | 87.254 | 90.355 |
| III. Random Forest Classifier | 94.736 | 100 | 90.196 | 94.845 |

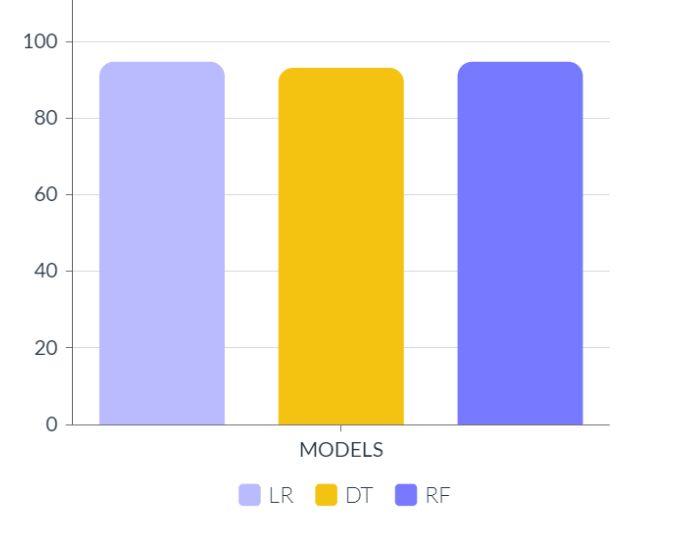


Figure 6.1: Graph representing the accuracy results of three models for Undersampling.

OVERSAMPLING: For oversampling, it is observed that **Random Forest Classifier** provides the best result on our data. Hence, it is best suited.

Accuracy- 99.991 Precision- 99.983

Recall- 100 F1-score- 99.991

Table 2: Results for Oversampling data

| **Algorithm used** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| --- | --- | --- | --- | --- |
| I. Logistic Regression | 94.531 | 97.289 | 91.607 | 94.362 |
| II. Decision Tree Classifier | 99.807 | 99.734 | 99.880 | 99.807 |
| III. Random Forest Classifier | 99.991 | 99.983 | 100 | 99.991 |

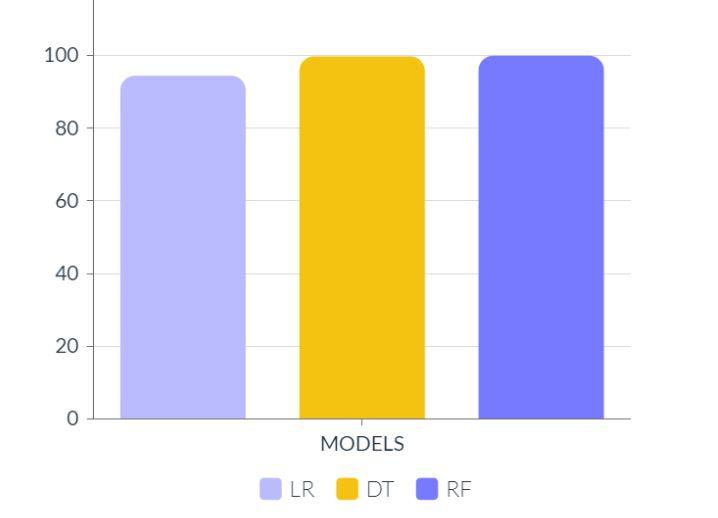


Figure 6.2: Graph representing the accuracy results of three models for Oversampling.

1. **Conclusion and Future Scope**

***Conclusion:***

This paper shows the comparative performance of Logistic regression, Decision tree and Random Forest. Increase in credit card frauds has alarmingly stressed the fraud management system in all the banks, so a machine learning based fraud detection system is used to provide both accuracy and All these three classifiers are trained based on real time credit card transactions which will help us to reduce at least 40-50% of total fraud losses [7] [8].The contribution in this project is summarized as follows :-

1. Three classifiers based on Decision Tree, Logistic regression and Random Forest are trained based on real time credit card transactions data and their performance is calculated based on several metrics.
2. The dataset is sampled based upon two categories undersampling and oversampling, achieving two sets of data distribution.
3. The performance of the dataset is examined based on the accuracy, precision, recall and F1-score. [9]

***Future Scope:***

Given the flexibility of this project, multiple algorithms can be integrated as modules, and their results can be combined to enhance the final result's accuracy. To further refine this model, additional algorithms can be integrated, as long as their output matches the others' format. This modular approach allows for increased versatility and flexibility in the project.

Another opportunity for improvement lies within the dataset. As demonstrated previously, the algorithms' precision improves as the dataset's size increases. Therefore, increasing the dataset's size is likely to enhance the model's ability to detect fraud and reduce false positives. However, gaining the necessary support from banks is essential to achieving this goal. [10]

**8. References**

[1] S P Maniraj, Aditya Saini, Swarna Deep Sarkar, Shadab Ahmed (2019). Credit Card Fraud Detection using Machine Learning and Data Science.

[2] Munira Ansari, Hashim Malik, Siddhesh Jadhav, Zaiyyan Khan (2021). Credit Card Fraud Detection.

[3] [Amlan Kundu](https://link.springer.com/chapter/10.1007/11961635_18#auth-Amlan-Kundu), Shamik Sural and A. K. Majumdar (2006). Two-Stage Credit Card Fraud Detection Using Sequence Alignment

[4] Aditya Raj Singh, Hrishikesh Morade , Sanket Sable , Sunil Hule Guide:- B.R. Patle (2021). Credit Card Fraud Detection using Machine Learning.

[5] [Ruttala Sailusha](https://ieeexplore.ieee.org/author/37088424980), [V. Gnaneswar](https://ieeexplore.ieee.org/author/37088426722), [R. Ramesh](https://ieeexplore.ieee.org/author/37088425380) and [G. Ramakoteswara Rao](https://ieeexplore.ieee.org/author/37088425150) (2020). Credit Card Fraud Detection Using Machine Learning.

[6] Obtained Dataset from Kaggle.

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

[7] Suman (2014). Survey Paper on Credit Card Fraud Detection.

[8] Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy published by IEEE transactions on Neural networks and learning systems vol. 29, no. 8. (2018)

[9] John Richard D. Kho, Larry A. Vea (2017). Credit Card Fraud Detection Based on Transaction Behaviour.

[10] Abrar Nadim , Ibrahim Mohammad Sayem , Aapan Mutsuddy ,Mohammad Sanaullah Chowdhury IEEE (2019). Analysis of Machine Learning Techniques for Credit Card Fraud Detection.