**Mini Project Report on**



**Credit Card Fraud Detection Using Machine Learning**

**Submitted in partial fulfillment of the requirement for the award of the degree of**

**BACHELOR OF TECHNOLOGY**

**IN**

**COMPUTER SCIENCE & ENGINEERING**

**Submitted by:**

**Student Name:**  **University Roll No.**

**Raj Singh Patel 2017530**

***Under the Mentorship of***

**Dr. Jyoti Aggarwal**

**Associate Professor**



**Department of Computer Science and Engineering**

**Graphic Era (Deemed to be University)**

**Dehradun, Uttarakhand**

**July 2023**



**CANDIDATE’S DECLARATION**

I hereby certify that the work which is being presented in the project report entitled **“Credit Card Fraud Detection Using Machine Learning”** in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Computer Science and Engineeringof the Graphic Era (Deemed to be University), Dehradun shall be carried out by the under the mentorship of **Dr. Jyoti Aggarwal, Associate Professor**, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun.

Name: - Raj Singh Patel University Roll no: -2017530 **\**

**Table of Contents**

|  |  |  |
| --- | --- | --- |
| **Chapter No.** | **Description** | **Page No.** |
| Chapter 1 | Introduction | **1** |
| Chapter 2 | Literature Survey | **2** |
| Chapter 3 | Methodology | **3-7** |
| Chapter 4 | Result and Conclusion | **8** |
| Chapter 5 | Conclusion and Future Work | **9** |
|  | References |  |

**Chapter 1**

**Introduction**

In recent years credit card became one of the essential parts of the people. Due to sudden increase in E-commerce, customer started using credit card for online purchasing therefore risk of fraud also increased. Instead of carrying a huge amount of cash in hand it is easier to keep credit cards. But nowadays that too has become unsafe. Nowadays we are facing a big problem on credit card fraud which is increasing in a good percentage. The main purpose is to the survey on the various methods applied to detect credit card frauds. From the abnormalities, in the transaction, the fraudulent one is identified. We address this issue in -order to implement some machine learning algorithm like Random Forest, Logistic Regression, Decision Tree in - order to detect this kind of fraud.

**Why do we need Credit Card Fraud Detection Model?**

The myriads of plastic cards in use worldwide are a gold mine for criminals. By 2027, financial service providers are expected to take a $40 billion hit globally in credit card losses, a significant increase compared to $27.85 bn in 2018.

This growth in losses is partially caused by the rise of electronic transactions. Just imagine that today the average American has more than three credit cards, which amounts to 1.5 billion cards in the US alone. While the number of plastic cards globally numbers an impressive 22.11 billion.

Another reason is that fraudulent methods are getting more sophisticated and thus harder to spot by traditional fraud detection software.

To address such a massive challenge effectively, companies involved in the card payment process need advanced approaches. No wonder that high hopes are placed on machine learning. ML-fueled technologies have made a name for their ability to handle huge amounts of data and discover anomalies that humans may overlook.

**Chapter 2**

**Literature Survey**

**2.1 Introduction**

Several cutting-edge techniques provide several algorithms that help in the identification of different credit card fraud activities. We can create a potent model for identifying credit card fraud if we have a solid understanding of these algorithms. This research aids us in recognising dubious credit card transactions by offering a machine learning approach. We introduce the traits True Positive and False Alarm in this section of our paper. Both of these characteristics are essential in identifying fraudulent behaviour because it can be done so quickly. Given the importance of networks in the modern world, it is crucial that our models be up-to-date in order to perform better detection functions.

This model should be far better to perform real-time analysis if fresh fraudulent activity is discovered. Neural Networks have been used to create a Fraudulent Detection System in addition to conventional machine learning techniques. Because there are numerous attacks on the system designed to steal someone's personal information in order to commit fraud, protecting personal information has grown to be a big challenge for financial companies. Our model includes two key characteristics that can aid in identifying anomalous behavior in charts for various columns, such as time, amount, etc.

**2.2 Models Used And Data Sampling**

Since the machine learning techniques used in the random forest and logistic regression are trained datasets are required for our method. The system's main memory is subsequently loaded with these trained datasets. Since our dataset contains approximately 300,000 values, loading the learned dataset into main memory is a challenging operation. We have eliminated the redundant datasets for this reason. We chose to train our dataset using prior data since our model should be able to identify fraudulent transactions occurring in the current month and will be useful in the real world.

**Chapter 3**

**Methodology**

**Introduction to Machine Learning:**

Machine learning is a field of artificial intelligence that focuses on the development of algorithms and models that can learn from data without being explicitly programmed. It has become an increasingly important part of many areas of technology, including image and speech recognition, natural language processing, and finance.

There are several types of machine learning, including:

1. Supervised learning: In supervised learning, a model is trained on a labeled dataset, where the correct output is provided for each example in the training set. The goal is to make predictions on new, unseen examples based on the patterns learned from the training data.
2. Unsupervised learning: In unsupervised learning, the model is not provided with labeled training examples. Instead, it must discover the underlying structure of the data through techniques like clustering.
3. Semi-supervised learning: Semi-supervised learning is a type of machine learning that combines both supervised and unsupervised learning. It is often used when a large amount of unlabeled data is available, but it is too expensive or time-consuming to label all of it.
4. Reinforcement learning: In reinforcement learning, an agent learns to interact with its environment in order to maximize a reward. The agent receives feedback in the form of rewards and punishments as it takes actions in the environment.

**Tools Used:**

**Google Colaboratory: -** The Colaboratory at Google Research, or "Colab" for short, is a product. Colab is a Python editor for the web that allows anyone to develop and run Python programmes. It's useful in fields like machine learning, data analysis, and teaching. Colab is a hosted Jupyter notebook service that doesn't require installation and gives you free access to computing resources like GPUs.

The availability of Colab resources is neither guaranteed or infinite, and usage limits are subject to change at any time. This is essential in order for Colab to provide free materials.

**Python: -** The language used to develop this project is Python. Python offers readable and concise codes. Since machine learning and artificial intelligence involve complex algorithms, the simplicity of Python adds value and enables the creation of reliable systems.

**Machine Learning Algorithm used:**

**Random Forest:**

A Random Forest Algorithm is a supervised machine learning algorithm which is extremely popular and is used for Classification and Regression problems in Machine Learning. We know that a forest comprises numerous trees, and the more trees more it will be robust. Similarly, the greater the number of trees in a Random Forest Algorithm, the higher its accuracy and problem-solving ability.  Random Forest is a classifier that contains several decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. It is based on the concept of ensemble learning which is a process of combining multiple classifiers to solve a complex problem and improve the performance of the model.

**Logistic Regression:**

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyze the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision-making. For example email spam or not.

**Decision Tree:**

A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. It is a tree-like structure where each internal nodetests on attribute, each branch corresponds to attribute value and each leaf node represents the final decision or prediction.

**Dataset Used:**

The dataset 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. Unfortunately, due to confidentiality issues, the original features and more background information about the data cannot be provided. 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 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.



**Fig: 3.1**

**Building The Model:**

**Data Preprocessing:**

Data preprocessing is a step in the data mining and data analysis process that takes raw data and transforms it into a format that can be understood and analyzed by computers and machine learning.

Raw, real-world data in the form of text, images, video, etc., is messy. Not only may it contain errors and inconsistencies, but it is often incomplete, and doesn’t have a regular, uniform design.

Machines like to process nice and tidy information – they read data as 1s and 0s. So calculating structured data, like whole numbers and percentages is easy. However, unstructured data, in the form of text and images must first be cleaned and formatted before analysis.

Timeline

Description automatically generated

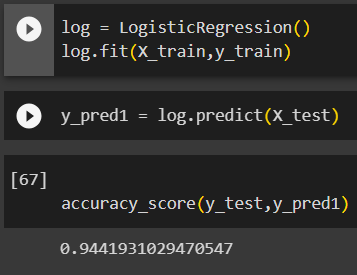
**Fig. 3.2**

**Training the Model:**

With the help of balanced data using techniques like oversampling and undersampling the model was first built using Logistic Regression and later the performance was improved using Decision Tree and Random Forest.

Using Logistic Regression:

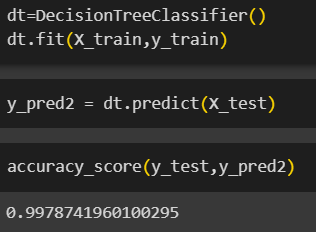
An accuracy of 94.41% was achieved using Logistic Regression Algorithm.



**Fig. 3.3**

Using Decision Tree:

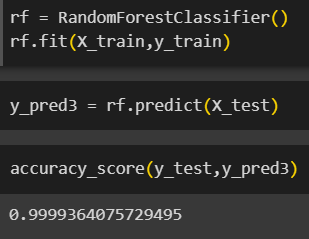
An increase in performance was observed when the model was trained using Decision Tree Algorithm. An accuracy of 99.78% was achieved.



**Fig. 3.4**

Using Random Forest:

An increase in performance was observed when the model was trained using Random Forest Algorithm. An accuracy of 99.99% was achieved.

****

**Fig. 3.4**

**Chapter 4**

**Result and Discussion**

Although all the three algorithms were close to each other in terms of performance but the Decision Tree (Accuracy: 99.78%) and Random Forest (Accuracy: 99.99%) performed slightly well than the Logistic Regression (Accuracy: 94.41 %).

**Why Random Forest is preferred over Logistic Regression?**

Logistic Regression is used widely for solving Industry scale problems as it’s easy to implement and it generally doesn't give you discrete output and gives Probability associated with each output. The logistic regression algorithm is robust to a small noise in the data and is not particularly affected by small cases of multi-collinearity. We can use many metrics for evaluation but let’s consider Recall for True Positive and F1 score (Harmonic Mean for Recall and Precision)

Random Forest Classifier is more of Accuracy focused algorithm and is best till it’s used with proper fit, else it gets overfat quickly. Random selection in individual decision trees of RFC can capture more complex feature patterns to provide the best accuracy. RFC can also tell us how much each feature contributes to class prediction using Feature Importance and tree graphs for better interpretation.

Overall saying Random Forest Classifier performs better with more categorical data than numeric and logistic regression is a little confusing when comes to categorical data So. If the dataset has more Categorical data and consists of outliers it is better to use Random Forest Classifier.

At last the model Predicted in terms of 0 and 1 whether the transaction is fraud or not.

**Chapter 5**

**Conclusion and Future Work**

From a proper analysis of positive points and constraints on the member, it can be safely concluded that the product is a considerably productive member. This use is working duly and meeting to all Banker requisites. This member can be freely plugged in numerous other systems. There have been mathematics cases of computer glitches, violations in content and most important weight of features is fixed in automated prophecy system, so in the near future the so – called software could be made more secure, trustworthy and dynamic weight conformation. In near future this module of prophecy can be integrated with the module of automated processing system

Further improvement may be done in this using Neural Networks. The system is trained on old training dataset in future software can be made resembling that new testing date should also take part in training data after some fix time.

**References:**

[1] Credit Card Fraud Detection Dataset. https://www.kaggle.com/datasets/kartik2112/fraud-detection

[2] Random Forest Tutotial from towardsdatascience. https://towardsdatascience.com/understanding-random-forest-58381e0602d2?gi=dcd300b32327

[3] Geeks For Geeks and Youtube support.