**[P-1]**

**A PRELIMINARY REPORT ON**

**CREDIT CARD FRAUD DETECTION**

SUBMITTED TO THE SAVITRIBAI PHULE PUNE UNIVERSITY, PUNE

IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS

OF

**THIRD YEAR COMPUTER ENGINEERING**

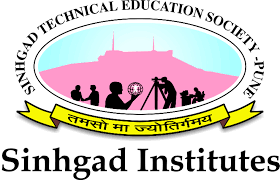
##### SUBMITTED BY

**STUDENT NAME ROLL NO:**

**NAMAN SHARMA 130**

**ADITI KHEDKAR 116**

**SONAM MAHAJAN 110**



## DEPARTMENT OF COMPUTER ENGINEERING

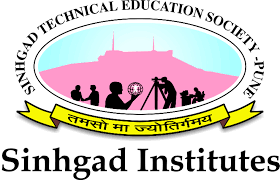
## STES’S SMT. KASHIBAINAVALCOLLEGE OF ENGINEERING

**VADGAON BK, OFF SINHGAD ROAD, PUNE 411041**

## [P-2]

## SAVITRIBAI PHULE PUNE UNIVERSITY

## 2019 -2020



**CERTIFICATE**

This is to certify that the project report entitles

**CREDIT CARD FRAUD DETECTION**

Submitted by

**STUDENT NAME ROLL NO:**

**NAMAN SHARMA 130**

**ADITI KHEDKAR 116**

**SONAM MAHAJAN 110**

is a bonafide work carried out by him/her under the supervision of **Prof. S. P. Patil** and it is approved for the partial fulfillment of the requirement of third year computer engineering.

**(Prof. S. P. Patil)** **(Dr. P. N. Mahalle)**

Guide, Head,

Department of Department of

Computer Engineering Computer Engineering

**(Dr. A. V. Deshpande)**

Principal,

Smt. KashibaiNavaleCollege of Engineering Pune – 41

Place: Pune Date: dd/mm/yyyy

**ACKNOWLEDGMENT**

We would like to express our special thanks of gratitude to our HOD, Dr. P. N. Mahalle sir and our project guide Prof. S. P. Patil sir**,** who gave us the opportunity to this wonderful project of SKILL DEVELOPMENT LAB on the topic CREDIT CARD FRAUD DETECTION.

This would not have completed without his enormous help & worthy experience. Whenever we were in need, he was there behind us. And we came to know about so many new things and we are really thankful to him.

Also, we would like to thank our friends who helped us a lot in finalizing this project within the limited time frame

**NAME OF THE STUDENTS:**

**NAMAN SHARMA  
ADITI KHEDKAR  
SONAM MAHAJAN**

**CONTENTS**

Sr.No. Chapters Page No.

1 **Abstract** 5

2 **Problem Definition** 6

3 **Motivation** 6

4 **Software Requirements specification**

4.1 Introduction 7

4.2 Project Scope 7

4.3 Functional Requirements 9

4.3.1 Libraries Used

4.3.2 Machine Learning

4.3.3 Types of Machine Learning Algorithms

4.3.4 Logistic Regression Algorithm

4.3.5 Random Forest Algorithm

4.3.6 Performance Measurement Metrics

4.4 System Requirements

4.4.1 Database Requirements 15

4.4.2 Software Requirements 15

5 **Other Specifications** 17

5.1 Advantages 17

5.2 Disadvantages 17

5.3 Applications 8

6 **Results & Discussion**

6.1 Screenshots 18

* 1. Conclusion 20

1. **ABSTRACT**

Online transactions have become an integral part of human life. It is used, not only for shopping online on websites like Amazon and Flipkart but has also become extremely prevalent in day to day transactions. Since the government of India has launched the BHIM UPI initiative a lot of transactions are now being done online using apps like Paytm or Google Pay instead of cash.

One extremely common method of online payment is through a credit card. Because a credit card offers various useful features such as one-click online payment and the ability to pay using EMIs it is becoming more and more common for people to use it for payments. Additionally, credit cards are not limited to online transactions and can be used to pay for things in a physical store as well. It is a convenient way of getting services and paying for them without any hassles.

Although credit cards are extremely popular, they do have some downsides as do all other methods online payment. In the same way that cash is susceptible to theft, credit cards are subject to financial fraud.

Credit card fraud is an epidemic that must be solved as it causes losses to the tune of billions of dollars to businesses, banks and other financial institutions. This is why it is important to find ways to detect credit card fraud.

In our project we try to detect fraudulent credit card transactions with the help of competent machine learning algorithms by studying large datasets of credit card transactions.

These algorithms such as Logistic Regression, Random Forest Algorithm allow the computer to decipher patterns which might be able to predict whether a certain transaction is fraudulent or not.

Once we train these Machine Learning Models and optimize programs to safely detect and block fraudulent transactions it is possible to save billions of dollars in losses each year in lost revenue.

Hence the proposed software program is not only extremely useful but also a necessity in today’s modern world where a large majority of transactions occur online.

1. **PROBLEM DEFINITION**

The goal is to deploy a software that creates models using two popularand efficient machine learning algorithms namely Logistic Regression and Random Forest and use it to predict whether a transaction is fraudulent or not by training the model with existing data.

1. **MOTIVATION**

The goal of this project is to develop a program that can help financial institutions or banks to save lost revenue through fraud. The program can test and classify fraudulent credit card transactions using Machine Learning Algorithms to train a ML model.

The program uses two very competent and popular algorithms for its ML model namely -

1. Logistic Regression Algorithm
2. Random Forest Algorithm

**4. SOFTWARE REQUIREMENTS AND SPECIFICATION**

**4.1**. **Introduction**

Online transactions have become an integral part of human life. It is used, not only for shopping online on websites like Amazon and Flip kart but has also become extremely prevalent in day to day transactions. Since the government of India has launched the BHIM UPI initiative a lot of transactions are now being done online using apps like Paytm or Google Pay instead of cash.

One extremely common method of online payment is through a credit card. Because a credit card offers various useful features such as one-click online payment and the ability to pay using EMIs it is becoming more and more common for people to use it for payments. Additionally, credit cards are not limited to online transactions and can be used to pay for things in a physical store as well. It is a convenient way of getting services and paying for them without any hassles.

Although credit cards are extremely popular, they do have some downsides as do all other methods online payment. In the same way that cash is susceptible to theft, credit cards are subject to financial fraud.

Credit card fraud is an epidemic that must be solved as it causes losses to the tune of billions of dollars to businesses, banks and other financial institutions. This is why it is important to find ways to detect credit card fraud.

**4.2. Project Scope**

Our program helps to stop online thieves and detect their nefarious activities that can cause losses to consumers and organizations alike. It helps to keep integrity of data and helps to reduce effort and time spent to manually classify whether a transaction is fraudulent or not.

**Why to use machine learning algorithms?**

Majority of businesses use manual reviews and checks to avoid fraud and process the integrity of a transaction. Involvement of humans in this process gives valuable insights about fraudsters and genuine customer behavior. But the manual review is costly, and leads to high false negatives. This is not only inaccurate as humans are emotional beings and cannot guarantee accuracy when it comes to highly logical tasks such as detecting patterns. But it is also extremely inefficient and time consuming.

Machine Learning helps avoid these obstacles and provide a reliable process for fraud detection that delivers high performance and accuracy every single time.

**4.3. Functional Requirements**

**4.3.1 Libraries Used**

**• NumPy**

Numpy is a general-purpose array-processing package. It provides a high-performance multidimensional array object, and tools for working with these arrays. It is the fundamental package for scientific computing with Python.

Besides its obvious scientific uses, Numpy can also be used as an efficient multi-dimensional container of generic data.

Array in Numpy is a table of elements (usually numbers), all of the same type, indexed by a tuple of positive integers. In Numpy, number of dimensions of the array is called rank of the array. A tuple of integers giving the size of the array along each dimension is known as shape of the array. An array class in Numpy is called as n- dimensional array. Elements in Numpy arrays are accessed by using square brackets and can be initialized by using nested Python Lists.

**• Pandas**

Pandas is a software library written for the Python programming language for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series. It is free software released under the three-clause BSD license. The name is derived from the term "panel data", an econometric term for data sets that include observations over multiple time periods for the same individuals.

**• SciKit Learn**

Scikit-learn is a free software machine learning library for the Python programming language. It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to operate with the Python numerical and scientific libraries NumPy and SciPy.

**4.3.2 Machine Learning**

Machine Learning is a sub-area of artificial intelligence, whereby the term refers to the ability of IT systems to independently find solutions to problems by recognizing patterns in databases. In other words: Machine Learning enables IT systems to recognize patterns on the basis of existing algorithms and data sets and to develop adequate solution concepts. Therefore, in Machine Learning, artificial knowledge is generated on the basis of experience.

In order to enable the software to independently generate solutions, the prior action of people is necessary. For example, the required algorithms and data must be fed into the systems in advance and the respective analysis rules for the recognition of patterns in the data stock must be defined. Once these two steps have been completed, the system can perform the following tasks by Machine Learning:

* Finding, extracting and summarizing relevant data
* Making predictions based on the analysis data
* Calculating probabilities for specific results
* Adapting to certain developments autonomously
* Optimizing processes based on recognized patterns

• **Machine Learning: How it works**

In a way, Machine Learning works in a similar way to human learning. For example, if a child is shown images with specific objects on them, they can learn to identify and differentiate between them. Machine Learning works in the same way: Through data input and certain commands, the computer is enabled to "learn" to identify certain objects (persons, objects, etc.) and to distinguish between them. For this purpose, the software is supplied with data and trained. For instance, the programmer can tell the system that a particular object is a human being (="human") and another object is not a human being (="no human"). The software receives continuous feedback from the programmer. These feedback signals are used by the algorithm to adapt and optimize the model. With each new data set fed into the system, the model is further optimized so that it can clearly distinguish between "humans" and "non-humans" in the end.

**4.3.3 Types of Machine Learning Algorithms**

Machine Learning algorithms can be classified into 3 different types, namely:

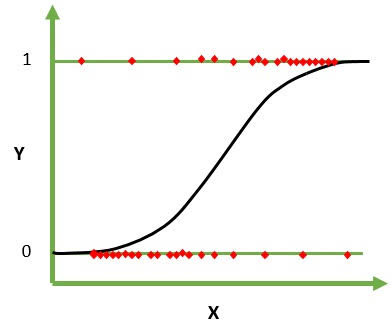
**• Supervised Machine Learning Algorithms:**  
Imagine a teacher supervising a class. The teacher already knows the correct answers but the learning process doesn’t stop until the students learn the answers as well (poor kids!). This is the essence of Supervised Machine Learning Algorithms. Here, the algorithm is the student that learns from a training dataset and makes predictions that are corrected by the teacher. This learning process continues until the algorithm achieves the required level of performance.

**• Unsupervised Machine Learning Algorithms:**  
In this case, there is no teacher for the class and the poor students are left to learn for themselves! This means that for Unsupervised Machine Learning Algorithms, there is no specific answer to be learned and there is no teacher. The algorithm is left unsupervised to find the underlying structure in the data in order to learn more and more about the data itself.

**• Reinforcement Machine Learning Algorithms:**  
Well, here are hypothetical students learn from their own mistakes over time (that’s like life!). So the Reinforcement Machine Learning Algorithms learn optimal actions through trial and error. This means that the algorithm decides the next action by learning behaviors that are based on its current state and that will maximize the reward in the future.

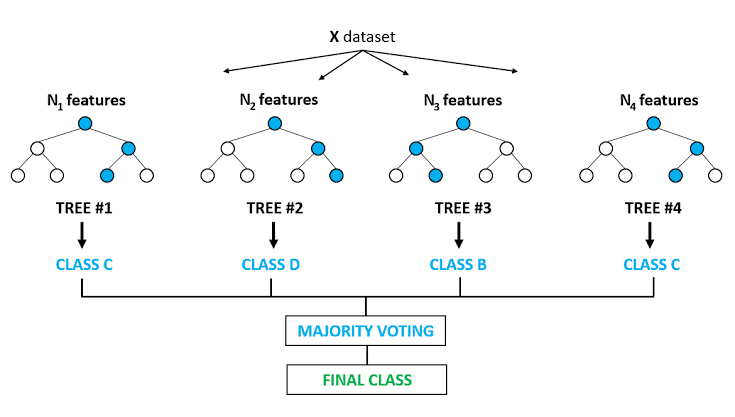
**4.3.4 Logistic Regression Algorithm**

The Logistic Regression Algorithm deals in discrete values whereas the Linear Regression Algorithm handles predictions in continuous values. So, Logistic Regression is suited for binary classification wherein if an event occurs, it is classified as 1 and if not, it is classified as 0. Hence, the probability of a particular event occurrence is predicted based on the given predictor variables. An example of the Logistic Regression Algorithm usage is in politics to predict if a particular candidate will win or lose a political election.



**4.3.5 Random Forest Algorithm**

The Random Forests Algorithm handles some of the limitations of Decision Trees Algorithm, namely that the accuracy of the outcome decreases when the number of decisions in the tree increases. So, in the Random Forests Algorithm, there are multiple decision trees that represent various statistical probabilities. All of these trees are mapped to a single tree known as the CART model. (Classification and Regression Trees). In the end, the final prediction for the Random Forests Algorithm is obtained by polling the results of all the decision trees. An example of the Random Forests Algorithm usage is in the automobile industry to predict the future breakdown of any particular automobile part.



**4.3.6 Performance Measurement Metrics**

**• Confusion Matrix**

In the field of machine learning and specifically the problem of statistical classification, a **confusion matrix**, also known as an error matrix, is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a **matching matrix**). Each row of the matrix represents the instances in a predicted class while each column represents the instances in an actual class (or vice versa). The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

It is a special kind of contingency table, with two dimensions ("actual" and "predicted"), and identical sets of "classes" in both dimensions (each combination of dimension and class is a variable in the contingency table).

• **Cohen Kappa Score**

Cohen's kappa coefficient is a statistic which measures inter-rater agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than simple percent agreement calculation, since k takes into account the agreement occurring by chance. Cohen's kappa measures the agreement between two raters who each classify N items into C mutually exclusive categories.

**• F1 Score**

In statistical analysis of binary classification, the F1 score (also F-score or F-measure) is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score: p is the number of correct positive results divided by the number of all positive results returned by the classifier, and r is the number of correct positive results divided by the number of all relevant samples (all samples that should have been identified as positive). The F1 score is the harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0.

• **Recall Score**

In pattern recognition, information retrieval and binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance.

**4.4 SYSTEM REQUIREMENTS**

**4.4.1 Database Requirements :**

As our project is based entirely on data science and processing a dataset with machine learning algorithms, there are no explicit database requirements as a database is not being used.

**4.4.2 Software Requirements :**

Following are the softwares used for the program.

**Software : Android Studio**

We have used Spyder 3.3.6 for the application development because Spyder IDE is officially integrated with Anaconda Navigator. This environment runs on a base installation of Anaconda Navigator and uses the latest version of Python 3 available at the time of writing this report which is Python 3.7.3. Spyder is an open source cross-platform integrated development environment for scientific programming in the Python language.

**Operating System :**

We have tested the program on two operating systems namely Windows 10 (64-bit) and Linux Mint 18.1 LTS (64-bit) which is a Linux distribution similar to Ubuntu and is derived from Debian. The softwares and libraries used are also compatible with almost all versions of Linux, Windows and macOS.

Following are the system requirements for Spyder 3.7..x

**Operating System :**

Windows 7 or newer, 64-bit macOS 10.10+, or Linux, including Ubuntu, Red Hat, CentOS 6+, and others.

**System Architecture :**

Windows- 64-bit x86, 32-bit x86; macOS- 64-bit x86; Linux- 64-bit x86, 64-bit Power8/Power9.

**Disk Space :**

Minimum 5 GB disk space to download and install.

**Python Version :**

Python 3 is used.

**Libraries Used :**

NumPy, Pandas, SciKit Learn are the libraries used and must be installed before the program can be executed.

**5. OTHER SPECIFICATIONS**

**8.1. Advantages**

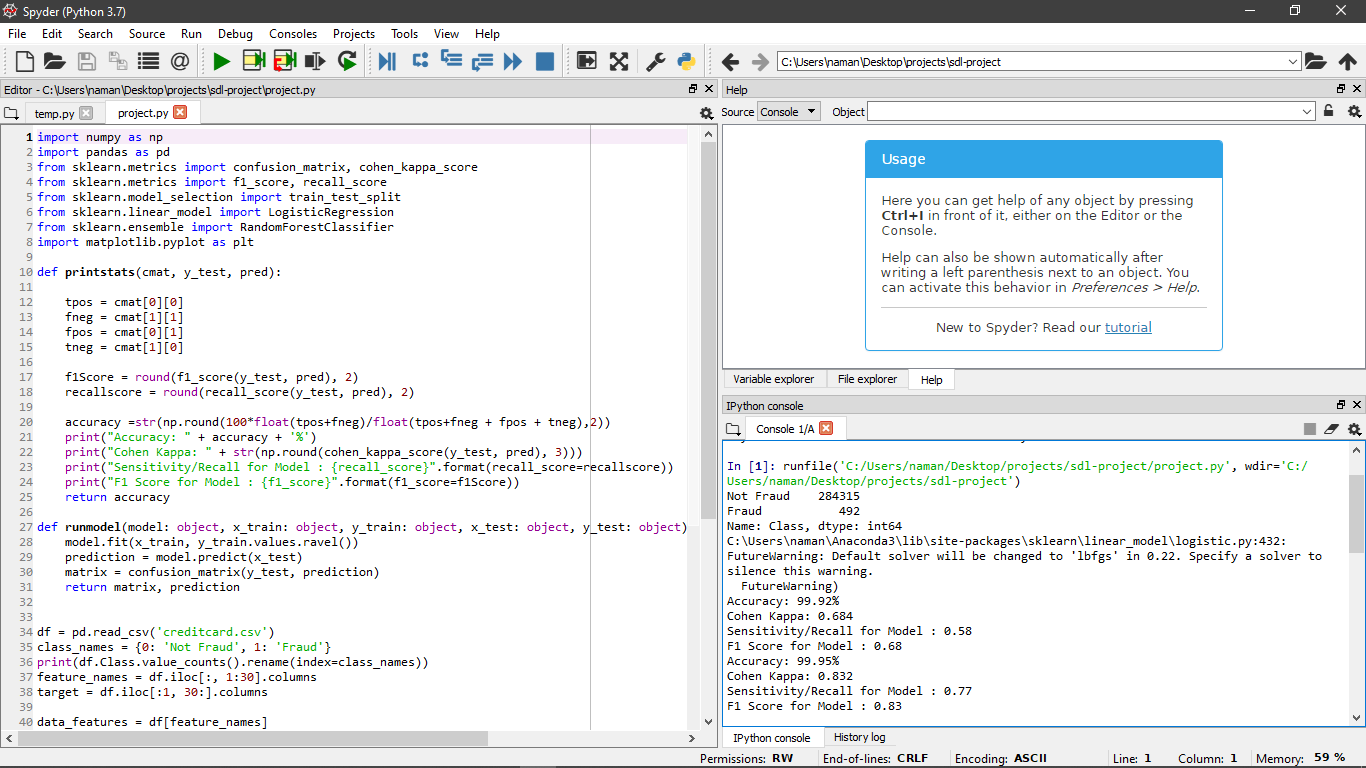
* One of the most notable advantages is reduction in human effort
* Extremely efficient when compared to manual review
* Very cost effective
* Helps regain lost revenue and prevent fraud
* Much higher accuracy as compared to manual review and detection

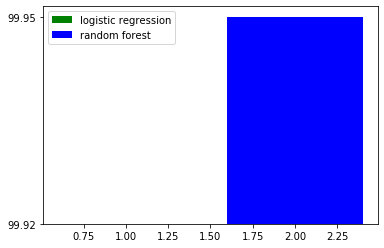
**8.2. Disadvantages:**

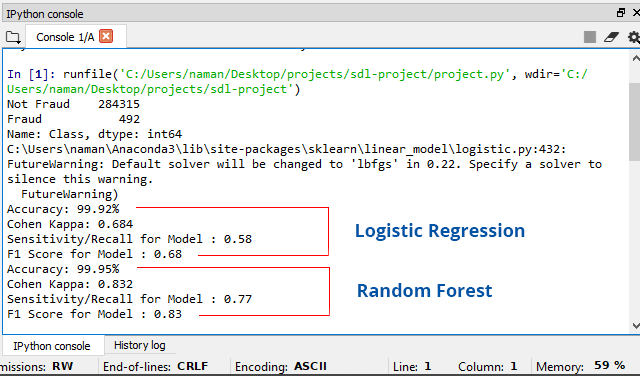
* There is a requirement of hardware in the form of computers.
* Skilled labor is required to use and maintain the program.
* Requires skill, effort and time for initial implementation.

**6. RESULTS AND DISCUSSION**

**6.1 Screenshots**







**6.2. Conclusion :**

Online payment methods and services make the lives of the consumer very easy and solve problems created by the use of cash. Majority of the transactions that occur in today’s modern world are online. Implementation of this project helps classify fraudulent transactions and recovers lost revenue. It is essential to avoid and block fraudulent transactions as online transactions are more susceptible to fraud than physical hard cash. However, it becomes worth the trade-off if implemented correctly because cash has some downsides and can be very slow, inefficient and time consuming.

Hence classifying fraud transactions and preventing loss of revenue becomes faster, more reliable and more efficient than manual review.