

# Credit Card Fraud Detection Using Machine Learning

Srinivasa Rao Dammavalam, Md Mukheed

*Department of IT  
VNRVJIET  
Hyderabad, India*

Date of Submission: 01-01-2023

Date of Acceptance: 08-01-2023

## ABSTRACT:

Credit risk is one of the main functions of banking. Banks classify risk according to their profile. Although many algorithms have emerged, there is still a problem to solve. Non-existence, data normalization is applied before Cluster Analysis, and obtained results from Cluster Analysis and Artificial Neural Networks related to fraud detection showed attribute clustering and neural inputs can be minimized. The importance of the work lies in finding an algorithm to reduce costs. The result obtained was 23% and the algorithm used was Minimum Bayesian-Risk (MBR). In the proposed system, Random Forest Algorithm is used for classification and regression. A random forest has an advantage over a decision tree because it corrects the habit of overfitting its training datasets. It is found to provide a good estimate of the generalization error and is robust to overfitting. In credit card fraud detection, credit card datasets are collected for training datasets and user credit card inquiries are collected for testing datasets. After the classification process, the Random Forest Algorithm is used to analyze the datasets and current datasets. Finally, optimization is done, and the accuracy obtained by Random Forest is 99.9 percent.

**Keywords:** Random Forest, Decision Tree, Plan Neutral Network, Weighted Neutral Network, Python.

## I. INTRODUCTION

Billions in losses are resulting from fraudulent credit score card transactions every yr. Fraud is as vintage as mankind itself and can take an unlimited variety of various forms. percent's 2017 global financial Crime Survey shows that approximately 48% of businesses have experienced monetary crime [3]. consequently, there is simply a requirement to address the issue of credit score card fraud detection. using credit scorecards is well-known in current society and credit card fraud has persevered in growth in the latest years [2]. good sized financial losses due to fraud have affected now not most

effective merchants and banks, but also men and women folks that use credit scores. Fraud can also affect a trader's recognition and photograph and cause non-economic losses which, although difficult to quantify within a short time, may be seen over a long time [4]. for example, if a cardholder turns into a sufferer of fraud with a precise business enterprise, the quantity no longer trusts their enterprise and opts for a competitor

The improvement of the financial system and the open economic market make the credit card business one of the maximum critical earnings of the bank. however, on side of the increase in issuance quantity, international credit fraud transactions are growing at an alarming price. economic agencies can't efficaciously discover fraudulent transactions; as a result, the loss will become increasingly critical. the way to perceive fraudulent credit card transactions efficaciously, fast, and appropriately is turning into a problem of trendy interest. In China, we're begun to apply online credit score cards to pay in recent years. The related take look split into two directions: fraudulent identity and organization packages.

The research's primary direction is Tong Fengru, which is primarily based on a mixture of classifiers and Yan Hua, Hu Mengliang, which uses a Bayesian category algorithm. within the second case, a 3rd-birthday party charge service provider – IPS has formally launched an anti-credit vehicle connection with credit card payment systems, effective inhibition of electronic credit card payment in opposition to diverse risks that could arise at some stage in transactions. Early research on credit card fraud prevention makes a specialty of type and identification strategies and fashions, which includes man or woman sample popularity methods along with decision timber and neural networks, combinatorial methods, and distributed facts mining. but, due to the complexity and sparseness of transaction records, these techniques often face the trouble of version choice, model parameter placing, and wrong selection when running with big transaction statistics, which

frequently results in debt look at, overfitting, and local most reliable trouble [2]. guide vector machine is a particularly new area in statistics mining. The system first maps the facts from the input area to the feature area and then constructs a linear discriminant function inside the characteristic area. although there are many similarities between a neural network and a guide vector system in shape, the latter is surprisingly simple.

#### Application:

The financial marketplace makes the credit card enterprise one of the financial institution's maximum essential sales. However, alongside the growth in issuance quantity, international credit score fraud transactions are increasing at an alarming rate. Economic corporations cannot efficaciously detect fraudulent transactions.

#### Motivation:

To discover fraudulent credit score card transactions, in this paper, we advise an optimized SVM version for online credit card fraud detection.

#### Problem Statement:

Credit card bills have elevated in recent years. it can be used for online and regular shopping. credit score card payments are necessary and handy today. because of the growth in fraudulent transactions, there may be a want to find a powerful fraud detection fashions fraud device called ANT [1]. The system uses a parallel anti-fraud model based on a neural community. a parallel anti-fraud model based on a neural community.

#### Objectives:

1. To study and analyze various credit score card fraud detection strategies.
  2. layout a brand-new statistics mining-based credit score card fraud detection using guide Vector Machines.
  - three. Use incremental mastering approach to reduce misclassification rate and false alarm technology.
  4. examine the proposed technique using numerous input and output parameters together with type errors, accuracy, and false alarms.
- of SVM and decision tree methods in credit card fraud detection against a real data set.

## II. RELATED WORK

Literature gives an overview of previous related works in the current domain. The Literature review can bring focus to the area of research and broaden your knowledge of the domain.

[1] the usage of predictive analytics technology to detect credit score card fraud in Canada. "Kosemani Temitayo Hafiz, Dr. Shaun Aghili, Dr. Pavol Zavorsky

This research paper focuses on creating a scorecard of the relevant assessment criteria, features, and capabilities of predictive analytics solutions currently used to detect credit card fraud. The Scorecard provides a side-by-side comparison of five predictive analytics solutions from credit card vendors accepted in Canada. From the following Based on the research results, a list of PAT card fraud issues, risks, and limitations were outlined.

[2] BLAST-SSAHA hybridization for credit card fraud detection. "Amlan Kundu, Suvasini Panerai, Shamil Sural, Senior Fellow, IEEE, and Arun K. Majumdar "

This paper proposes the use of a two-stage sequence alignment in which a profile analyzer (PA) first determines the similarity of an incoming sequence of transactions on a given credit card to the past spending sequences of the actual cardholder. Unusual transactions traced by the profile analyzer are further passed to the anomaly analyzer (DA) for possible matching with past fraudulent behavior. The final decision on the nature of the transaction is made based on the observations of these two analyzers. To achieve online response time for both PA and DA, we propose a new approach to combine the two sequence alignment algorithms BLAST and SSAHA

[3] Research on a Sum-of-Distance Credit Card Fraud Detection Model.

Along with the increasing volume of credit cards and the increasing volume of trade in China, the number of credit card frauds is skyrocketing. How to improve the detection and prevention of credit card fraud is becoming a focus of risk bank control. It proposes a model for credit card fraud detection using outlier detection based on the sum of distance according to the rarity and unusualness of fraud in credit card transaction data, applying outlier mining to credit card fraud detection. Experiments show that this model is feasible and accurate in detecting credit card fraud.

[4] Fraud detection in credit card system using SVM and decision tree. "Vijayshree B. Nipane, Poonam S. Kalinga, Dipali Vidhate, Kunal War, Bhagyashree P. Deshpande".

With the increasing advancement of e-commerce, frauds are spreading all over the world and causing huge financial losses. In the current scenario, the major cause of financial loss is credit card fraud; it does not only apply to traders but also to individual clients. Decision Tree, Genetic Algorithm, Meta-learning strategy, Neural Network, and HMM are presented methods used to detect credit card fraud. In

the considered fraud detection system, the artificial intelligence concept Support Vector Machine (SVM) and decision tree are used to solve the problem. Thus, financial losses can be reduced to a greater extent by implementing this hybrid approach.

[5] Supervised Machine Learning (SVM) for Credit Card Fraud Detection. "Sitaram patela, Sunita Gond.

This work proposes a method based on SVM (Support Vector Machine) with a multi-joining kernel, which also includes several user profiles fields instead of only an expenditure profile. The simulation result shows an improvement in the TP (True Positive), and TN (True Negative) rates and reduces the FP (False Positive) and FN (False Negative) rates.

[6] Credit card fraud detection using decision trees and support vector machines. "Y. Sahin and E. Duman"

In this study, classification models based on decision trees and support vector machines (SVM) are developed and applied to the problem of credit card fraud detection. This study is one of the first to compare the performance of SVM and decision tree methods in credit card fraud detection against a real data set.

Descent (GD) is a cost-minimization technique that examines the coefficients of a function (f). It is a key optimization approach for determining the minimal cost function. The model may be conveniently stored in memory with little noise using the GD technique. Computational linguistic rule-based human language modeling is combined with statistical, deep learning models, and machine learning in NLP. These technologies work together to allow computers to analyze human language in the form of text or speech data and comprehend its entire meaning, including the speaker's or writer's purpose and mood. This chatbot answers questions about hospital information, such as specialist availability, OPD hours, room registration, bed capacity, doctor availability, and emergency information, among other things. The suggested chatbot acts as if it were a genuine hospital receptionist,

#### **Machine Learning Tasks typically fall into several broad categories:**

**Supervised learning:** The computer is presented with examples of inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs. In special cases, the input signal may be only partially available or limited to special feedback.

**Partially supervised learning:** The computer is given only an incomplete training signal: a training set with some (often many) target outputs missing.

**Active learning:** The computer can only obtain training labels for a limited set of instances

(based on the budget) and must also optimize its selection of objects for which to obtain labels. When used interactively, they can be presented to the user for marking.

**Unsupervised learning:** No labels are given to the learning algorithm, so it is left on its own to find structure in its input. Unsupervised learning can be an end (discovering hidden patterns in data) or a means to an end (feature learning).

**Reinforcement learning:** Data (in the form of rewards and punishments) is provided only as feedback on the program's actions in a dynamic environment, such as driving a vehicle or playing a game against an opponent.

The main objective of this programming language is as follows:

- Python is a simple, object-oriented programming language.
- The language and implementation should provide support for software engineering principles such as pre-set strong type library for various machine learning algorithms and all other algorithms in a simple way.
- Coding will be smooth in python and data analysis can be easily done in python.

So much so that now we have modules and APIs available, and you can very easily get involved in machine learning with almost no knowledge of how it works. With default values from Scikit-learn, you can get 90-95% accuracy on many tasks right out of the gate. Machine learning is a lot like a car, you don't need to know much about how it works to get an incredible amount of benefit from it.

Despite the apparent age and maturity of machine learning, I'd say there's no better time than now to learn it because you can use it. The machines are quite powerful, the one you're working on can probably handle most of this series quickly. There is also a lot of data lately.

In the paper by LekhaAthota[5], N-gram, which is a series of N words, is used to construct the chatbot application. So, for example, "Final demo" is a 2-gram (a bigram), "This is a final demo" is a 4-gram, and "Good to go" is a 3-gram (trigram). The TF-IDF (Term Frequency-Inverse Document Frequency) works by examining whether the word belongs to a document in a large collection of documents. This can be examined by multiplying two metrics: the word's inverse document frequency over a collection of documents and the number of times a word occurs in a document. It's used to get the keyword out of the user query. To get the best response for the inquiry, each term is weighted down. The Web-interface is designed for users to enter their queries. The program is enhanced with security and effectiveness modifications that ensure user protection and integrity

when getting answers to queries. This chatbot assists users with basic health information. When a person initially visits the website, they must register before asking the questions to the chatbot. If the answer is not available in the database, the system employs an expert system to respond to the queries.

In the paper by Dammavalam Srinivasa Rao [6], an AI chatbot for college activities is developed using Deep Neural networks. The data regarding college activities is being collected in the JSON format and the Bag of word technique is used in preprocessing of data. Gradient Descent is used for optimizing the model to process the patterns and give the best possible response to the question asked by the user. pyttsx3 python library is used for speech recognition to enable users to give input questions using voice. The model accuracy is found to be around 93 percent for 1200 epochs of training the model.

### III. PROPOSED SYSTEM

In the proposed system, we use a random forest algorithm to classify the credit card data set. Random Forest is a classification and regression algorithm. In short, it is a collection of decision tree classifiers. A random forest has an advantage over a decision tree because it corrects the habit of overfitting its training set. A subset of the training set is randomly selected to train everyone

And then a decision tree is created, and each node is then split into an element selected from a random subset of the full set of elements. Even for large datasets with many features and data instances, training is extremely fast in a random forest because each tree is trained independently of the others. The Random Forest algorithm was found to provide a good estimate of the generalization error and to be robust to overfitting.

#### Advantages of the proposed system:

- Random Forest evaluates the importance of variables in a regression or classification problem in a natural way that can be done using Random Forest.
- The "amount" function is the amount of the transaction. The 'class' element is the target class for binary classification and has a value of 1 for the positive case (fraud) and 0 for the negative case (not fraud).

### SYSTEM DESIGN

#### Introduction

The system design document describes the system requirements, operating environment, system and subsystem architecture, file and database design, input formats, output layout, human-machine

interface, detailed design, processing logic, and external interfaces.

#### Dataset:

The data used in this paper is a collection of product reviews collected from credit card transaction records. This step is about selecting a subset of all available data to work with. ML problems are best started with data, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labeled data.

#### Pre-Processing:

- Organize your selected data by formatting, cleaning, and sampling from it. Three common steps in data preprocessing are Formatting: The data you have selected may not be in a format that is convenient for you. The data may be in a relational database, and you would like it in a flat file, or the data may be in a proprietary file format, and you would like it in a relational database or text file. Cleaning: Data cleaning is the removal or correction of missing data. There may be instances of data that are incomplete and do not carry the data you think you need to solve the problem. These instances may need to be removed. In addition, some attributes may contain sensitive information and these attributes may need to be completely removed from the data. Sampling: There ma

#### Feature Extraction:

Another thing to do is Feature extraction is a process of attribute reduction. Unlike feature selection, which ranks existing attributes according to their predictive value, feature extraction transforms the attributes. Transformed attributes or elements are linear combinations of the original attributes. Finally, our models are trained using the Classifier algorithm. We use the classification module in the Natural Language Toolkit library on Python. We use a collected labeled data set. The rest of our labeled data will be used to evaluate the models. Some machine learning algorithms were used to classify the pre-processed data. The chosen classifiers were random forests. These algorithms are very popular in text classification tasks.

#### Evaluation Model:

Model evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the selected model will perform in the future. Evaluating a model's performance with the training data is not acceptable.

**Algorithm:**

**Random Forest:**

A random forest is a type of supervised machine-learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you combine different types of algorithms or the same algorithm multiple times to create a more powerful prediction model. A random forest algorithm combines multiple algorithms of the same type, i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks

**Working on random forest**

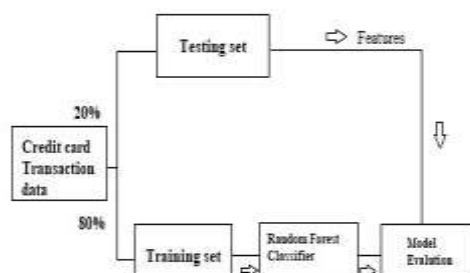
The following are the basic steps involved in implementing a random forest algorithm

1. Select N random records from the dataset.
2. Create a decision tree based on these N records.
3. Select the desired number of trees in your algorithm and repeat steps 1 and 2.
4. For a classification problem, each tree in the forest predicts the category to which the new record belongs. Finally, the new record is assigned to the category that receives the most votes.

**System Architecture**

categorical and numeric features. The random forest algorithm also works well when the data has missing values or has not been scaled well.

A system architecture or systems architecture is a conceptual model that defines a system's structure, behavior, and other views. An architecture description is a formal description and representation of a system. Organized in a way that encourages thinking about system structures and behavior.



**3-Tier Architecture:**

Three-tier software architecture (three-tier architecture) emerged in the 1990s to overcome the limitations of two-tier architecture. The third layer (middle layer server) is between the user interface (client) and the data management components (server). This middle layer provides process

management where business logic and rules are executed and can accommodate hundreds of users (compared to only 100 users with a two-tier architecture) by providing functions such as queuing, application startup, and database preparation.

A three-tier architecture is used when an efficient distributed client/server design is needed that provides (compared to a two-tier) increased performance, flexibility, maintainability, reusability, and scalability while hiding the complexity of distributed processing from the user. These characteristics have made three-tier architectures a popular choice for Internet applications and network-oriented information systems. Can be much faster for exploring and prototyping solutions than considering the entire data set.

**IV. IMPLEMENTATION**

Conduct studies and analyzes of an operational and technological nature and support the exchange and development of methods and tools for operational analysis in solving defense problems.

**Inputs and outputs designs:**

**Logical proposal**

Logical system design refers to the abstract representation of data flows, inputs, and outputs of the system. This is often done through modeling using an overly abstract (and sometimes graphical) model of the actual system. In the context of systems design, they are included. The logical design includes ER diagrams, i.e., entity relationship diagrams

**Physical design**

Physical design refers to the actual input and output processes of the system. This is determined by how data is entered into the system, how it is verified/authenticated, how it is processed, and how it is displayed as output. The following system requirements are decided in the physical design.

1. Entry Requirement,
2. Output requirements,
3. Storage Requirements,
4. Processing requirements,
5. System management and backup or recovery.

In other words, the physical part of systems design can generally be divided into three sub-tasks:

1. User interface design
2. Data design
3. Process design

User interface design deals with how users add information to the system and how the system provides information back to them. Data design deals with how data is represented and stored in a system. Finally, Process Design addresses how data moves through the system and how and where it is

authenticated, secured, and/or transformed as it flows into, through, and out of the system. At the end of the systems design phase, documentation describing the three subtasks is created and made available for use in the next phase.

Physical design in this context does not refer to the material physical design of the information system. To use an analogy, the physical design of a personal computer includes input through the keyboard, processing in the CPU, and output through the monitor, printer, etc. This would not refer to the actual layout of the tangible hardware, which for a PC would be the monitor, CPU, motherboard, hard drive, modems, video/graphics cards, USB slots, etc. This is a detailed design of the user and product database processor and control processor. A personal H/S specification is developed for the proposed system.

### Entry and exit representation Entry design

The input design is the connecting link between the information system and the user. It includes the evolving specification and procedures for data preparation, and these steps are necessary to put transaction data into a usable form for processing, which can be achieved by controlling a computer to read the data from a written or printed document, or this can happen through human keying. data directly into the system. Input design focuses on controlling the amount of input required, controlling errors, avoiding delays, avoiding redundant steps, and keeping the process simple. The entrance is designed to provide security and ease of use while maintaining privacy. The draft entry considered the following:

- What data should be given as input?
- How should the data be organized or coded?
- Dialog to guide service personnel when providing input.
- Methods for preparing input validation and steps to follow when an error occurs.

### Objectives

Input design is the process of converting a user-oriented description of input into a computer system. This suggestion is important to avoid errors in the data entry process and to show management the right direction to get the right information from the computer system.

This is achieved by creating user-friendly data entry screens for processing large volumes of data. The goal of input design is to facilitate data entry and avoid errors. The data entry screen is designed to allow all data manipulation. It also provides facilities for viewing records.

After entering the data, it checks its validity. Data can be entered using screens. Relevant messages

are provided as needed, so the user will not be in the corn immediately. So the goal of entry design is to create an entry layout that is easy to follow

### Output design

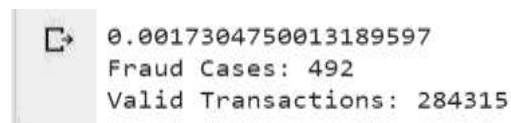
Quality output meets the requirements of the end-user and understandably presents information. In each system, the results of the processing are communicated to the users and the other system through outputs. In the design of the output, it is determined how the information is to be relocated for immediate use, as well as the output in paper form. It is the most important and direct source of information for users. Effective and intelligent output design improves system relationships and helps users make decisions.

- a. Designing computer output should be done in an organized and well-thought-out manner; the right output must be developed while ensuring that each output element is designed so that people find the system easy and effective to use. When analyzing the design of the computer output, they should determine the specific output that is needed to meet the requirements.
- b. Choose ways to present information.
- c. Create a document, report, or other formats that contain system-generated information.

## V. RESULTS

The data for this article may be discovered right here. This dataset consists of the real bank transactions made by way of EU cardholders in 2013. As a safety subject, the real variables are not being shared, however — they have been transformed into variations of PCA. As a result, we can locate 29 characteristic columns and 1 final class column.

We will try different machinelearning models one by one. Defining the model is much easier. You can define your model with a single line of code. Similarly, you can fit a model to your data with a single line of code. You can also tune these models by choosing different optimized parameters. However, if less parameter tweaking can lead to better accuracy, there is no need to complicate it.



```
0.0017304750013189597
Fraud Cases: 492
Valid Transactions: 284315
```

Only 0.17% of all transactions are fraudulent. The data are grossly imbalanced. First, let's apply the model without balancing. If the accuracy is not good, you can find a way to balance this dataset. But first, let's implement the model

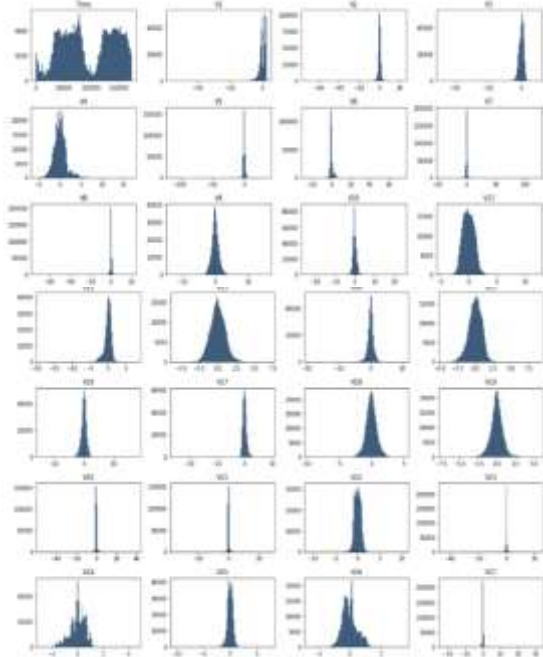
without using it, and balance the data only when necessary.

**Output screens:**

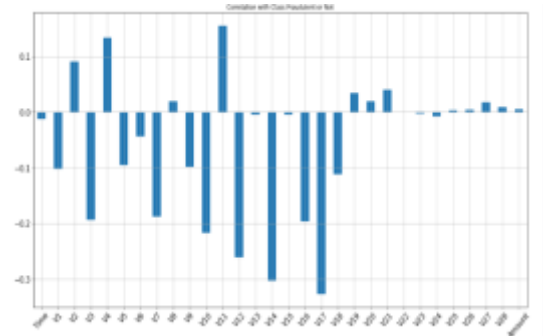


**Dataset:**

**Histograms of Numerical Values:**



**Correlation with class fraud or not:**



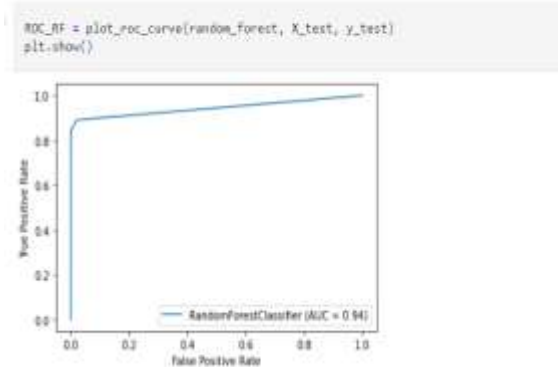
**The main challenges involved in credit card fraud detection are:**

1. With vast amounts of data processed daily, model building must be fast enough to respond to fraud in a timely manner.
2. imbalanced data, i.e. most transactions (99.8%) are not fraudulent and it is very difficult to find fraudulent ones.
3. data availability as most data is private.

4. Misclassified data can be another big problem as not all fraudulent transactions are captured and reported.
5. Adaptation techniques used by fraudsters on their models.

**Models Accuracy**

Model	Accuracy	FalseNegRate	Recall	Precision	F1 Score
0 RandomForest	0.999544	0.27440	0.77551	0.95	0.853933



Model	Accuracy	FalseNegRate	Recall	Precision	F1 Score
0 Random Forest	0.999860	0.069106	0.930894	0.987069	0.958159
1 Decision Tree	0.990860	0.073171	0.920981	0.967061	0.933470
2 XG-boost	0.998860	0.074171	0.900896	0.982069	0.944978
3 KNN	0.989860	0.070171	0.910897	0.937052	0.914894
4 SVM	0.98860	0.083171	0.890894	0.927031	0.902347

**Web APP Screenshots:**



**VI. CONCLUSION**

The Random Forest algorithm will carry out higher with greater training information however will go through in speed throughout testing and application. the usage of extra preprocessing strategies might also assist. The SVM algorithm still suffers from the imbalanced dataset trouble and calls for extra pre-processing to offer higher outcomes

within the outcomes SVM indicates are high-quality however might have been higher if greater pre-processing become achieved at the records.

## VII. FUTURE SCOPE

We haven't been able to achieve the goal of 100% fraud detection accuracy, we've ultimately created a system that, given enough time and data, can come very close to that goal. As with any such project, there is some room for improvement

The very nature of this project allows multiple algorithms to be integrated as modules and their results can be combined to increase the accuracy of the result.

This model can be further improved by adding more algorithms to it. However, the output of these algorithms must be in the same format as the others. Once this condition is met, modules can be easily added as shown in the code. This gives the project a great degree of modularity and versatility.

Further room for improvement can be found in the dataset. As shown earlier, the accuracy of the algorithms increases as the size of the data set increases. So, more data will certainly make the model more accurate in detecting fraud and reduce the number of false alarms. However, this requires official support from the banks themselves.

## REFERENCES

- [1]. Sudhamathy G: Credit Risk Analysis and Prediction Modelling of Bank Loans Using R, vol. 8, no-5, pp. 1954-1966.
- [2]. LI Changjian, HU Peng: Credit Risk Assessment for ural Credit Cooperatives based on Improved Neural Network, International Conference on Smart Grid and Electrical Automation vol. 60, no. - 3, pp 227-230, 2017.
- [3]. Wei Sun, Chen-Guang Yang, Jian-Xun Qi: Credit Risk Assessment in Commercial Banks Based On Support Vector Machines, vol.6, pp 2430-2433, 2006.
- [4]. Amlan Kundu, SuvasiniPanigrahi, Shamik Sural, Senior Member, IEEE, "BLAST-SSAHA Hybridization for Credit Card Fraud Detection", vol. 6, no. 4 pp. 309-315, 2009.
- [5]. Y. Sahin and E. Duman, "Detecting Credit Card Fraud by Decision Trees and Support Vector Machines, Proceedings of International Multi-Conference of Engineers and Computer Scientists, vol. I, 2011.
- [6]. Sitaram Patel, Sunita Gond, "Supervised Machine (SVM) Learning for Credit Card Fraud Detection, International of engineering trends and technology, vol. 8, no. -3, pp. 137- 140, 2014.
- [7]. Snehal Patil, HarshadaSomavanshi, Jyoti Gaikwad, Amruta Deshmane, Rinku Badgujar," Credit Card Fraud Detection Using Decision Tree Induction Algorithm, International Journal of Computer Science and Mobile Computing, Vol.4 Issue.4, April-2015, pg. 92-95 [8] Dahee Choi and Kyungho Lee, "Machine Learning based Approach to Financial Fraud Detection Process in Mobile Payment System", vol. 5, no. - 4, December 2017, pp. 12-24.