

Identification Of Fraudulent Credit Card Transactions Using Ensemble In Learning

Ms. Mariam, SK Shabana, R Sanjani, T Sankeerthana

¹ Assistant Professor, Department Of ECE, Bhoj Reddy Engineering College For Women, India.

^{2,3,4}B. Tech Students, Department Of ECE, Bhoj Reddy Engineering College For Women, India.

ABSTRACT

Machine learning is an emerging technique for building analytic models for machines to "learn" from data and be able to do predictive analysis. The ability of machines to "learn" and do predictive analysis is very important in this era of big data and it has a wide range of application areas. For instance, banks and financial institutions are sometimes faced with the challenge of what risk factors to consider when advancing credit/loans to customers, for several features/attributes of the customers are normally taken into consideration, but most of these features have little predictive effect on the credit worthiness or otherwise of the customer. Furthermore, a robust and effective automated bank credit risk score that can aid in the prediction of customer credit worthiness very accurately is still a major challenge facing many banks. In this paper, we examine a real bank credit data and conduct several machine learning algorithms on the data for comparative analysis and to choose which algorithms is the best fit for learning bank credit data. The algorithms gave over 80% of accuracy in prediction. Furthermore, the most important features that determine whether a customer will default or otherwise in paying his/her credit the next month are extracted from a total of 23 features. We then applied these most important features on some selected machine learning algorithms and compare their predictive accuracy with the other algorithms that used all the 23 features. The results show no significant difference, signifying that these features

can accurately determine the credit worthiness of the customers. Finally, we formulate a predictive model using the most important features to predict the credit worthiness of a given customer.

1-INTRODUCTION

The growing volumes, varieties and velocity of data due to the emergence of the Internet in particular and the cheaper data sharing and storage facilities coupled with the cheaper but more powerful computational tools have opened a new frontier in the field of data science. And thus, there is currently an active ongoing research within the fields of data mining (discovering patterns in data) and machine learning's (building analytical models using algorithms for machine to "learn" from data), both aim at using algorithms and concepts to extract knowledge and pattern from data.

2- LITERATURE REVIEW:

Past Research and Approaches

In earlier studies, many approaches have been proposed to bring solutions to detect fraud from supervised approaches, unsupervised approaches to hybrid ones; which makes it a must to learn the technologies associated in credit card frauds detection and to have a clear understanding of the types of credit card fraud. As time progressed fraud patterns evolved introducing new forms of fraud making it a keen area of interest for researchers. The remainder of this section describes single machine learning algorithms, machine learning models and

fraud detection systems that were used in fraud detection. The problems that came across the review have analyzed for the later use of implementing an efficient machine learning model.

Challenges in Fraud Detection

With the analysis of various detection models, past researchers have found many problems regarding fraud detection. And they have mentioned Lack of real life data as a huge issue. Real life data are lacking because of the data sensitivity and privacy issues. Papers and have studied Imbalance data or skewed distribution of data. The reason behind this is having quite a less amount of frauds when compared to non-frauds in the transaction datasets. Paper states that data mining techniques take time to execute when dealing with big data. Overlapping of data is another major drawback in preparation of credit card transaction data. According to paper and the issue occurs due to some scenarios when the legitimate transactions look exactly like fraudulent transactions. In another way, fraudulent transactions may appear as legitimate transactions. Also, they have come across the difficulty in dealing with categorical data. When considering the credit card transaction data, most of the features have categorical values. In this case, almost all the machine learning algorithms do not support the categorical values. they have mentioned choice of detection

algorithms and feature selection as a challenge in detecting frauds since most of the machine learning algorithms take much time for training purposes than predicting. Another key issue that affects financial fraud detection is the feature selection. It aims to filter out the attributes that most describes the aspects of fraud detection and its characters. In paper they have highlighted fraud detection cost and lack of adaptability as challenges in the fraud detection process. When considering a system, the cost of fraudulent behaviour and the prevention cost should

be taken into consideration. Lack of adaptability occurs when the algorithm is exposed to new types of fraud patterns and normal transactions. Effectiveness can change according to the problem definition and its specifications, so having a good understanding of the performance measure is necessary.

Summary of Algorithms used in Literature

There are different kinds of models implemented for credit card fraud detections. In those models, different algorithms have been used.

Adapting the fraud detection system to newly introduced frauds can be problematic whether to retrain the machine learning model due to drastic changes in the fraud patterns, also may be costly and risky. For instance, Tyler et al. extended a framework proposed in, implemented the model and the model was applied to a real-world transaction log. To address the classification problem Logistic Regression (LR) has been used. The instances of fraudulent transactions have been discretized into strategies by using Gaussian Mixture Models (GMMs). Here synthetic minority oversampling technique was used to address the class imbalance. To stand out the significance of estimates in economic value sensitivity analysis has been used. The results have proven that a practical method which uses minimal steps to retrain a model could function as same as a classifier that typically retrains every round. There is another model called Risk-Based Ensemble (RBE) that can handle the data consisting of issues and give outstanding results. For handling imbalanced data, a highly efficient bagging model has been used. To handle the implicit noise in the transaction dataset they have used Naive Bayes algorithm. Peter et al. evaluated several deep learning algorithms with respect to their efficacy. The four topologies are Recurrent Neural Networks (RNNs), Gated Recurrent Units (GRUs), Long

Short-term Memory (LSTMs), and Artificial Neural Networks (ANNs). In their project in addition to data cleaning and other data

preparation steps, they have overcome class imbalance and scalability problems by using under sampling. To discover which hyper-parameters had the highest influence on the performance of the model, the sensitivity analysis was carried out. They have discovered that the performance of the model was affected by the size of the network. They concluded that larger the network it showed better performance.

Insights from Ensemble and Deep Learning Methods

Credit card data have the issue of skewed distribution which is also known as the class imbalance. According to Andrea et al., their project addresses class imbalance including other issues such as concept drift and verification latency. They have also illustrated the most relevant performance matrix that can be used in credit card fraud detection. The achievement of the research also includes a formal model and a powerful learning strategy for addressing the 'verification latency' and an 'alert and feedback' mechanism. According to experiments they have declared the precision of the alerts as the most important measure.

Chee et al. used twelve standard models and hybrid methods which use AdaBoost and majority voting methods to achieve better accuracy rates in credit card fraud detection. They were evaluated using both benchmark and real-world data. A summary of the strengths and limitations of the methods were evaluated. The Matthews Correlation Coefficient metric (MCC) has been taken as the performance measure. To evaluate the robustness of the algorithms noise was added to the data. Also, they have proved that the majority voting method was not

affected by the added noise.

3-BASELINE AND TARGET SYSTEM

Baseline System over view

The application of machine learning and data approaches to study financial data are comprehensively described below, conducted research on using attributes of customers to assess credit risk by using a weighted-selected attribute bagging method. They benchmarked their result experimentally by using two credit databases and reported outstanding performance both in term of prediction accuracy and stability as compared with another state of the art methods.

Boundaries of Baseline System

A data mining approach is also proposed by Moro to predict the success or otherwise of a Portuguese retail bank in telemarketing. They applied various data mining models on the bank telemarketing data and reported that the neural network data mining method was the best for analyzing the data. The role of machine learning techniques in business data mining is outlined. Their work described the strengths and weaknesses of various machine learning techniques within the context of business data mining approach.

Target System Features

The dataset used in this paper is taken from the UCI machine learning data repository submitted by I-Cheng Yeh. The attributes of the dataset are described in the repository. The "default payment next month" coded as 'no' or 'yes' in the dataset is treated as the response variable represented as y in this paper. We conducted data exploratory techniques on the dataset in order to understand the nature of the dataset.

Dataset used and Problem Framing

The exploratory analysis revealed that there seems to be some relationship between Age of the customers, their bank balance and their ability to pay their credit in the following month. Banks customer between the

ages of 20 and 60 years with small limited bank balance are seen to be the highest defaulters in paying their bank loans.

4-SYSTEM MODULES:

Admin module

Admin can login to the Application and add to the credit holder Information and his limit to use. And admin can monitor the suspects and he can block/activate users.

User module

User can register to the application and he can perform the activities of credit card usage like buy products, paymoney

Data Collection module

The dataset collected for predicting loan default customers is predicted into Training set and testing set. Generally 80:20 ratios are applied to split the training set and testing set.

Data Classification and Prediction module

The data model which was created using Decision tree is applied on the training set and based on the test result accuracy, Test set prediction is done. Following are the attributes

5-DATA PREPROCESSING AND CLEANING

Data Collection and Splitting

The data which was collected might contain missing values that may lead to inconsistency. To gain better results data need to be preprocessed so as to improve the efficiency of the algorithm. The outliers have to be removed and also variable conversion need to be done. In order to overcoming these issues we use map function

Preprocessing Techniques

Based on the correlation among attributes it was observed more likely to pay back their loans. The attributes that are individual and significant can

include Property area, education, loan amount, and lastly credit History, which is since by intuition it is considered as important. The correlation among attributes can be identified using corplot and box plot in Python platform.

Model using Classification Algorithms for predicting the loan defaulter's and non defaulter's problem LSTM algorithm is used. It is effective because it provides better results in classification problem. It is extremely intuitive, easy to implement and provide interpretable predictions. It produces out of bag estimated error which was proven to be unbiased in many tests. It is relatively easy to tune with. It gives highest accuracy result for the problem

We noticed that 299 cases in the test set are predicted as "Y", which is more than 81%, whereas in the training set only about 69% had this status.

Pandas:

Pandas is an open-source Python Library providing high-performance data manipulation and analysis tool using its powerful data structures. The name Pandas is derived from the word Panel Data – an Econometrics from Multidimensional data. In 2008, developer Wes McKinney started developing pandas when in need of high performance, flexible tool for analysis of data. Prior to Pandas, Python was majorly used for data munging and preparation. It had very little contribution towards data analysis. Pandas solved this problem. Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data — load, prepare, manipulate, model, and analyze. Python with Pandas is used in a wide range of fields including academic and commercial domains including finance, economics, Statistics, analytics, etc. Standard Python distribution doesn't come bundled with Pandas module. A lightweight alternative is to install NumPy using popular Python

6-MACHINE LEARNING ALGORITHMS

Machine Learning Process

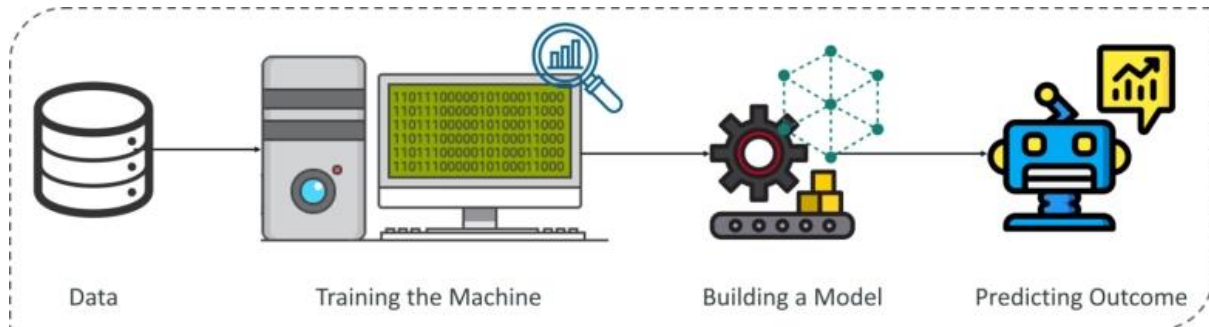


Figure: 1 : Machine Learning Workflow

How does Machine Learning Work?

Machine Learning algorithm is trained using a training data set to create a model. When new input data is introduced to the ML algorithm, it makes a prediction on the basis of the model.

The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine Learning algorithm is deployed. If the accuracy is not acceptable, the Machine Learning algorithm is

trained again and again with an augmented training data set.

The Machine Learning process involves building a Predictive model that can be used to find a solution for a Problem Statement. To understand the Machine Learning process let's assume that you have been given a problem that needs to be solved by using Machine Learning.

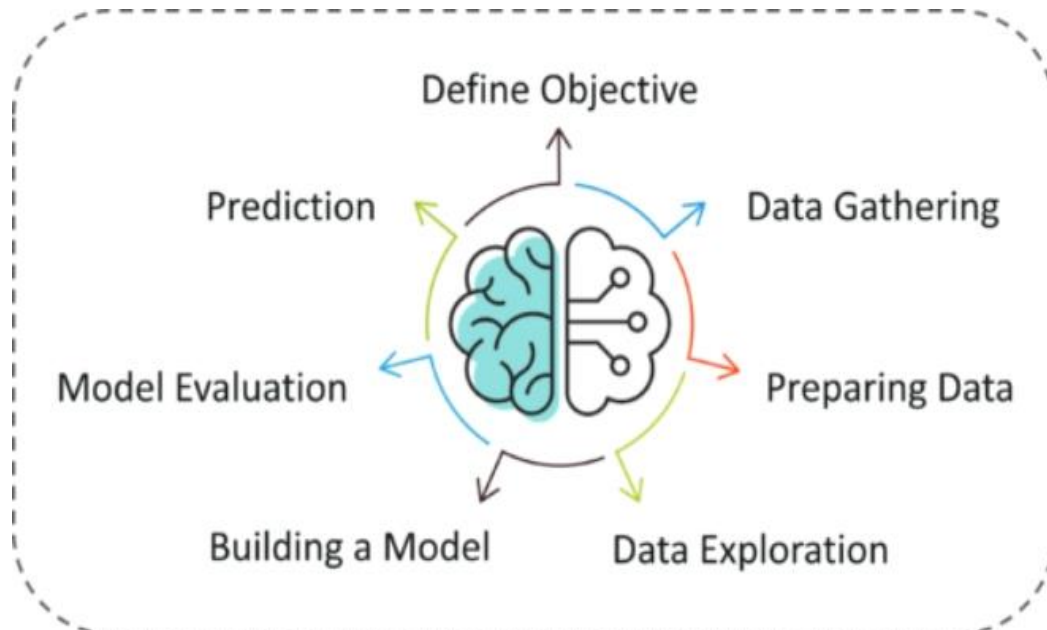


Figure : 2 : Phases of the Machine Learning Process

The below steps are followed in a Machine Learning process:

Step 1: Define the Objective of the Problem Statement

At this step, we must understand what

exactly needs to be predicted. In our case, the objective is to predict the possibility of rain by studying weather conditions. At this stage, it is also essential to take mental notes on what kind of data can be used to solve this problem or the type of

approach you must follow to get to the solution.

Step 2: Data Gathering

Step 3: Data Preparation

The data you collected is almost never in the right format. You will encounter a lot of inconsistencies in the data set such as missing values, redundant variables, duplicate values, etc. Removing such inconsistencies is very essential because they might lead to wrongful computations and predictions. Therefore, at this stage, you scan the data set for any inconsistencies and you fix them then and there.

Step 4: Exploratory Data Analysis

Grab your detective glasses because this stage is all about diving deep into data and finding all the hidden data mysteries. EDA or Exploratory Data Analysis is the brainstorming stage of Machine Learning. Data Exploration involves understanding the patterns and trends in the data. At this stage, all the useful insights are drawn and correlations between the variables are understood.

For example, in the case of predicting rainfall, we know that there is a strong possibility of rain if the temperature has fallen low. Such correlations must be understood and mapped at this stage.

Step 5: Building a Machine Learning Model

All the insights and patterns derived during Data Exploration are used to build the Machine Learning Model. This stage always begins by splitting the data set into two parts, training data, and testing data. The training data will be used to build and analyze the model. The logic of the model is based on the Machine Learning Algorithm that is being implemented.

Choosing the right algorithm depends on the type of problem you're trying to solve, the data set and the level of complexity of the problem. In the upcoming sections, we will discuss the different types of problems that can be solved by using Machine Learning.

Step 6: Model Evaluation & Optimization

After building a model by using the training data set, it is finally time to put the model to a test. The testing data set is used to check the efficiency of the model and how accurately it can predict the outcome. Once the accuracy is calculated, any further improvements in the model can be implemented at this stage. Methods like parameter tuning and cross-validation can be used to improve the performance of the model.

Step 7: Predictions

Once the model is evaluated and improved, it is finally used to make predictions. The final output can be a Categorical variable (eg. True or False) or it can be a Continuous Quantity (eg. the predicted value of a stock).

In our case, for predicting the occurrence of rainfall, the output will be a categorical variable.

7-SYSTEM DESIGN

The input design is the link between the information system and the user. It comprises the developing specification and procedures for data preparation and those steps are necessary to put transaction data in to a usable form for processing can be achieved by inspecting the computer to read data from a written or printed document or it can occur by having people keying the data directly into the system. The design of input focuses on controlling the amount of input required, controlling the errors, avoiding delay, avoiding extra steps and keeping the process simple. The input is designed in such a way so that it provides security and ease of use with retaining the privacy. Input Design considered the following things:

- What data should be given as input?
- How the data should be arranged or coded?
- The dialog to guide the operating personnel in providing input.

➤ Methods for preparing input validations and steps to follow when error occur.

The data entry screen is designed in such a way that all the data manipulates can be performed. It also provides record viewing facilities.

Output Design

A quality output is one, which meets the requirements of the end user and presents the information clearly. In any system results of processing are communicated to the users and to other system through outputs. In output design it is determined how the information is to be displaced for immediate need and also the hard copy output. It is the most important and direct source information to the user. Efficient and intelligent output design improves the system's relationship to help user decision-making.

1. Designing computer output should proceed in an organized, well thought out manner; the right output

must be developed while ensuring that each output element is designed so that people will find the system can use easily and effectively. When analysis design computer output, they should Identify the specific output that is needed to meet the requirements.

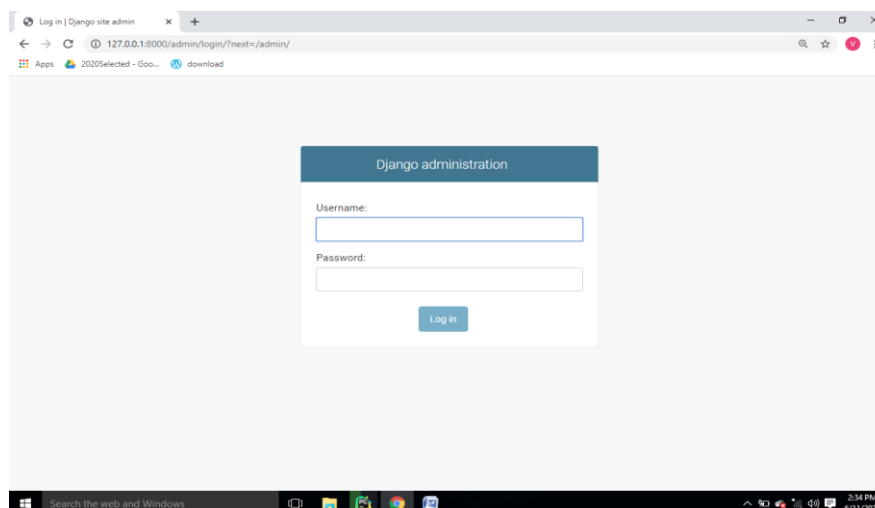
2. Select methods for presenting information.

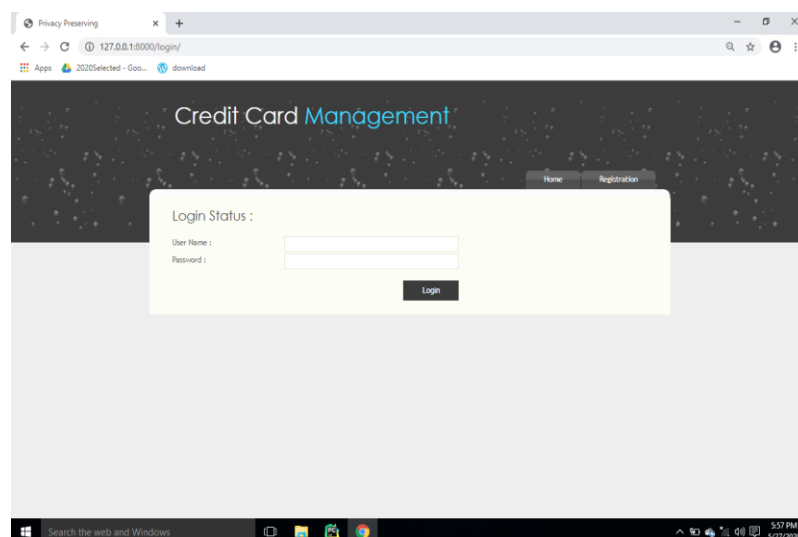
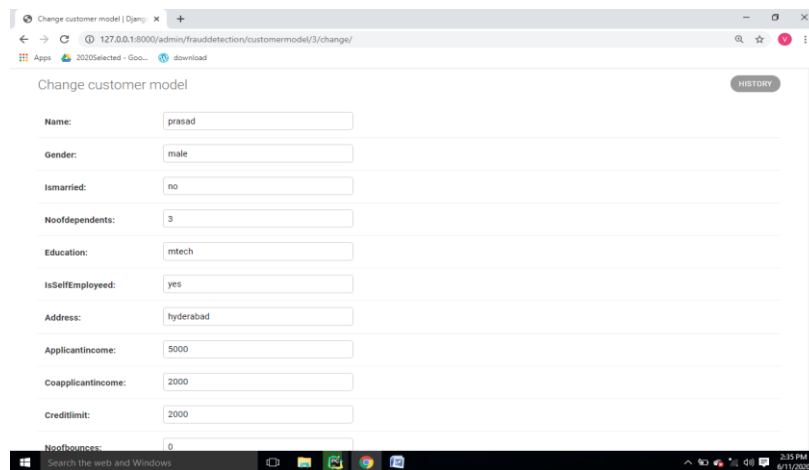
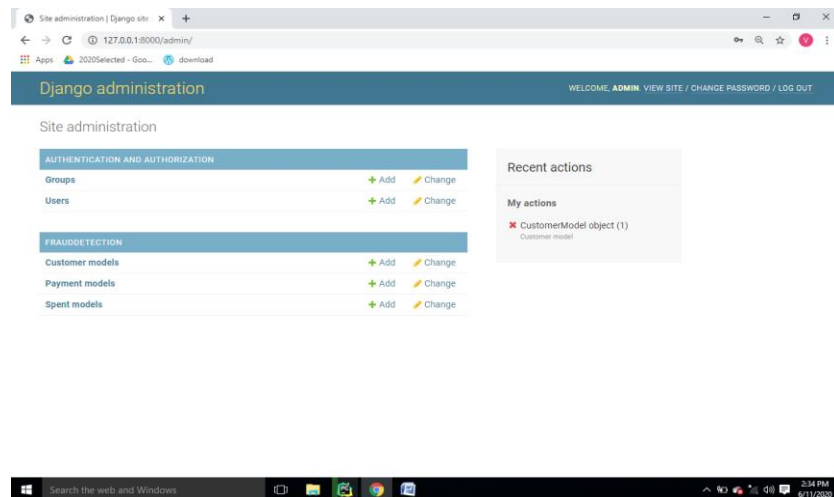
3. Create document, report, or other formats that contain information produced by the system.

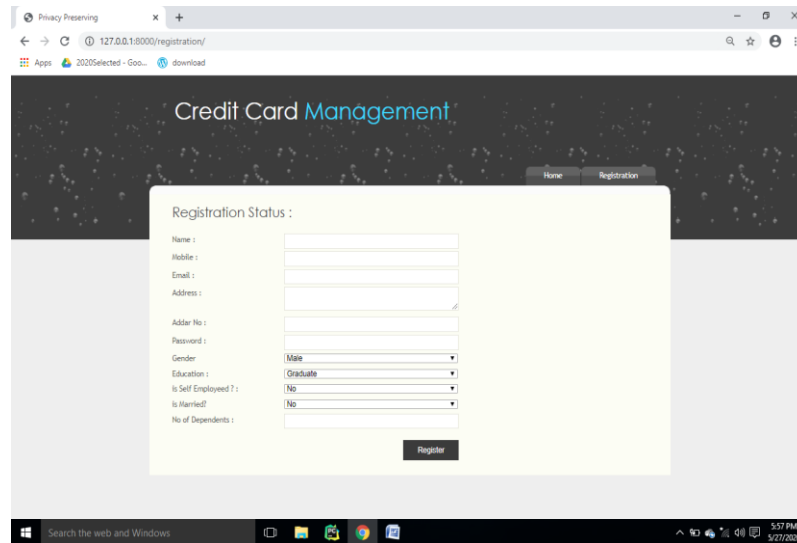
The output form of an information system should accomplish one or more of the following objectives.

- ❖ Convey information about past activities, current status or projections of the
- ❖ Future.
- ❖ Signal important events, opportunities, problems, or warnings.
- ❖ Trigger an action.
- ❖ Confirm an action.

8-SNAPSHOTS







Credit Card Management

Registration Status :

Name :

Mobile :

Email :

Address :

Addr No :

Password :

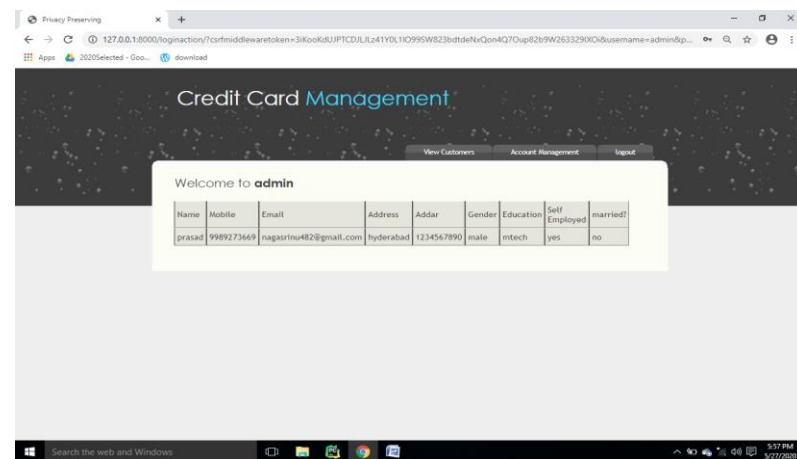
Gender :

Education :

Is Self Employed ? :

Is Married? :

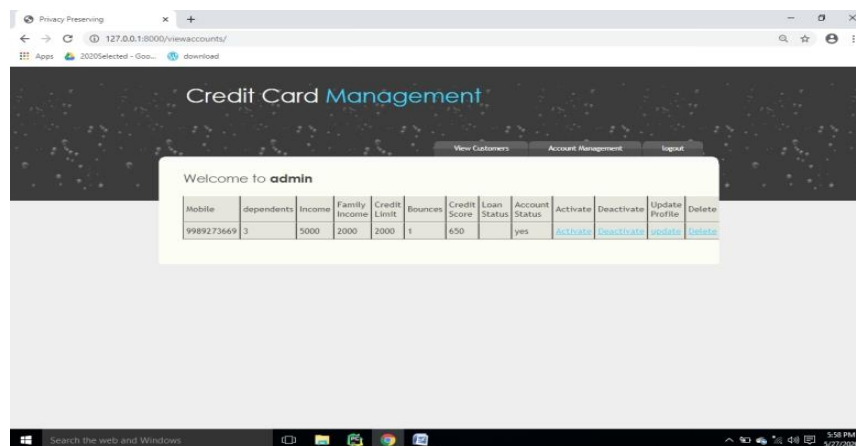
No of Dependents :



Credit Card Management

Welcome to **admin**

Name	Mobile	Email	Address	Addr	Gender	Education	Self Employed	married?
prasad	9989273669	nagasrinu482@gmail.com	hyderabad	1234567890	male	intech	yes	no



Credit Card Management

Welcome to **admin**

Mobile	dependents	Income	Family Income	Credit Limit	Bounces	Credit Score	Loan Status	Account Status	Activate	Deactivate	Update Profile	Delete
9989273669	3	5000	2000	2000	1	650		yes	Activate	Deactivate	Update	Delete

CONCLUSION AND FUTURE SCOPE:

In this paper, we applied machine learning approach to study bank credit dataset in order to predict customers' credit worthiness (their ability to pay

their loan in the next month). We employed 15 different machine learning algorithms on the dataset in order to determine which algorithms is the best fit for studying bank credit dataset. The experiment

revealed that, apart from the Nearest Centroid and Gaussian Naive Bayes, the rest of the algorithms perform credibly well in term of their accuracy and other performance evaluation metrics. Each of these algorithms achieved an accuracy rate between 76% to over 80%. We also determined the most important features that influence the credit worthiness of customers. These most important features are then used on some selected algorithms and their performance accuracy compared with the instance of using all the 23 features. The experimental results showed no significance difference in their predictive accuracy and other metrics. We further formulated a predictive model using linear regression that composed of the 3 most important features, for predicting customer's credit worthiness. These findings have a lot of implications. The model can be used as a tool to advice banks as which factors are important in determining the credit worthiness of customers. Furthermore, the result showed which machine learning algorithms are not suitable for studying bank credit dataset. We intend to develop a hybrid machine learning system that will incorporate the most important features that determine credit worthiness of customers in order to formulate banks' risk automated system.

REFERENCES

- [1] G. McLachlan, K.-A. Do, and C. Ambrose, Analyzing microarray gene expression data, vol. 422. John Wiley & Sons, 2005.
- [2] E. Ngai, Y. Hu, Y. Wong, Y. Chen, and X. Sun, "The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature," Decision Support Systems, vol. 50, no. 3, pp. 559–569, 2011.
- [3] R. O. Duda, P. E. Hart, and D. G. Stork, Pattern classification. John Wiley & Sons, 2012.
- [4] I. H. Witten and E. Frank, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2005.
- [5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.
- [6] J. Li, H. Wei, and W. Hao, "Weight-selected attribute bagging for credit scoring," Mathematical Problems in Engineering, vol. 2013, 2013.
- [7] S. Moro, P. Cortez, and P. Rita, "A data-driven approach to predict the success of bank telemarketing," Decision Support Systems, vol. 62, pp. 22–31, 2014.
- [8] I. Bose and R. K. Mahapatra, "Business data mininga machine learning perspective," Information & management, vol. 39, no. 3, pp. 211–225, 2001.
- [9] C.-F. Tsai and M.-L. Chen, "Credit rating by hybrid machine learning techniques," Applied soft computing, vol. 10, no. 2, pp. 374–380, 2010.
- [10] K. Y. Tam and M. Y. Kiang, "Managerial applications of neural networks: the case of bank failure predictions," Management science, vol. 38, no. 7, pp. 926–947, 1992.
- [11] M. Ghazanfar and A. Prugel-Bennett, "Building switching hybrid recommender system using machine learning classifiers and collaborative filtering," IAENG International Journal of Computer Science, vol. 37, no. 3, 2010.
- [12] M. Er, L. Zhai, X. Li, and L. San, "A hybrid online sequential extreme learning machine with simplified hidden network," IAENG International Journal of Computer Science, vol. 39, no. 1, pp. 1–9, 2012.
- [13] M. Lichman, "UCI machine learning repository," 2013.