CMR UNIVERSITY

Private University Established in Karnataka State by Act No. 45 of 2013

SCHOOL OF ENGINEERING AND TECHNOLOGY

A Machine Learning Project Report

On

"CREDIT CARD FRAUD DETECTION"

Submitted in partial fulfillment of the requirements for the award of degree in

Bachelor of Technology

in

Computer Science and Engineering

Of CMR University, Bangalore

Submitted by:

SARANYA T
16UG08047

Under the Guidance of:

Mr. Naveen Ghorpade
Assistant Professor
Dept. of CSE, SOET

Department of Computer Science and Engineering

Off Hennur - Bagalur Main Road,

Near Kempegowda International Airport, Chagalahatti,Bangalore, Karnataka-562149

2019-2020

CMR UNIVERSITY
Private University Established in Karnataka State by Act No. 45 of 2013

SCHOOL OF ENGINEERING AND TECHNOLOGY

Department of Computer Science and Engineering

CERTIFICATE

This is to Certify that the project work, entitled “Human Activity recognition”, submitted to the CMR University, Bangalore, in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a record of work done by Ms.Saranya T bearing university register number 16UG08047 during 2019-20 at School of Engineering and Technology, CMR University, Bangalore under my supervision and guidance. The Contents of this project work, in full or in parts, have not been submitted to any other Institute or University for the award of a degree.

Signature of the Guide
Mr. Naveen Ghorpade
Assistant Professor
Dept. of CSE, SOET, CMRU

Signature of the HOD
Dr. Arun Biradar
Professor and Head
Dept. of CSE, SOET, CMRU

Signature of the Dean
Dr. Jayaprasad M
Professor and Dean
SOET, CMRU

Examiners Signature with date

DECLARATION

I, Saranya. T (16UG08047) student of 7th semester B.Tech. Computer Science and Engineering, School of Engineering and Technology, Bangalore, hereby declare that the project work entitle" Credit card fraud detection" has been carried out by me under the guidance of Mr. Naveen Ghorpade, assistant professor, Department of Computer Science and Engineering, School of Engineering and Technology. This report is submitted in partial fulfillment of the requirement for award of Bachelor of Technology in Computer Science and Engineering by CMR University, Bangalore during the academic year 2019-2020. The matter embodied in the dissertation has not been submitted previously by anybody for the award of any degree or diploma to any other university.

Place: Bangalore
Date:

SARANYA T
16UG08047

ACKNOWLEDGEMENT

The satisfaction that accompanies the successful completion of this project would be incomplete without the mention of the people who made it possible, without whose constant guidance and encouragement would have made efforts go in vain. I consider myself privileged to express gratitude and respect towards all those who guided me through the completion of the project. I express my heartfelt sincere gratitude to Dr. Jayaprasad M, Dean, School of Engineering and Technology, CMR University for his support. I would like to express my thanks to Dr. Arun Biradar, Professor and Head, Department of Computer Science and Engineering, School of Engineering and Technology, CMR University, Bangalore, for his encouragement that motivated me for the successful completion of Project work. I express my thanks to my Project Guide Mr. Naveen Ghorpade, assistant professor, Department of Computer Science and Engineering, School of Engineering and Technology, CMR University for his constant support.

I would like to thank all the professors and staff of Computer Science and Engineering Department for their co-operation and timely guidance.

SARANYA.T
16UG08028

ABSTRACT

Activity recognition systems are a large field of research and development, currently with a focus on advanced machine learning algorithms, innovations in the field of hardware architecture, and on decreasing the costs of monitoring while increasing safety. This project aims at identifying the different activities like standing, sitting, walking etc. and find out the percentage values of each activity performed by each human being and is being represented using a visual representation.

This project uses four different algorithms for activity recognition and finds out which algorithm is best suited for the activity recognition. The average values for each algorithm is found and the algorithm with the highest average value is considered as the best algorithm.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

Fraud is one of the major ethical issues in the credit card industry. The main aims are, firstly, to identify the different types of credit card fraud, and, secondly, to review alternative techniques that have been used in fraud detection. The sub-aim is to present, compare and analyze recently published findings in credit card fraud detection. This article defines common terms in credit card fraud and highlights key statistics and figures in this field. Depending on the type of fraud faced by banks or credit card companies, various measures can be adopted and implemented. The proposals made in this paper are likely to have beneficial attributes in terms of cost savings and time efficiency. The significance of the application of the techniques reviewed here is in the minimization of credit card fraud. Yet there are still ethical issues when genuine credit card customers are misclassified as fraudulent.

1.2 Problem Statement

In today’s world data analysis plays a vital role where anything in the world is considered to be a data generator. Online payment does not require physical card. Anyone who knows the details of card can make fraud transaction. Currently card holder comes to know only after the fraud transaction carried out. No mechanism track the fraud transaction.

1.3 Objectives

Logistic Regression is a supervised classification method that returns the probability of binary dependent variable that is predicted from the independent variable of dataset that is logistic regression predict the probability of an outcome which has two values either zero or one, yes or no and false or true. Logistic regression has similarities to linear regression but as in linear regression a straight line is obtained, logistic regression shows a curve. The use of one or several predictors or independent variable is on what prediction is based, logistic regression produces logistic curves which plots the values between zero and one. Regression is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. There are many types of logistic regression model such as binary logistic model, multiple logistic model, binomial logistic models. Binary Logistic Regression model is used to estimate the probability of a binary response based on one or more predictors.

CHAPTER 2

LITERATURE REVIEW

This section reviews the research works carried out by different researchers that are related to the proposed work. The following research papers includes papers on languages used for web development and their pros and cons. This section also includes papers published by students on their college website development projects which will be briefed in the later part of this section. This section reviews the research works carried out by different researchers that are related to the proposed work. The data used for the website or processed by the website are stored in the data bases.

Ln[1]: Raj S.B.E., Portia A.A., Analysis on credit card fraud detection methods.

This paper represents a research about a case study involving credit card fraud detection, where data normalization is applied before Cluster Analysis and with results obtained from the use of Cluster Analysis and Artificial Neural Networks on fraud detection has shown that by clustering attributes neuronal inputs can be minimized. And promising results can be obtained by using normalized data and data should be MLP trained. This research was based on unsupervised learning. Significance of this paper was to find new methods for fraud detection and to increase the accuracy of results.

Ln[2] A.Shen etal (2007)

demonstrate the efficiency of classification models to credit card fraud detection problem and the authors proposed the three classification models ie., decision tree, neural network and logistic regression. Among the three models neural network and logistic regression outperforms than the decision tree.

Ln [3] M.J.Islam et al (2007)

Proposed the probability theory frame work for making decision under uncertainty. After reviewing Bayesian theory, naïve bayes classifier and k-nearest neighbor classifier is implemented and applied to the dataset for credit card system.

Ln[4] Y. Sahin and E. Duman(2011)

Has cited the research for credit card fraud detection and used seven classification methods took a major role .In this work they have included decision trees and SVMs to decrease the risk of the banks. They have suggested Artificial Neural networks and Logistic Regression classification models are more helpful to improve the performance in detecting the frauds.

Ln[5] Y. Sahin, E. Duman(2011)

has cited the research , used Artificial Neural Network and Logistic Regression Classification and explained ANN classifiers outperform LR classifiers in solving the problem under investigation. Here the training data sets distribution became more biased and the distribution of the training data sets became more biased and the efficiency of all models decreased in catching the fraudulent transactions.

Ln[6] Huang, S. (2013)..

Fraud Detection Model by Using Support Vector Machine Techniques developed two models based on logistic regression and SVM.

Ln[7] Ng, G., & Singh, H. (1997).

developed models based on an individual and combined machine learning techniques for handwritten digits' recognition. The results showed that the classification accuracy of the combined classifier model outperformed the individual classifier model.

Ln[8] Maes, S., Tuyls, K., Vanschoenwinkel, B., & Manderick, B. (2002).

the authors tried Bayesian Belief Networks (BBN) and Artificial Neural Networks (ANN) on a real dataset obtained from Euro pay International. Their experiment showed that the Bayesian Belief networks out performs ANN in terms of classification accuracy and training time. It was found that ANN may need several hours for training while BNN takes only 20 minutes. However, the trained ANN was found to be faster in classifyi

Ln [9] M.J.haseem et al (2008)

Proposed the probability theory frame work for making decision under uncertainty. After reviewing Bayesian theory, naïve bayes classifier and k-nearest neighbor classifier is implemented and applied to the dataset for credit card system.

Ln[10] Mk, T., & Rao, H. (1998).

developed models based on an individual and combined machine learning techniques for handwritten digits' recognition. The results showed that the classification accuracy of the combined classifier model outperformed the individual classifier model. …). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc...) as well as executable documents which can be run to perform data analysis.

Ln[11] Manderick (1999).

The use of one or several predictors or independent variable is on what prediction is based, logistic regression produces logistic curves which plots the values between zero and one. Regression is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. There are many types of logistic regression model such as binary logistic model, multiple logistic model, binomial logistic models. Binary Logistic Regression model is used to estimate the probability of a binary response based on one or more predictors.

Ln[12] Bahnsen, A. Stojanovic, D. Aouada, B. Ottersten,

Depending on the type of fraud faced by banks or credit card companies, various measures can be adopted and implemented. The proposals made in this paper are likely to have beneficial attributes in terms of cost savings and time efficiency. The significance of the application of the techniques reviewed here is in the minimization of credit card fraud. Yet there are still ethical issues when genuine credit card customers are misclassified as fraudulent. ANN in terms of classification accuracy and training time. It was found that ANN may need several hours for training while BNN takes only 20 minutes. However, the trained ANN was found to be faster in classification.

Ln[13] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland

The use of one or several predictors or independent variable is on what prediction is based, logistic regression produces logistic curves which plots the values between zero and one. Regression is a regression model where the dependent variable is categorical and analyzes the relationship between multiple independent variables. There are many types of logistic regression model such as binary logistic model, multiple logistic model, binomial logistic models. Binary Logistic Regression model is used to estimate the probability of a binary response based on one or more predictors.

Ln[14] Selvani Deepthi Kavila,LAKSHMI S.V.S.S.,RAJESH B

The sub-aim is to present, compare and analyze recently published findings in credit card fraud detection. This article defines common terms in credit card fraud and highlights key statistics and figures in this field. Depending on the type of fraud faced by banks or credit card companies, various measures can be adopted and implemented. The proposals made in this paper are likely to have beneficial attributes in terms of cost savings and time efficiency. The significance of the application of the techniques reviewed here is in the minimization of credit card fraud. Yet there are still ethical issues when genuine credit card customers are misclassified as fraudulent.

Ln[15] K. Chaudhary, B. Mallick, "*Credit Card Fraud*

machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n- dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. Svm classifier treated as one of the dominant classification algorithms.

CHAPTER 3

SOFTWARE AND HARDWARE REQUIREMENTS

SOFTWARE REQUIREMENTS:

Operating system:

- Windows
- Linux- Debian, Fedora, Ubuntu etc

Software tools:

- Anaconda navigator
- Jupiter notebook

HARDWARE REQUIREMENTS:

- Processor (CPU) with 2 gigahertz (GHz) frequency or above
- A minimum of 2 GB of RAM
- Monitor Resolution 1024 X 768 or higher

4.5 SOFTWARE DESCRIPTION:

1. **Anaconda Navigator:** Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows users to launch applications and manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository, install them in an environment, run the packages and update them. It is available for Windows, macOS and Linux.

   The following applications are available by default in Navigator:

- JupyterLab
- Jupyter Notebook
- QtConsole
- Spyder
- Glueviz
- Orange
- Rstudio
- Visual Studio Code

2. **Jupyter Notebook:** The Jupyter Notebook is an open source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning and much more. Notebook documents: It is the document produced by the Jupyter Notebook App, which contain both computer code (e.g. python) and rich text elements (paragraph, equations, figures, links, etc…). Notebook documents are both human-readable documents containing the analysis description and the results (figures, tables, etc...) as well as executable documents which can be run to perform data analysis.

Notebook Dashboard: *It* is the component which is shown first when you launch Jupyter Notebook App. The Notebook Dashboard is mainly used to open notebook documents, and to manage the running kernels (visualize and shutdown). The Notebook Dashboard has other features similar to a file manager, namely navigating folders and renaming/deleting files. Notebook *kernel*: It is a “computational engine” that executes the code contained in a Notebook document. The ipython *kernel*, referenced in this guide, executes python code. Kernels for many other languages exist (official kernels).

When you open a Notebook document, the associated kernel is automatically launched. When the notebook is executed (either cell-by-cell or with menu Cell -> Run All), the kernel performs the computation and produces the results. Depending on the type of computations, the kernel may consume significant CPU and RAM. Note that the RAM is not released until the kernel is shut-down.

CHAPTER 4

SYSTEM DEVELOPMENT PROCESSS

4.1MODEL USED

System design is the phase that bridges the gap between problem domain and the existing system in a manageable way. This phase focuses on the solution domain, i.e. “how to implement?”

Specification talks about the hardware and software requirements needed to fulfill or run the project. Basically, the minimum requirements on which the program or app will run.

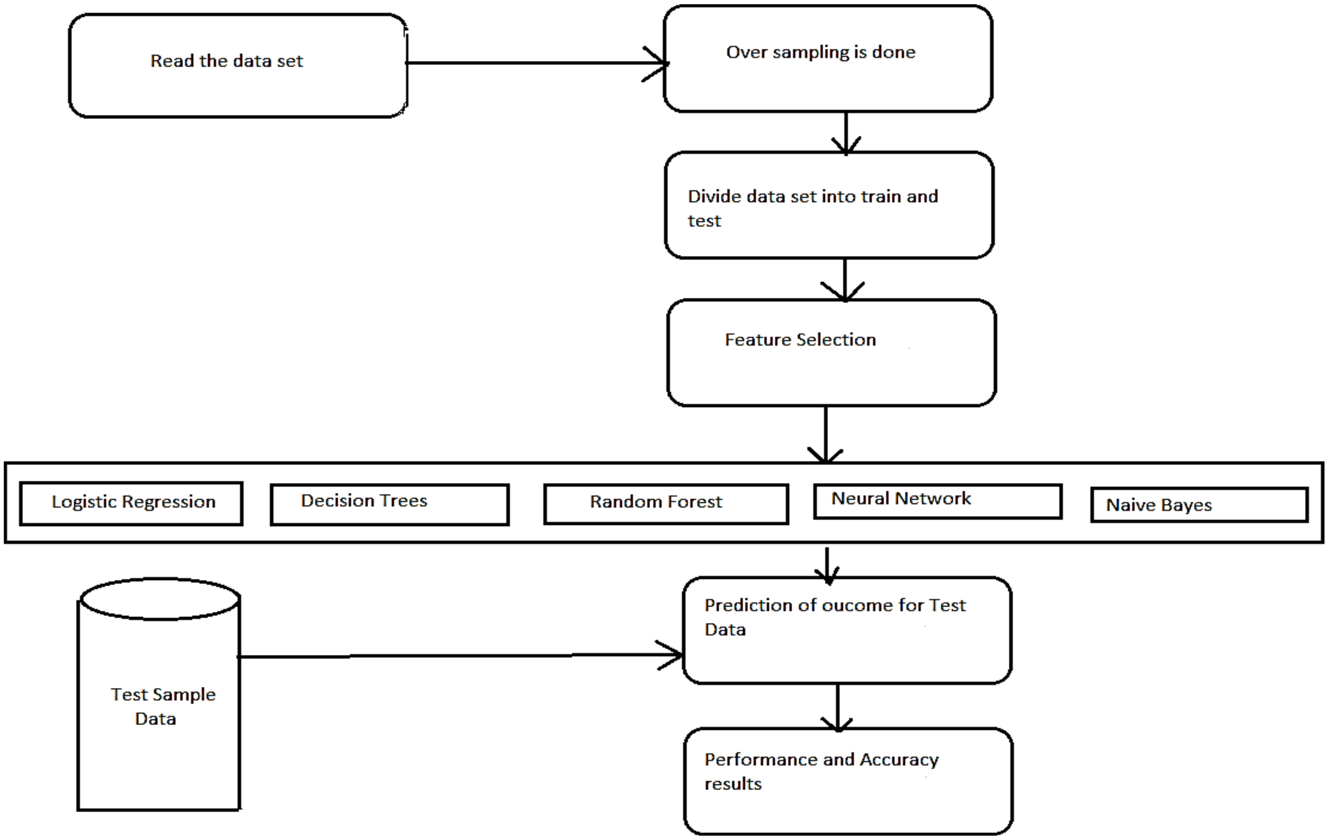


Fig 4.1 Design

- **Processing Steps:**

Step 1: Read the dataset.

Step 2: Random Sampling is done on the data set to make it balanced.

Step 3: Divide the dataset into two parts i.e., Train dataset and Test dataset.

Step 4: Feature selection are applied for the proposed models.

Step 5: Accuracy and performance metrics has been calculated to know the efficiency for different algorithms.

Step6: Then retrieve the best algorithm based on efficiency for the given dataset.

4.1.1 Requirements

The first phase involves understanding what needs to design and what is its function, purpose, etc. Here, the specifications of the input and output or the final product are studied and marked.

4.1.2 System Design

The requirement specifications from the first phase are studied in this phase and system design is prepared. System Design helps in specifying hardware and system requirements and also helps in defining overall system architecture. The software code to be written in the next stage is created now.

4.1.3 Implementation

With inputs from system design, the system is first developed in small programs called units, which are integrated into the next phase. Each unit is developed and tested for its functionality which is referred to as Unit Testing.

4.1.4 Integration and Testing

All the units developed in the implementation phase are integrated into a system after testing of each unit. The software designed, needs to go through constant software testing to find out if there are any flaw or errors. Testing is done so that the client does not face any problem during the installation of the software.

4.1.5 Deployment of System

Once the functional and non-functional testing is done, the product is deployed in the customer environment or released into the market.

4.1.6 Maintenance

This step occurs after installation, and involves making modifications to the system or an individual component to alter attributes or improve performance.

4.2 SYSTEM DESIGN

It is the phase where the SRS document is converted into a format that can be implemented and decides how the system will operate.

In this phase, the complex activity of system development is divided into several smaller sub activities, which coordinate with each other to achieve the main objective of system development.

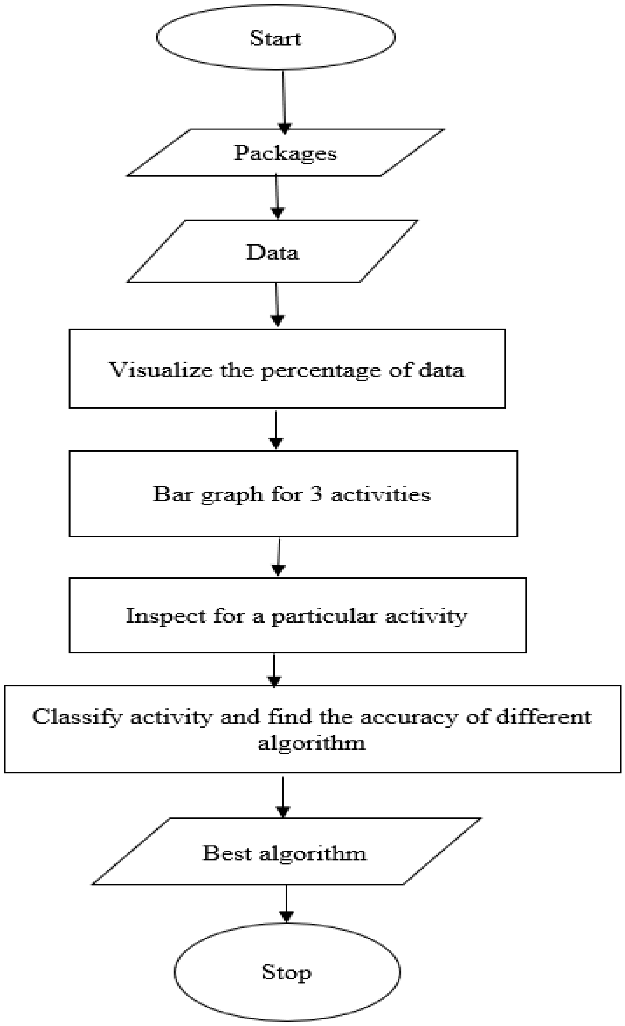

Start
Packages
Data
Visualize the percentage of data
Bar graph for 3 activities
Inspect for a particular activity
Classify activity and find the accuracy of different algorithm
Best algorithm
Stop

Fig 4.2 Flow chart of Credit card fraud detection

4.3 ALGORITHM EXPLAINATION

4.3.1 Support Vector Machine

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, we plot each data item as a point in n- dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiate the two classes very well (look at the below snapshot).

SVM or Support Vector Machine is a linear model for classification and regression problems. It can solve linear and non-linear problems and work well for many practical problems. The idea of SVM is simple: The algorithm creates a line or a hyperplane which separates the data into classes. Svm classifier treated as one of the dominant classification algorithms.

Advantages of SVM algorithm

- SVMs are effective when the number of features is quite large.
- It works effectively even if the number of features are greater than the number of samples.
- Non-Linear data can also be classified using customized hyper planes built by using kernel trick.
- It is a robust model to solve prediction problems since it maximizes margin.

Disadvantage of SVM algorithm

- The biggest limitation of Support Vector Machine is the choice of the kernel. The wrong choice of the kernel can lead to an increase in error percentage.
- With a greater number of samples, it starts giving poor performances.
- SVMs have good generalization performance but they can be extremely slow in the test phase.

- SVMs have high algorithmic complexity and extensive memory requirements due to the use of quadratic programming.

Application of SVM algorithm

SVMS are a by-product of Neural Network. They are widely applied to pattern classification and regression problems. Here are some of its applications:

- Facial expression classification: SVMs can be used to classify facial expressions. It uses statistical models of shape and SVMs.
- Speech recognition: SVMs are used to accept keywords and reject non-keywords them and build a model to recognize speech.
- Handwritten digit recognition: Support vector classifiers can be applied to the recognition of isolated handwritten digits optically scanned.
- Text Categorization: In information retrieval and then categorization of data using labels can be done by SVM.

4.3.2 Logistic regression

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labeled "0” and "1". In the logistic model, the log- odds (the logarithm of the odds) for the value labeled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value). The corresponding probability of the value labeled "1" can vary between 0 (certainly the value "0") and 1 (certainly the value "1"), hence the labeling; the function that converts log-odds to probability is the logistic function, hence the name. The unit of measurement for the log-odds scale is called a logit.

Analogous models with a different sigmoid function instead of the logistic function can also be used, such as the probity model; the defining characteristic of the logistic model is that increasing one of the independent variables multiplicatively scales the odds of the given outcome at a constant rate, with each independent variable having its own parameter; for a binary dependent variable this generalizes the odds ratio.

Advantages of logistic regression

- Logistic Regression performs well when the dataset is linearly separable.
- Logistic regression is less prone to over-fitting but it can over fit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over- fitting in these scenarios.
- Logistic Regression not only gives a measure of how relevant a predictor (coefficient size) is, but also its direction of association (positive or negative).
- Logistic regression is easier to implement, interpret and very efficient to train.

Disadvantage of logistic regression

- Main limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables. In the real world, the data is rarely linearly separable. Most of the time data would be a jumbled mess.
- If the number of observations are lesser than the number of features, Logistic Regression should not be used, otherwise it may lead to over fit.
- Logistic Regression can only be used to predict discrete functions. Therefore, the dependent variable of Logistic Regression is restricted to the discrete number set. This restriction itself is problematic, as it is prohibitive to the prediction of continuous data.

Application of logistic regression

- Image Segmentation and Categorization
- Geographic Image Processing
- Handwriting recognition
- Healthcare: Analysing a group ofover million people for myocardial infarction within a period of 10 years is an application area of logistic regression.
- Prediction whether a person is depressed or not based on bag of words from the corpus seems to be conveniently solvable using logistic regression and SVM.

4.3.3 K nearest neighbor

K-Nearest Neighbors is one of the most basic yet essential classification algorithms in Machine Learning. It belongs to the supervised learning domain and finds intense application in pattern recognition, data mining and intrusion detection.

It is widely disposable in real-life scenarios since it is non-parametric, meaning, it does not make any underlying assumptions about the distribution of data (as opposed to otheralgorithms such as GMM, which assume a Gaussian distribution of the given data).

We are given some prior data (also called training data), which classifies coordinates into groups identified by an attribute.

Advantage of K nearest neighbor

- K-NN is pretty intuitive and simple: K-NN algorithm is very simple to understand and equally easy to implement. To classify the new data point K-NN algorithm reads through whole dataset to find out K nearest neighbours.
- K-NN has no assumptions: K-NN is a non-parametric algorithm which means there are assumptions to be met to implement K-NN. Parametric models like linear regression has lots of assumptions to be met by data before it can be implemented which is not the case with K-NN.

- No Training Step: K-NN does not explicitly build any model, it simply tags the new data entry-based learning from historical data. New data entry would be tagged with majority class in the nearest neighbour.
- It constantly evolves: Given it's an instance-based learning; k-NN is a memory-based approach. The classifier immediately adapts as we collect new training data. It allows the algorithm to respond quickly to changes in the input during real-time use.
- Very easy to implement for multi-class problem: Most of the classifier algorithms are easy to implement for binary problems and needs effort to implement for multi class whereas K-NN adjust to multi class without any extra efforts.
- Can be used both for Classification and Regression: One of the biggest advantages of K-NN is that K-NN can be used both for classification and regression problems.
- One Hyper Parameter: K-NN might take some time while selecting the first hyper parameter but after that rest of the parameters are aligned to it.

Disadvantage of K nearest neighbor

- K-NN slow algorithm: K-NN might be very easy to implement but as dataset grows efficiency or speed of algorithm declines very fast.
- Curse of Dimensionality: KNN works well with small number of input variables but as the numbers of variables grow K-NN algorithm struggles to predict the output of new data point.
- K-NN needs homogeneous features: If you decide to build k-NN using a common distance, like Euclidean or Manhattan distances, it is completely necessary that features have the same scale, since absolute differences in features weight the same, i.e., a given distance in feature 1 must means the same for feature 2.
- Optimal number of neighbours: One of the biggest issues with K-NN is to choose the optimal number of neighbours to be consider while classifying the new data entry.
- Imbalanced data causes problems: k-NN doesn't perform well on imbalanced data. If we consider two classes, A and B, and the majority of the training data is labelled as A, then

the model will ultimately give a lot of preference to A. This might result in getting the less common class B wrongly classified.

- Outlier sensitivity: K-NN algorithm is very sensitive to outliers as it simply chose the neighbours based on distance criteria.
- Missing Value treatment: K-NN inherently has no capability of dealing with missing value problem.

4.3.4 Random forest classification

Random forest algorithm is a supervised classification algorithm. As the name suggest, this algorithm creates the forest with a number of trees. In general, the more trees in the forest the more robust the forest looks like. In the same way in the random forest classifier, the higher the number of trees in the forest gives the high accuracy results.

The same random forest algorithm or the random forest classifier can use for both classification and the regression task. Random forest classifier will handle the missing values. When we have more trees in the forest, random forest classifier won’t over fit the model. Can model the random forest classifier for categorical values also.

Advantage of K Random forest classification

- Random forest can solve both type of problems that is classification and regression and does a decent estimation at both fronts.
- One of benefits of Random Forest which exists me most is, the power of handle large data sets with higher dimensionality. It can handle thousands of input variables and identity most significant variables so it is considered as one of the dimensionality reduction method. Further, the model outputs importance of variable, which can be a very handy feature.
- It has an effective method for estimating missing data and maintains accuracy when large proportion of the data are missing.
- It has methods for balancing errors in data sets where classes are imbalanced.

- The capability of the above can be extended to unlabelled data, leading to unsupervised clustering, data views and outlier detection.
- Random forest involves sampling of the input data with replacement called as bootstrap sampling. Here one third of data is not used for training and can be used to testing. These are called the OUT OF BAG samples. Error estimated on these output bag samples is known as out of bag error. Study of error estimates by out of bag, gives evidence to show that the out of bag estimate is as accurate as using a test set of the same size as the training set. Therefore, using the out of bag error estimate removes the need for a set aside test set.

Disadvantage of Random forest classification

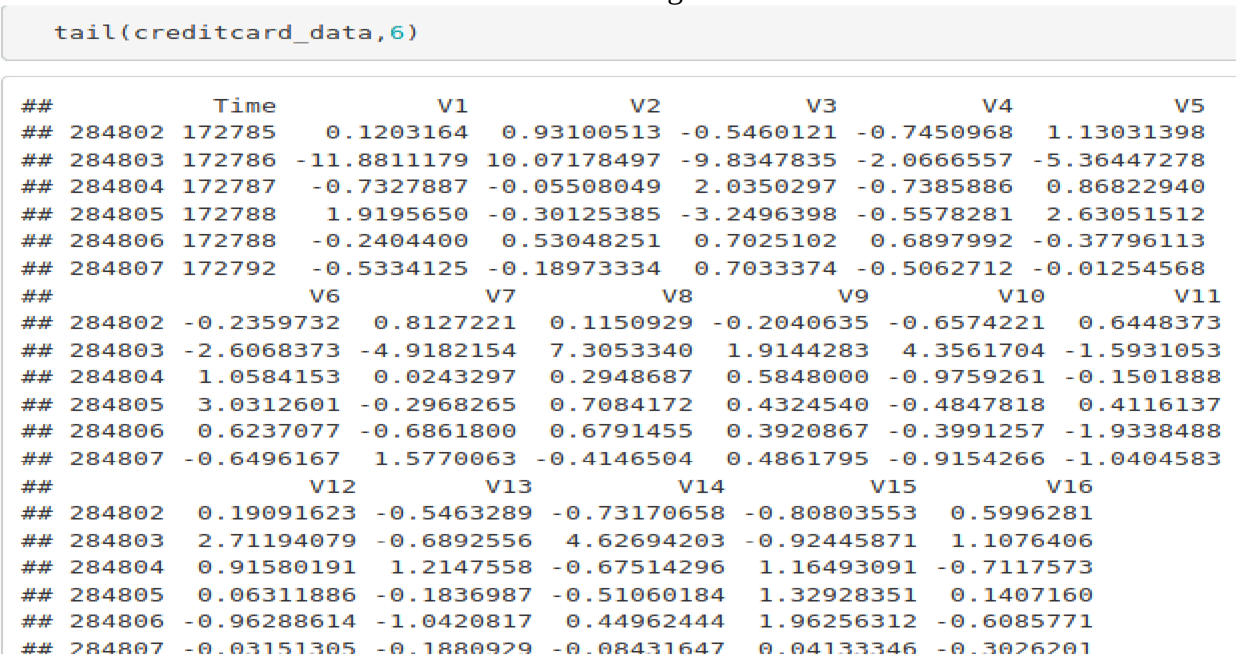
- It surely does a good job at classification but not as for regression problem as it does not give precise continuous nature prediction. In case of regression, it doesn't predict beyond the range in the training data, and that they may over fit data sets that are particularly noisy.
- Random forest can feel like a black box approach for a statistical modeller we have very little control on what the model does. You can at best try different parameters and random seeds.

Application of Random forest classification

- Banking
- Medicine
- Stock Market
- E-commerce

4.4 DATASET

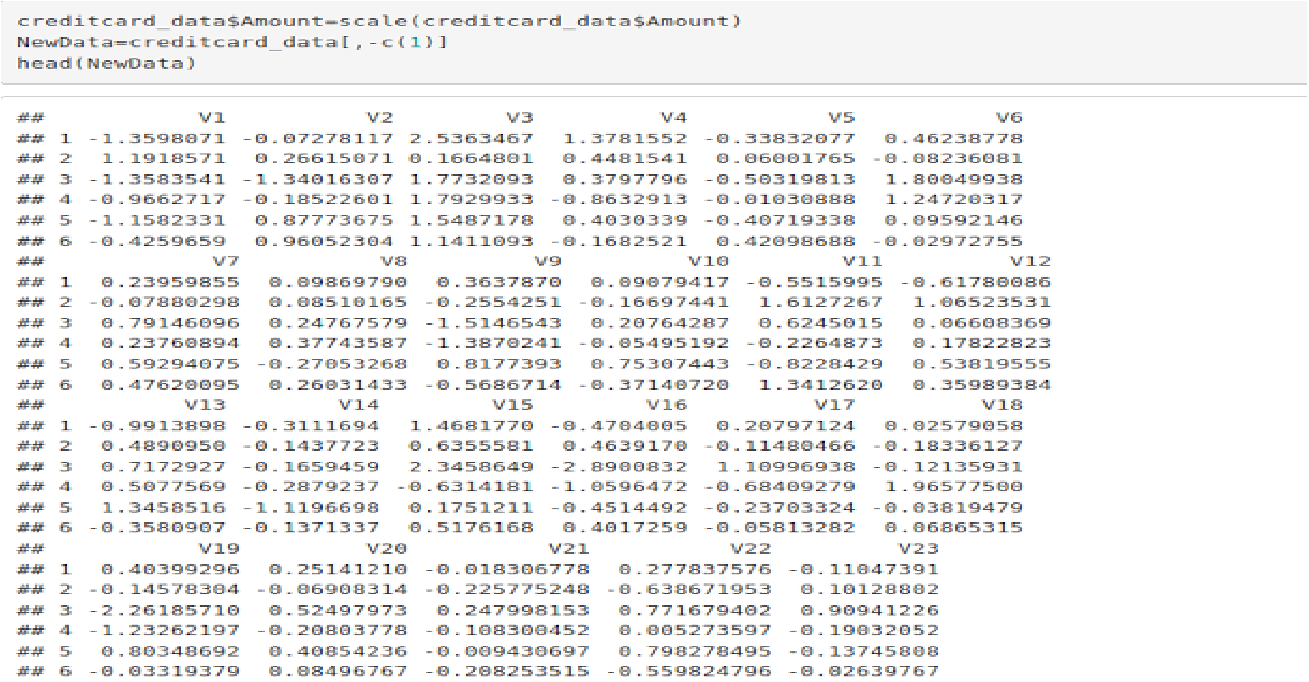
4.1 Testing dataset



```
tail(creditcard_data,6)
```

```
##           Time          V1          V2         V3         V4          V5
## 284802 172785   0.1203164  0.93100513 -0.5460121 -0.7450968  1.13031398
## 284803 172786 -11.8811179 10.07178497 -9.8347835 -2.0666557 -5.36447278
## 284804 172787  -0.7327887 -0.05508049  2.0350297 -0.7385886  0.86822940
## 284805 172788   1.9195650 -0.30125385 -3.2496398 -0.5578281  2.63051512
## 284806 172788  -0.2404400  0.53048251  0.7025102  0.6897992 -0.37796113
## 284807 172792  -0.5334125 -0.18973334  0.7033374 -0.5062712 -0.01254568
##                V6         V7         V8         V9        V10        V11
## 284802 -0.2359732  0.8127221  0.1150929 -0.2040635 -0.6574221  0.6448373
## 284803 -2.6068373 -4.9182154  7.3053340  1.9144283  4.3561704 -1.5931053
## 284804  1.0584153  0.0243297  0.2948687  0.5848000 -0.9759261 -0.1501888
## 284805  3.0312601 -0.2968265  0.7084172  0.4324540 -0.4847818  0.4116137
## 284806  0.6237077 -0.6861800  0.6791455  0.3920867 -0.3991257 -1.9338488
## 284807 -0.6496167  1.5770063 -0.4146504  0.4861795 -0.9154266 -1.0404583
##                V12        V13         V14         V15        V16
## 284802  0.19091623 -0.5463289 -0.73170658 -0.80803553  0.5996281
## 284803  2.71194079 -0.6892556  4.62694203 -0.92445871  1.1076406
## 284804  0.91580191  1.2147558 -0.67514296  1.16493091 -0.7117573
## 284805  0.06311886 -0.1836987 -0.51060184  1.32928351  0.1407160
## 284806 -0.96288614 -1.0420817  0.44962444  1.96256312 -0.6085771
## 284807 -0.03151305 -0.1880929 -0.08431647  0.04133346 -0.3026201
```

4.2 Training dataset



```
creditcard_data$Amount=scale(creditcard_data$Amount)
NewData=creditcard_data[,-c(1)]
head(NewData)
```

```
##           V1          V2        V3         V4          V5          V6
## 1 -1.3598071 -0.07278117 2.5363467  1.3781552 -0.33832077  0.46238778
## 2  1.1918571  0.26615071 0.1664801  0.4481541  0.06001765 -0.08236081
## 3 -1.3583541 -1.34016307 1.7732093  0.3797796 -0.50319813  1.80049938
## 4 -0.9662717 -0.18522601 1.7929933 -0.8632913 -0.01030888  1.24720317
## 5 -1.1582331  0.87773675 1.5487178  0.4030339 -0.40719338  0.09592146
## 6 -0.4259659  0.96052304 1.1411093 -0.1682521  0.42098688 -0.02972755
##            V7          V8         V9         V10        V11         V12
## 1  0.23959855  0.09869790  0.3637870  0.09079417 -0.5515995 -0.61780086
## 2 -0.07880298  0.08510165 -0.2554251 -0.16697441  1.6127267  1.06523531
## 3  0.79146096  0.24767579 -1.5146543  0.20764287  0.6245015  0.06608369
## 4  0.23760894  0.37743587 -1.3870241 -0.05495192 -0.2264873  0.17822823
## 5  0.59294075 -0.27053268  0.8177393  0.75307443 -0.8228429  0.53819555
## 6  0.47620095  0.26031433 -0.5686714 -0.37140720  1.3412620  0.35989384
##          V13        V14        V15        V16         V17         V18
## 1 -0.9913898 -0.3111694  1.4681770 -0.4704005  0.20797124  0.02579058
## 2  0.4890950 -0.1437723  0.6355581  0.4639170 -0.11480466 -0.18336127
## 3  0.7172927 -0.1659459  2.3458649 -2.8900832  1.10996938 -0.12135931
## 4  0.5077569 -0.2879237 -0.6314181 -1.0596472 -0.68409279  1.96577500
## 5  1.3458516 -1.1196698  0.1751211 -0.4514492 -0.23703324 -0.03819479
## 6 -0.3580907 -0.1371337  0.5176168  0.4017259 -0.05813282  0.06865315
##           V19         V20          V21          V22         V23
## 1  0.40399296  0.25141210 -0.018306778  0.277837576 -0.11047391
## 2 -0.14578304 -0.06908314 -0.225775248 -0.638671953  0.10128802
## 3 -2.26185710  0.52497973  0.247998153  0.771679402  0.90941226
## 4 -1.23262197 -0.20803778 -0.108300452  0.005273597 -0.19032052
## 5  0.80348692  0.40854236 -0.009430697  0.798278495 -0.13745808
## 6 -0.03319379  0.08496767 -0.208253515 -0.559824796 -0.02639767
```

CHAPTER 5

METHODOLOGY

EXISTING SYSTEM

In case of the existing system the fraud is detected after the fraud is done that is, the fraud is detected after the complaint of the card holder. And so the card holder faced a lot of trouble before the investigation finish. And also as all the transaction is maintained in a log, we need to maintain a huge data. And also now a day’s lot of online purchase are made so we don’t know the person how is using the card online, we just capture the IP address for verification purpose. So there need a help from the cyber crime to investigate the fraud. To avoid the entire above disadvantage we propose the system to detect the fraud in a best and easy way.

PROPOSED SYSTEM

In proposed system, we present a new system FDS Which does not require fraud signatures and yet is able to detect frauds by considering a cardholder’s spending habit. The details of items purchased in Individual transactions are usually not known to any Fraud Detection System(FDS) running at the bank that issues credit cards to the cardholders. Hence, we feel that FDS is an ideal choice for addressing this problem. Another important advantage is a drastic reduction in the number of False Positives transactions identified as malicious by an FDS although they are actually genuine. An FDS runs at a credit card issuing bank. Each incoming transaction is submitted to the FDS for verification. FDS receives the card details and the value of purchase to verify, whether the transaction is genuine or not. The types of goods that are bought in that transaction are not known to the FDS. It tries to find any anomaly in the transaction based on the spending profile of the cardholder, shipping address, and billing address, etc. If the FDS confirms the transaction to be of fraud, it raises an alarm, and the issuing bank declines the transaction.

CHAPTER 6

SYSTEM IMPLEMENTATION

Project implementation (or project execution) is the phase where visions and plans become reality. This is the logical conclusion, after evaluating, deciding, visioning, planning, and choosing the right method to go about this particular plan of action. Technical implementation is one part of executing a project.

```
import numpy as np

import sklearn as sk
import pandas as pd
import matplotlib.pyplot as plt
from pandas_ml import ConfusionMatrix
import pandas_ml as pdml
from sklearn.preprocessing import scale
import random

# May have to do this...
#!pip install imblearn
#!pip install --upgrade sklearn

df = pd.read_csv('creditcard.csv', low_memory=False)
df = df.sample(frac=1).reset_index(drop=True)
df.head()
frauds = df.loc[df['Class'] == 1]
non_frauds = df.loc[df['Class'] == 0]
print("We have", len(frauds), "fraud data points and", len(non_frauds), "nonfraudu
lent data points.")
We have 492 fraud data points and 284315 nonfraudulent data points.
```

Fig 6.1 Importing packages and understanding data

```
In [5]:
     ax = frauds.plot.scatter(x='Amount', y='Class', color='Orange', label='Fraud')
non_frauds.plot.scatter(x='Amount', y='Class', color='Blue', label='Normal', ax=ax
)
plt.show()
print("This feature looks important based on their distribution with respect to cl
ass.print
```

Fig 6.2 Visualize the dataset

```
_ax = frauds.plot.scatter(x='V22', y='Class', color='Orange', label='Fraud')
non_frauds.plot.scatter(x='V22', y='Class', color='Blue', label='Normal', ax=a
x)
plt.show()
print("This feature may not be very important because of the similar distribut
ion.")
```

ax = frauds.plot.scatter(x='V22', y='Class', color='Orange', label='Fraud')

```
non_frauds.plot.scatter(x='V22', y='Class', color='Blue', label='Normal', ax=a
x)
plt.show()
```

Fig 6.3 Observe readings in the dataset

```
non_frauds.plot.scatter(x='V22', y='Class', color='Blue', label='Normal', ax=a
x)
plt.show()
print("This feature may not be very important because of the similar distribut
ion.")
from sklearn import datasets, linear_model
from sklearn.preprocessing import PolynomialFeatures
from sklearn.model_selection import train_test_split
                                                                        In [9]:
X = df.iloc[:,:-1]
y = df['Class']

print("X and y sizes, respectively:", len(X), len(y))
```

Fig 6.4 Iterating list of subjects in the standing activity

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.35)
print("Train and test sizes, respectively:", len(X_train), len(y_train), "|", len(
X_test), len(y_test))
print("Total number of frauds:", len(y.loc[df['Class'] == 1]), len(y.loc[df['Class
'] == 1])/len(y))
print("Number of frauds on y_test:", len(y_test.loc[df['Class'] == 1]), len(y_test
.loc[df['Class'] == 1]) / len(y_test))
print("Number of frauds on y_train:", len(y_train.loc[df['Class'] == 1]), len(y_
X and y sizes, respectively: 284807 284807
Train and test sizes, respectively: 185124 185124 | 99683 99683
Total number of frauds: 492 0.001727485630620034
Number of frauds on y_test: 154 0.0015448973245187243
Number of frauds on y_train: 338 0.0018258032453922777
```

Fig 6.5 Classifying activity

```
logistic = linear_model.LogisticRegression(C=1e5)
logistic.fit(X_train, y_train)
print("Score: ", logistic.score(X_test, y_test))
Score:  0.998966724517
In [11]:
y_predicted = np.array(logistic.predict(X_test))
y_right = np.array(y_test)
In [12]:
confusion_matrix = ConfusionMatrix(y_right, y_predicted)
print("Confusion matrix:\n%s" % confusion_matrix)
confusion_matrix.plot(normalized=True)
plt.show()
confusion_matrix.print_stats()
Confusion matrix:
Predicted      0    1  __all__
Actual
0          99486   43    99529
1             60   94      154
__all__    99546  137    99683
```

Fig 6.6 Visualize the output

CHAPTER 7

TESTING

Machine Learning models would also need to be tested as conventional software development . Due to this scarcity of real dataset, not many fraud detection models have been developed and described in the academic literature, and even fewer are known to have been implemented in actual detection systems. Still we can find some successful applications of various data mining techniques like, neural network, Bayesian classifier, support vector machine, artificial immune system, fuzzy systems, genetic algorithm, K-nearest neighbor, and hidden Markov model, Logistic Regression in fraud detection.

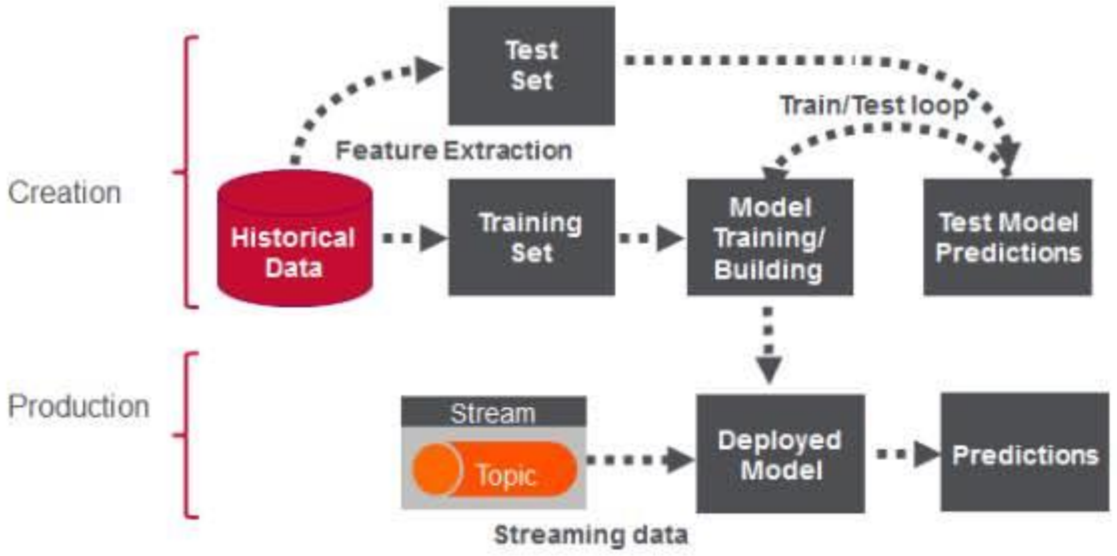


Fig 7.1 Testing method

When applied to Machine Learning models, streaming data testing would mean testing Machine Learning models without knowing the internal details such as features of the Machine Learning model, the algorithm used to create the model etc. The challenge, however, is to identify the test oracle which could verify the test outcome against the expected values known beforehand. This is discussed in the following section.

Testing techniques

- Model performance
- Metamorphic testing
- Dual coding
- Coverage guided fuzzing
- Comparison with simplified, linear models
- Testing with different data slices

In this project we are using the model performance method of testing. Testing model performance is about testing the models with the test data/new data sets and comparing the model performance in terms of parameters such as accuracy/recall etc., to that of pre-determined accuracy with the model already built and moved into production. This is the most trivial of different techniques which could be used for testing.

Advantage of testing

- Well suited and efficient for large code segments.
- Code access is not required.
- Clearly separates user's perspective from the developer's perspective through visibly defined roles.
- Large numbers of moderately skilled testers can test the application with no knowledge of implementation, programming language, or operating systems.

Disadvantage of testing

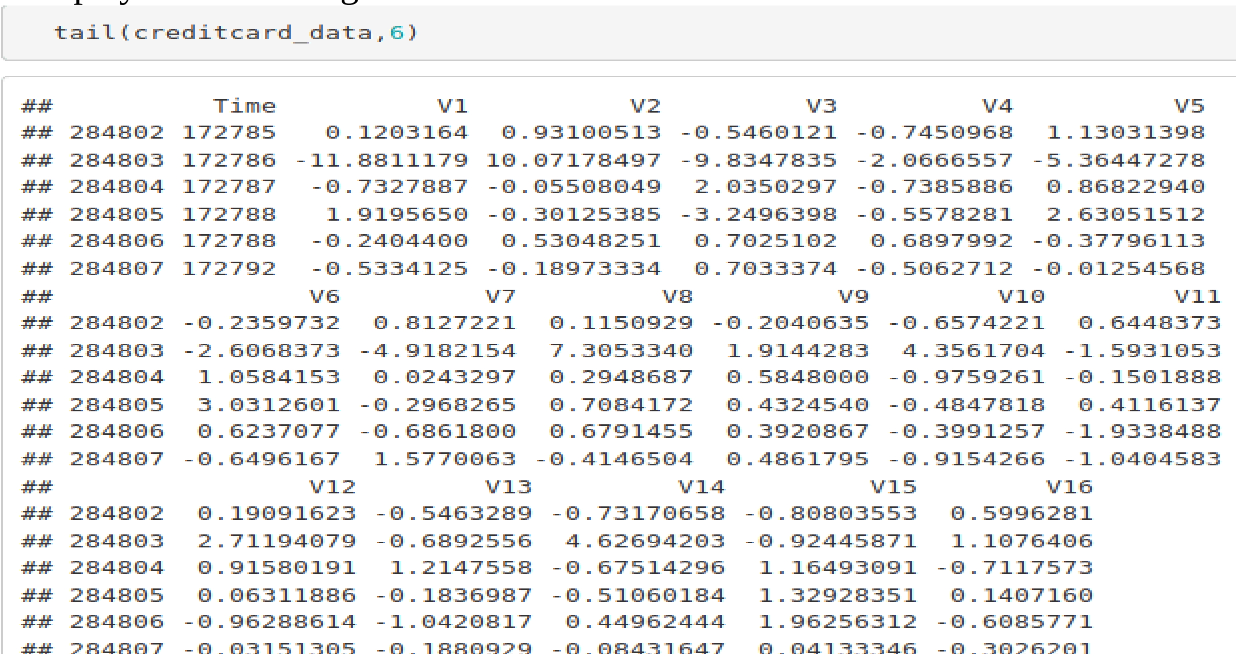
- Limited coverage, since only a selected number of test scenarios is actually performed.
- Blind coverage, since the tester cannot target specific code segments.
- The test cases are difficult to design.
- Inefficient testing, due to the fact that the tester only has limited knowledge about an application.

CHAPTER 8

RESULT AND DISCUSSION

8.1 Fraudulent activities in finance can be detected by looking at on-surface and evident signals. Unusually, large transactions or the ones that happen in atypical locations obviously deserve additional verification. Purely rule-based systems entail using algorithms that perform several fraud detection scenarios, manually written by fraud analysts.

Display of the training dataset



```
tail(creditcard_data,6)
```

```
##          Time          V1          V2         V3         V4          V5
## 284802 172785   0.1203164  0.93100513 -0.5460121 -0.7450968  1.13031398
## 284803 172786 -11.8811179 10.07178497 -9.8347835 -2.0666557 -5.36447278
## 284804 172787  -0.7327887 -0.05508049  2.0350297 -0.7385886  0.86822940
## 284805 172788   1.9195650 -0.30125385 -3.2496398 -0.5578281  2.63051512
## 284806 172788  -0.2404400  0.53048251  0.7025102  0.6897992 -0.37796113
## 284807 172792  -0.5334125 -0.18973334  0.7033374 -0.5062712 -0.01254568
##                V6         V7         V8         V9        V10        V11
## 284802 -0.2359732  0.8127221  0.1150929 -0.2040635 -0.6574221  0.6448373
## 284803 -2.6068373 -4.9182154  7.3053340  1.9144283  4.3561704 -1.5931053
## 284804  1.0584153  0.0243297  0.2948687  0.5848000 -0.9759261 -0.1501888
## 284805  3.0312601 -0.2968265  0.7084172  0.4324540 -0.4847818  0.4116137
## 284806  0.6237077 -0.6861800  0.6791455  0.3920867 -0.3991257 -1.9338488
## 284807 -0.6496167  1.5770063 -0.4146504  0.4861795 -0.9154266 -1.0404583
##                V12        V13         V14         V15        V16
## 284802  0.19091623 -0.5463289 -0.73170658 -0.80803553  0.5996281
## 284803  2.71194079 -0.6892556  4.62694203 -0.92445871  1.1076406
## 284804  0.91580191  1.2147558 -0.67514296  1.16493091 -0.7117573
## 284805  0.06311886 -0.1836987 -0.51060184  1.32928351  0.1407160
## 284806 -0.96288614 -1.0420817  0.44962444  1.96256312 -0.6085771
## 284807 -0.03151305 -0.1880929 -0.08431647  0.04133346 -0.3026201
```

Fig 8.1 Training dataset

8.1 Accuracy of four algorithms

Support Vector Classifier accuracy: 94.02782490668477%

Logistic Regression accuracy: 98.19952494061758%

K Nearest Neighbors Classifier accuracy:
90.02375296912113% Random Forest Classifier accuracy:
89.68442483881914%

CHAPTER 9

SNAPSHOTS

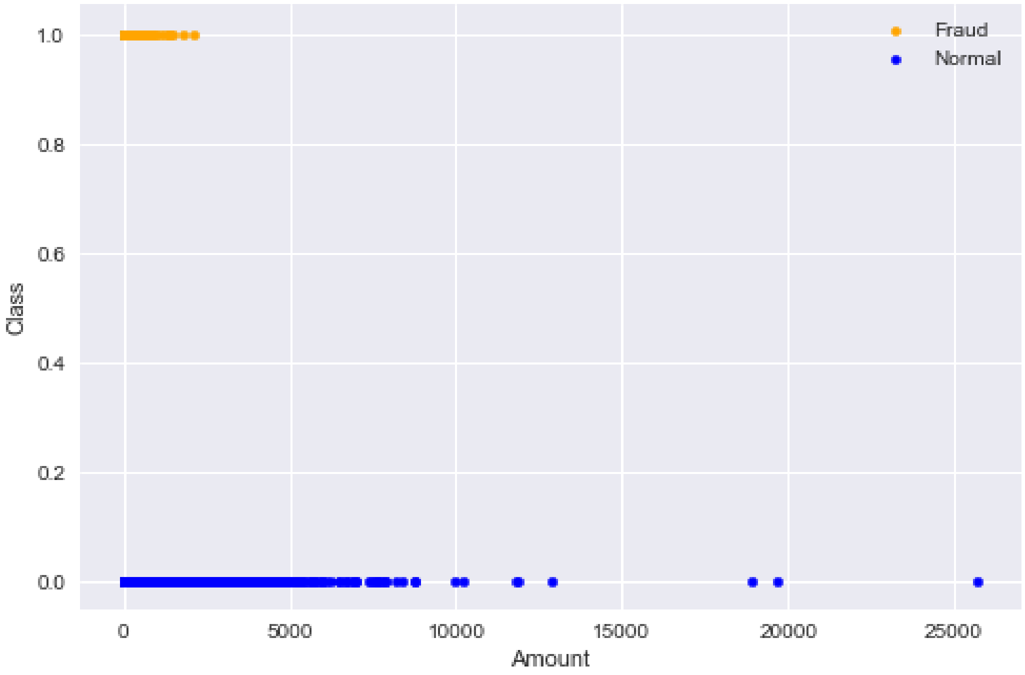


Fig 9.1 Fraud and normal dataset

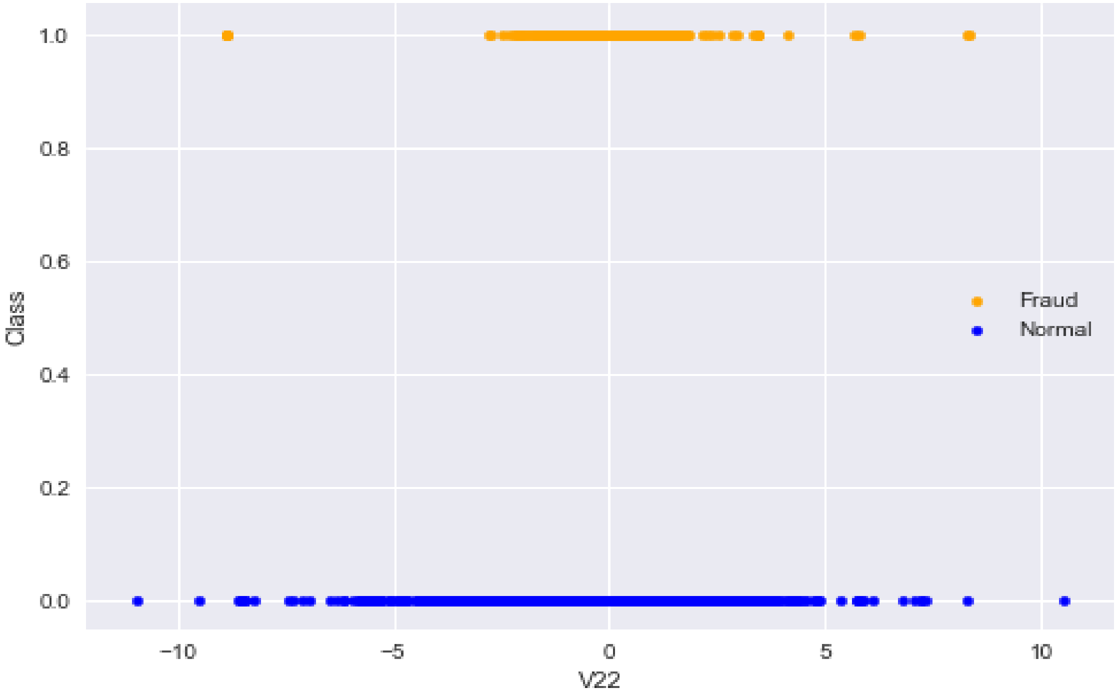


Fig 9.2 Bar graph for observing values in the dataset

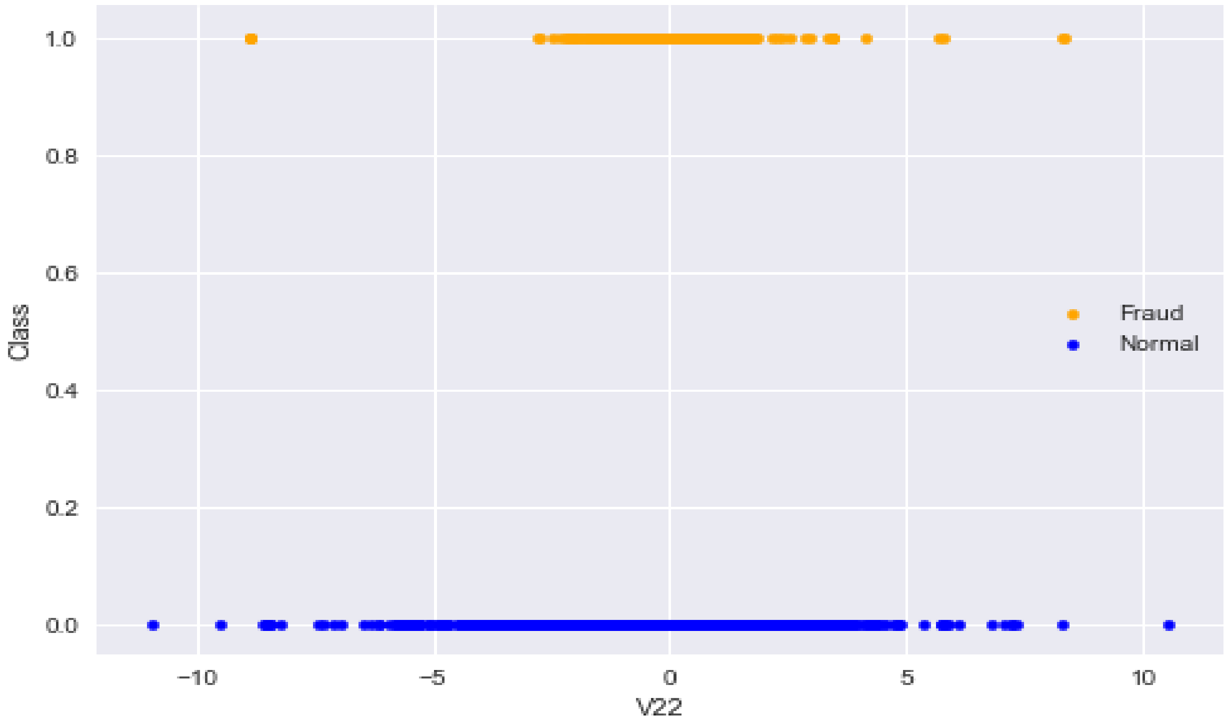


Fig 9.3 Line graph for list of subjects in standing activity

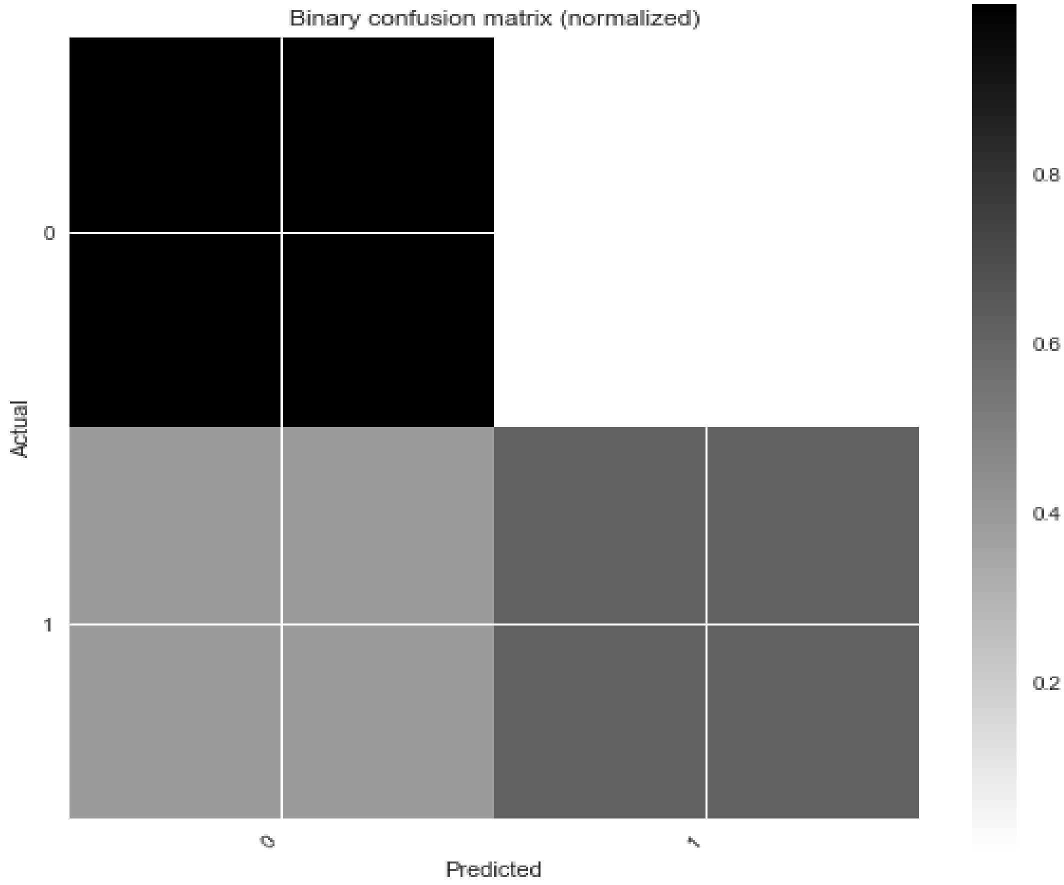


Fig 9.4 Binary confusion matrix

CONCLUSION

Since humans tend to exhibit specific behaviorist profiles, every cardholder can be represented by a set of patterns containing information about the typical purchase category, the time since the last purchase, the amount of money spent, etc. Deviation from such patterns is a potential threat to the system to detect and block from fraud transactions using a credit card. Here, I kindly convey that special feature of this software is the geniality and it can be worded on the personal computer, since the web page gives a variety option and the message gives clear understanding of the next page it is easy to follow and use .We are sure that this software will be useful for all company. this software, there is no need of knowledge of the computer operating method because, to enter into the menu just enter into windows and type the particular directory, in which the project is stored.

Machine learning technique like Logistic regression, Decision Tree and Random forest were used to detect the fraud in credit card system. Sensitivity, Specificity, accuracy and error rate are used to evaluate the performance for the proposed system. The accuracy for logistic regression, Decision tree and random forest classifier are 95.5, 90.3, and 94.0 respectively. By comparing all the three method, found that Logistic Regression is better than the Random Forest Classifier and Decision Tree.

FUTURE ENHANCEMENT

Credit card fraud is a massive problem for ecommerce retailers. To better understand, let’s do a quick review of how credit card fraud happens. If you’re an ecommerce business, you’re likely all-too-familiar with these problems, so feel free to skip ahead. Merchants lose out on valuable Customer Lifetime Value, whereby the customer has the potential to make multiple orders after the first successful order

- Merchants lose out on valuable referrals by that customer, who now is not a brand ambassador
- The customer is often times extremely frustrated and goes to a competitor. All this value is not only lost, but now also handed to a competitor.
- The customer is a black mark, and may talk or post negatively about their experience on the merchant’s site

REFERENCES

[1] [1A. C. Bahnsen, A. Stojanovic, D. Aouada, B. Ottersten, "*Cost sensitive credit card fraud detection using Bayes minimum risk*", Machine Learning and Applications (ICMLA). 2013 12th International Conference, vol. 1, pp. 333-338, 2013.

[2] B.Meena, I.S.L.Sarwani, S.V.S.S.Lakshmi," *Web Service mining and its techniques in Web Mining*" IJAEGT,Volume 2,Issue 1 , Page No.385-389.

[3] F. N. Ogwueleka, "*Data Mining Application in Credit Card Fraud Detection System*", Journal of Engineering Science and Technology, vol. 6, no. 3, pp. 311-322, 2011.

[4] G. Singh, R. Gupta, A. Rastogi, M. D. S. Chandel, A. Riyaz, "*A Machine Learning Approach for Detection of Fraud based on SVM*", International Journal of Scientific Engineering and Technology, vol. 1, no. 3, pp. 194-198, 2012, ISSN ISSN: 2277-1581.

[5] K. Chaudhary, B. Mallick, "*Credit Card Fraud: The study of its impact and detection techniques*", International Journal of Computer Science and Network (IJCSN), vol. 1, no. 4, pp. 31-35, 2012, ISSN ISSN: 2277-5420.

[6] M. J. Islam, Q. M. J. Wu, M. Ahmadi, M. A. Sid- Ahmed, "*Investigating the Performance of Naive-Bayes Classifiers and KNearestNeighbor Classifiers*", IEEE International Conference on Convergence Information Technology, pp. 1541-1546, 2007.

[7] R. Wheeler, S. Aitken, "Multiple algorithms for fraud detection" in Knowledge-Based Systems, Elsevier, vol. 13, no. 2, pp. 93-99, 2000.

[8] S. Patil, H. Somavanshi, J. Gaikwad, A. Deshmane, R. Badgujar, "*Credit Card Fraud Detection Using Decision Tree Induction Algorithm*", International Journal of Computer Science and Mobile Computing (IJCSMC), vol. 4, no. 4, pp. 92-95, 2015, ISSN ISSN: 2320-088X.

[9] S. Maes, K. Tuyls, B. Vanschoenwinkel, B. Manderick, "*Credit card fraud detection using Bayesian and neural networks*", Proceedings of the 1st international naiso congress on neuro fuzzy technologies, pp. 261-270, 2002.

[10] S. Bhattacharyya, S. Jha, K. Tharakunnel, J. C. Westland, "*Data mining for credit card fraud: A comparative study*", Decision Support Systems, vol. 50, no. 3, pp. 602-613, 2011.

[11] Y. Sahin, E. Duman, "*Detecting credit card fraud by ANN and logistic regression*", Innovations in Intelligent Systems and Applications (INISTA) 2011 International Symposium, pp. 315-319, 2011.

[12] Selvani Deepthi Kavila,LAKSHMI S.V.S.S.,RAJESH B " *Automated Essay Scoring using Feature Extraction Method* " IJCER ,volume 7,issue 4(L), Page No. 12161-12165.