Analysis of Credit Card Fraud Detection using Data Mining and Machine Learning Techniques

Authors: Farhaan Patel, Jessica Mathias, Priya Yadav, Rama Tejaswini Thotapalli, Sai Ashrith Aduwala

Abstract **--One of the main and a popular payment mode is the use of credit cards. This advent of cashless transactions has made it easy and convenient for users to make payments. Increased growth in its usage has directly affected the fraudulent transactions which increased the frequency of illegal activities. Statistics suggest that the losses incurred by these illegal activities surmount up to billions of dollars each year. The fraud transactions are carried out so gracefully that, to the common eye they look like any other genuine transaction. Thus, many banking and economic sectors have started to depend on technology to fight back these illegal transactions. Many machine learning applications have been developed to counter the fraud, but along with the improvement of these applications, the tricks used by the masterminds behind these activities have also improved. In this project, we try to implement machine learning algorithms like Logistic Regression, SVM, Decision Tree and K-Means along with unsupervised learning models and deep neural networks to analyse which of these implemented applications better suits the purpose for solving the problem of credit card fraud detection.**

I. INTRODUCTION

Loss due to credit card fraud is a major ongoing problem for companies across the world. Companies lose billions of dollars in revenue as a result of fraud. Even though several technologies such as Card Verification Code, Chip and Pin Verification etc. exists, the fraudsters find their way around this advanced system and hence, automatic detection of fraudulent methods is of utmost importance.

Even though several machine learning algorithms exist for fraud detection, the percentage of fraudulent transactions when compared to legitimate transactions is trivial. Hence training the algorithms with such an unbalanced set is a challenge. The model designed also needs to be dynamic enough to adapt to new fraud patterns on a constant basis.

Several approaches in the past have extensively applied machine learning algorithms to data from banking and credit card operations, but very few studies have been directed towards credit card fraud detection from the perspective of an online merchant. The difference between the banking sector and online merchants is that the online merchant has inadequate information about the customer who makes transactions on his site.

For this project we try to exploit the available data from the credit card transaction logs to implement multiple fraud detection systems. The aim is to examine the combination of these manual and automatic classification based on data. The proposed approach involves building various risk scoring system founded on machine learning techniques which will approximate a fraud suspicion score for each order. Also, the purpose of the risk scoring system is to evaluate the orders that are approved in spite of falling below a certain threshold. The results obtained from these machine learning techniques will be subjected to various validation methods. An analysis of which of these implemented techniques is better applicable for fraud detection, is done based on the validation results obtained.

The following sections of the report are organized as follows: Section II and III will cover the problem statement and literature review respectively. Section IV will cover in detail the different techniques and

methodologies used to implement this project. Section V contains the results obtained and the findings of the analysis performed. Section VI and VII will talk about various tools used and the salient knowledge gained from the project. Section VIII concludes the report.

II. PROBLEM STATEMENT

With the advent of modernization, the advancement of technology in our everyday life has exponentially increased. In the finance sector with the introduction of credit cards, sales and purchases of goods has become easy and convenient. Everyone nowadays has his own credit card. While some people use technology for the betterment of society, there are others who use it for their own personal good. Similarly is the case of credit cards. Identity thefts and fraudulent transactions are very common forms of credit card fraud. Even with many strict and complex security measures taken, people always find a way to trick the security into making them think it is a validated access. Thus, many financial establishments have turned to machine learning and artificial intelligence to provide with an efficient and cost effective solution for this unending problem. Our project is to implement and analyse these solutions and come up with an estimate of which solution is better suited for this problem.

III. LITERATURE REVIEW

[1] N.Malini et al [1] the author uses K-Nearest Neighbor Algorithm (KNN) along with unsupervised outlier detection technique to identify fraudulent credit card transactions in the banking domain. The author uses the distance metric to calculate the nearest point of any incoming credit card transaction to the other transactions. The Fraudulent transaction would be an anomaly on the KNN technique. The author has preferred the unsupervised method for anomaly detection to capture the unseen type of illegitium accuracy with limited memory.

K. R. Seeja et al [2] the author put forward a model to detect fraudulent credit card activity in a high degree of disproportionate and anonymous sample datasets. The author has used the frequent itemset mining technique to find legal versus fraudulent transactions. This technique is presented as similar to Apriori algorithms, which returns the set of recurring activity sets for each mate activity. The author also compares the different fraud detection techniques based on their advantages and disadvantages to proves KNN has maxim customer. A matching algorithm is used to identify the best matching pattern for the incoming transaction. The paper has used the Matthews correlation coefficient to ensure binary classification. According to the author, this model has high accuracy for fraud detection and less false positive or false negative classification rate in imbalanced datasets.

Sahil Dhankhad and three other writers [3] given a solution to a serious financial service issue costing billions of dollars annually. Using supervised machine learning methods with the real-world dataset, they have implemented many techniques to identify credit card fraud. In this article, they employed these algorithms to introduce super classifier using ensemble methods. This article offers the distinction between 10 algorithms of classification and their accuracy comparison. Authors used to evaluate the model using Accuracy, F1-Score, Recall, Precision, G-Mean, FPR, TRP techniques. Authors recognized the most significant features that in credit card fraudulent transaction detection may lead to greater precision of assessment models.

Samuel A. Oluwadare and three other writers [4] presented a model for implementing and improving credit card fraud detection model precision as the number of transactions and information size grows. In this paper, the dataset is sourced from European cardholders containing 284,807 transactions. To detect credit card fraud and provide accuracy, sensitivity, specificity, Matthews correlation coefficient and balanced classification rate, three Naïve Bayes, k-nearest neighbor 4 and logistic regression models are implemented. Based on the results, they claimed the KNN works better than other methods based on the outcomes.

Linda Delamaire and two other writers [5] have identified different types of credit card fraud and reviewed techniques used for fraud detection. In this paper, various published findings in credit card fraud detection have been compared and

analysed. Depending on the type of fraud that has taken place, different measures proposed can be adapted. The proposed methods are cost effective and time efficient. The main aim of these techniques put forth is to minimize credit card fraud, but the authors still faced a problem when genuine transactions were also classified as fraudulent.

Yiğit Kültür and one other writer [6] have analysed the cardholder spending behaviour and proposed a card holder behaviour model to detect credit card fraud. The paper started off talking about the existing rule based models for credit card fraud detection and how many of these models are ignoring a crucial aspect of the problem i.e. the cardholder behaviour. Then a behaviour based model was proposed, which inferred the transaction rules determined by the human fraud experts. They called this model Cardholder Behaviour Model(CBM).

IV. RESEARCH METHODOLOGY

4.1 Data Collection

4.1.1 Volume

The dataset used for the analysis of credit card detection in this paper contains data from European credit card holders consisting of rows of transactions made by credit card transactions. The total number of transactions captured were 500,000 and the number of features captured were 320. Data preprocessing was done to drop the missing values. Principal component analysis was done to determine the most relevant features. The result of data preprocessing yielded 284,807 records and 31 most prominent features were chosen. The features included 28 masked features which are intentionally masked by the data source, and 'time' and 'amount' of the transaction.

4.1.2 Velocity

Velocity of a dataset signifies the speed at which data is generated. The number of credit card transactions in a country at any given day could be considered as data generated with high velocity. This project too works with high velocity data of credit card transactions captured over a period of 2 days from European card holders

4.1.3 Variety

The data for credit cards transactions falls into the sensitive category which makes it difficult to find an authentic dataset on the internet. Also, all the credit card transaction datasets which are available have the majority of features masked which makes it difficult to combine datasets from different sources to increase the variety. The dataset used in this paper is provided by IEEE-CIS in conjunction with Vesta Corporation which is a payment service company.

4.1.4 Veracity

Veracity refers to the truthfulness of the data being analyzed. It also refers to the quality of data that is being processed and analyzed. Veracity plays a major role in contributing to a meaningful output which could actually make accurate predictions of future data because of the quality of data that was used as input. The veracity of this dataset is high owing to the fact that it has been used in several IEEE research papers which can be authenticated. Also the dataset is prepared by a payment service company called Vesta whose existence and authenticity are validated

4.2 Data Visualization

4.2.1 Data Exploration

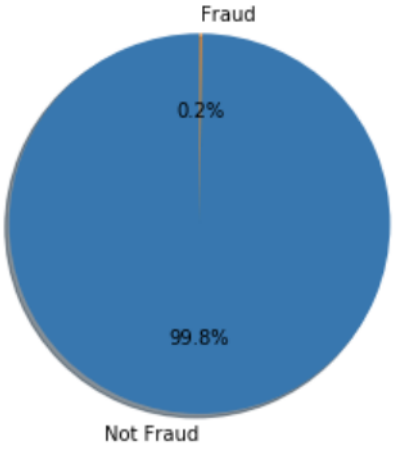

Fig 4.2.1

The dataset after preprocessing consisted of 284,000 rows out which only 0.2% or 492 rows consisted of fraud transactions whereas the rest of the rows had

genuine transactions. This shows that the data is highly imbalanced with respect to the target Class.

4.2.2 Time Density Plot

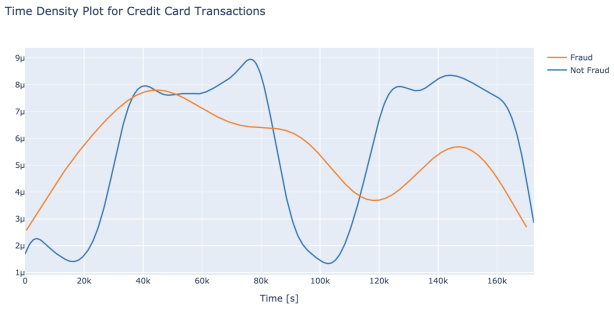

Fig 4.2.2

The time density plot was used to visualize the distribution of the transactions with respect to time. From the graph, the distribution of transactions can be seen over a continuous time period. From the graph it can be observed that transactions which are fraudulent, have a distribution more than that of authentic transactions transactions i.e they are distributed in the European time zone equally over a time period which include the dip in the number of transactions during the night.

4.2.3 Transaction amount Vs Transaction time

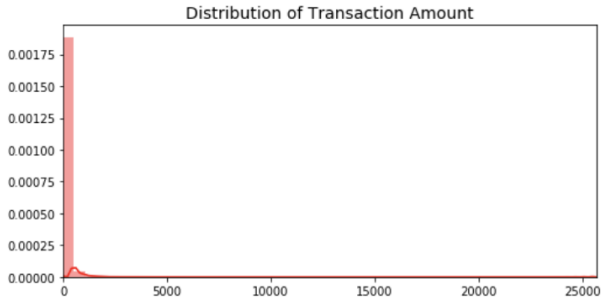

Fig 4.2.3

The above plot shows us that most of the transactions are below 5000 pounds with a long tail tending towards 25000 pounds.

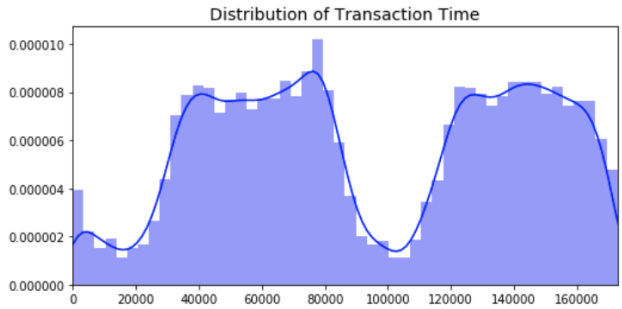

Fig 4.2.4

The transaction amount shows a continuous distribution of transactions over time with transactions dipping at night

4.2.4 Boxplots showing fraudulent transactions

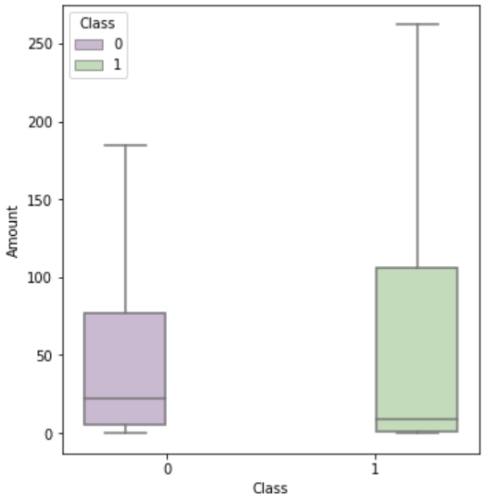

Fig 4.2.4

The above figure indicates that most of the fraud happens for amounts above 100 pounds which points to the proportional relationship between the transaction amount and the happening of fraud

4.2.5 Correlation Plot

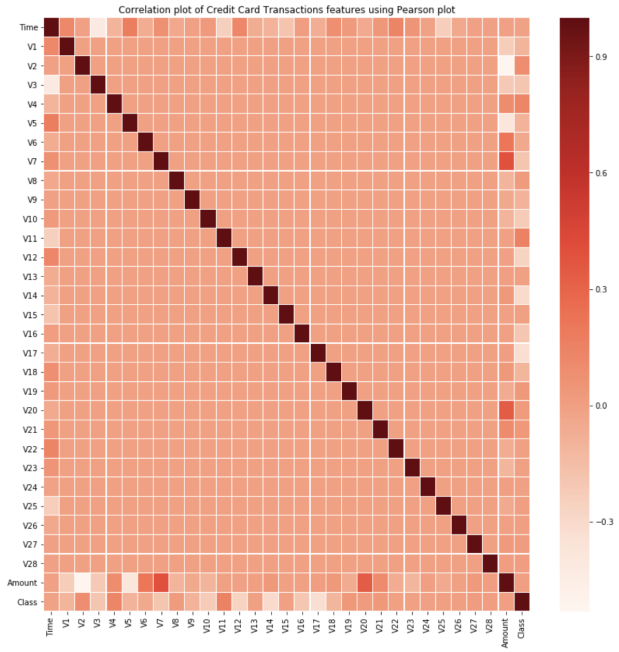

Fig 4.2.5

The correlation plot shows the correlation between various features. A correlation close to 1 indicates a strong positive relationship between the variables. A correlation close to -1 indicates a

strong negative relationship. If the value is 0, there exists no relationship between the variables. From the graph based on the darker shade of cubes, it can be made out that there are certain correlations between features such as Time (which has an inverse correlation with V3) and Amount (which has a direct correlation with V7 and V20, and an inverse correlation with V1 and V5).

4.2.6 Feature Importance

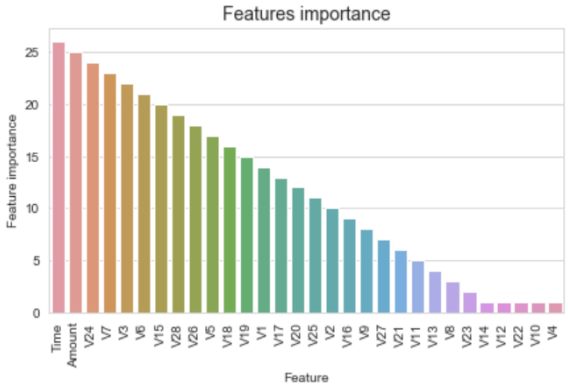

Fig 4.2.6

A regression model such as logistic regression was used to plot the Feature Importance graph. Using the feature importance graph it can be observed that the Time, Amount, V24, V7, V3, V5, V15, V28 and V26 are the more prominent features to be considered after data reduction

4.3. Data Preprocessing

4.3.1. Data Cleaning:

In the Data Cleaning step, handling of null values, smoothing, reducing the noisy data, dropping of correlated columns, dropping some rows with more number of null values was taken care of. In the dataset, there are many null values, for which the columns have been dropped by keeping the null value threshold greater than 0.7 and number of null values in a column as 250000. The correlated columns have also been dropped

4.3.2. Memory Reduction:

As the dataset is very large, it needs a storage of at least 150 MB. In order to reduce the memory usage, memory reduction techniques such as converting the all-in columns into 8, 16, and 32 bits by comparing the columns values within the range of max and min values of columns have been performed.. So, if the cell values contain more memory values in between, those cell values, they are changed to the corresponding max and min types of the column. By doing this memory reduction technique, the memory usage is reduced to 94.61 MB that is 74.4% reduction in the data.

4.3.3. Sampling:

Data Sampling is the statistical analysis technique used to select, manipulate and subset the data points. In this dataset, there are only 0.17% of fraudulent transactions whereas 99.83% of the dataset are non-fraud transactions, so, sampling helps to balance the class field of the dataset to increase the accuracy. In this dataset, under sampling of the data was performed as there are a smaller number of fraud transactions. So the stratified shuffle split sampling was used to select an equal number of samples of fraud cases and non - fraud samples. This will create a dataset with 50% of fraud cases and 50% as non-fraud cases.

Fig 4.3.3

4.3.4. Dimensionality Reduction/Feature reduction:

Feature Reduction is the same as dimensionality reduction which is the process of reducing the number of features or random variables. It can be divided into feature selection and feature extraction. As there are so many unwanted features in this dataset, the feature reduction techniques like Principle component analysis were used. Using the Principle component analysis on the dataset, the features were reduced from 82 to 43. Based on the decreasing values of eigenvalues the first 28 features were chosen as the principal components of the data.

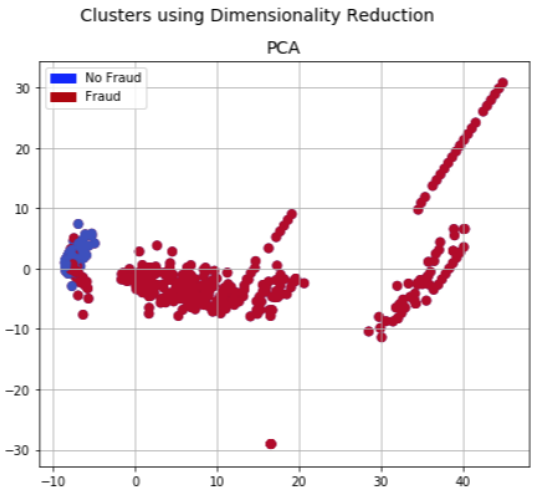

Fig 4.3.4

4.3.5. Feature selection:

Feature selection is a technique of selecting the most important features of the dataset. In this dataset, after doing the dimensionality reduction, 43 columns of the data were obtained. From those features first 28 columns were selected as prominent data since most of the data was represented by these columns.

4.3.6. Outliers detection:

Outlier detection is the main part of the credit card fraud detection. If there are more number of outliers there is more risk of loss of information and it reduces the accuracy of models. So, for this dataset, the extreme outliers were detected using box plots and removed by using the formula of interquartile range.

4.4 Implementation Methods

4.4.1. Logistic Regression:

Logistic Regression models the probabilities for classification problems with two possible outcomes. It is an extension of linear regression model for classification problems. Instead of fitting a straight line or hyperplane, the model uses the sigmoid function to map the output of linear function between 0 and 1. Sigmoid function is defined as S(x) = ex /ex+1. It is used when the dependent variable i.e. the target class is categorical, or it is used to predict the probability of categorical dependent variable

4.4.2. SVM:

Support Vector Machine is a discriminative classifier defined by a separating hyperplane. Given the training data, the algorithm gives an optimal hyperplane which categorizes any further data given to it. There will be infinite possible hyperplanes separating the two classes of data, but SVM searches for the hyperplane with the maximum margin that is maximum marginal hyperplane. This is one of the most efficient algorithms present because it is unaffected by errors in the data like outliers etc. as the classifying hyperplane depends only on a few points on the plane called Support Vectors.

4.4.3. Decision Trees:

Decision trees is the most intuitive one among all the other machine learning algorithms. It is a supervised learning algorithm that can be used for solving both regression and classification problems. It solves the problem by representing the given data and the attributes as a tree. The internal nodes of the tree are the attributes, the branches are the conditional statements and the leaf nodes are the target classes. The order of attributes in the tree is decided by calculating the contribution of that attribute using methods like Information Gain and Gini Index.

4.4.4. K-Means:

K-means is one of the simplest unsupervised machine learning algorithms which classifies the given data into k different clusters based on a distance metric. It identifies k different centroids and assigns each point to different centroid based on how close the point is to the centroid. These clusters are aggregation if such points that have some similarity. The goal of this algorithm is to minimize the intra-cluster distance and maximize the inter-cluster distance or to maximize intra-cluster similarity and minimize inter-cluster similarity.

4.4.5. KNN:

KNN algorithm assumes that similar things are in close proximity to each other. KNN is a lazy learner algorithm, which means it simply stores the data and waits till the test data arrives to train the model. In this approach the data instances are represented as points on a Euclidean space. We classify the point based on the majority of the classes of its neighbors. The classes are defined based on a distance metric. The distance from the test point to its nearest k neighbour points are calculated and it is classified based on the point which is closest.

4.4.6. Neural Networks:

Neural Networks are a collection of algorithms, which are designed to work as a human brain, that learn to recognise patterns. They take sensory real-world data like images, text, audio etc. which are then converted into machine perceptible data like numbers, vectors etc.. It stores this existing knowledge and uses them when needed. In general, a neural network has three layers: Input Layer, which takes the data; Hidden Layer, which performs the learning operation; and Output Layer, which gives the result obtained from the operation.

V. EXPERIMENTAL RESULTS AND ANALYSIS

The machine learning models which were mentioned above were trained and subjected to evaluation metrics such as Accuracy Score, Area Under Curve, Confusion Matrix, Cross-Validation score, Precision and Recall. In this project, the accuracy of all the methods implemented are compared using plots. The ROC curve , Precision and Recall graphs, Confusion matrix of all the methods are plotted.

5.1. ROC -AUC Curve:

AUC-ROC curve is an evaluation metric for classification problem at various thresholds settings. ROC is a probability curve and AUC (Area under curve) represents degree or measure of separability. Higher the AUC value, the model is better at classifying positives and negatives. True positive rate and False positive rate can be calculated from the below table.

In (Fig 5.1) is the ROC-AUC comparison of all the supervised methods implemented.

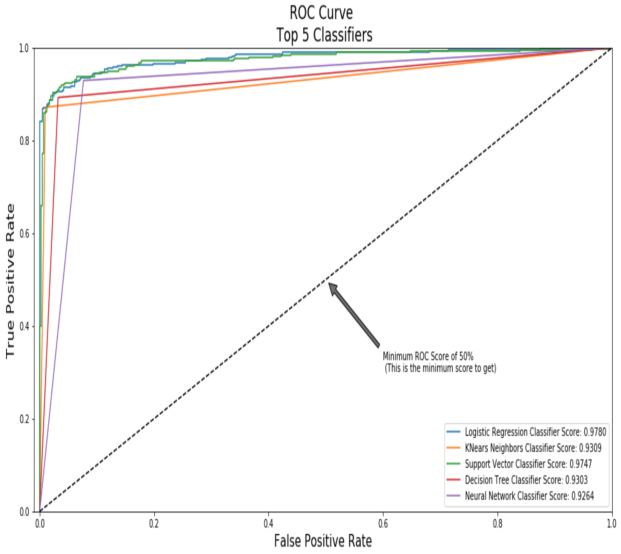

Fig 5.1

5.2. Confusion Matrix :

Confusion matrix is the statistical form of representing the classification by providing the number of correct and incorrect classifications. The matrix will have true positives, true negatives, false positives and false negatives.

The confusion matrix comparisons for the implemented methods are as below.

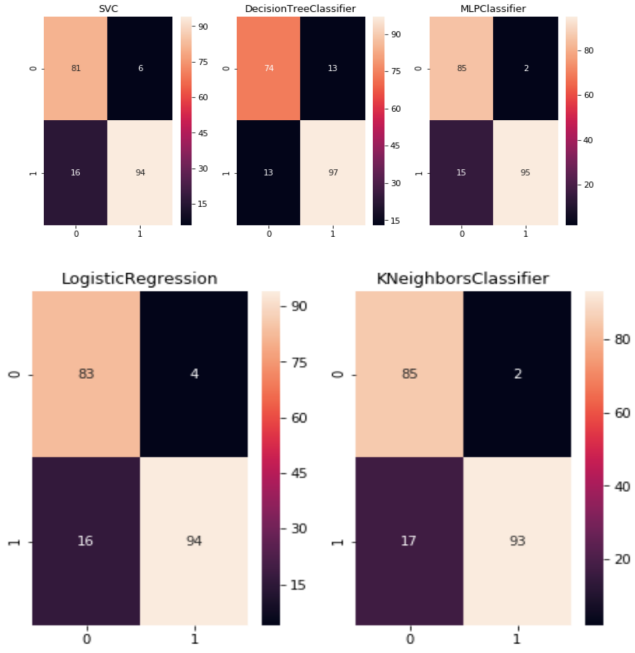

Fig 5.2

5.3. Precision and Recall Curve:

Recall: Ability of a classification model to identify all relevant instances

Precision: Ability of a classification model to return only relevant instances

Precision and Recall can also be calculated using above confusion matrix with the below formulae.

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Recall} = \frac{tp}{tp + fn}$$

Below is the plot of comparison of precision vs Recall of the methods implemented for credit card fraud detection.

Among all the classifiers logistic regression has score of 0.9829 which is very close to the optimum value.

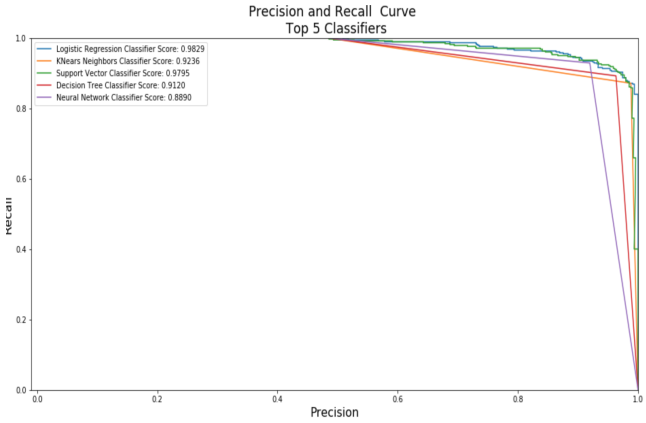

Fig 5.3

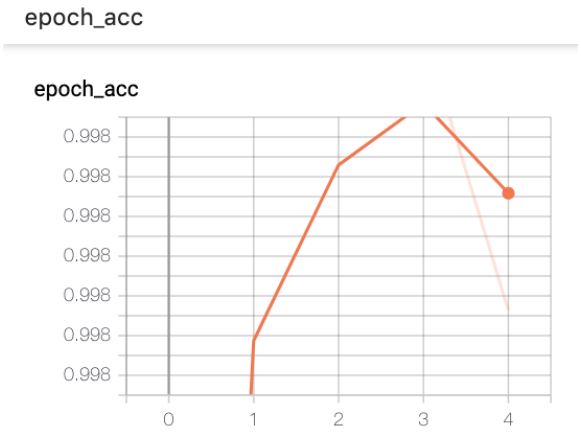

Fig 5.4

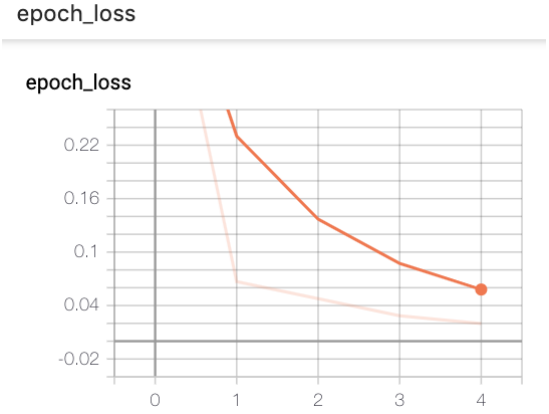

Fig 5.5

Parameter Tuning: Based on the preprocessing and model training, The accuracy of around 93% is obtained using the above 5 models. To increase the accuracy of the models GridSearchCV sklearn tools were used to tune the parameters of all the models and 5-fold cross validation. Then the accuracy of all the algorithms increased by 5%.

Based on the comparisons of the efficiency of the algorithms and analysis, In this dataset the Logistic regression classifier had highest training accuracy among all other classifiers that is 98.29%. Below is the screenshot of accuracy comparison of all the models. The training also obtained highest test accuracy of 91.5% for the neural network classifier and an accuracy of 99.83% using tensorflow keras model. As the out of sample error is more important than the in-sample error neural networks is the best algorithm.

VI. DEPLOYMENT (AWS Sage Maker)

In In this project, Amazon sage make is usedr for the cloud machine learning model deployment. Amazon Sage Maker provides the ability to build, train, and deploy machine learning models. Amazon Sage Maker is a service that covers machine learning workflow to label, data preprocessing, choosing an algorithm, train the model, parameter tuning and optimization for the deployment. For this project a sage maker instance was created on Amazon web services. Using this instance, a docker file was created in jupyter notebook and a decision tree model was used to train the dataset. Once the training was completed the docker was deployed and the resulting model classified the data with more accuracy than the models used in supervised learning. The model can also be fetched using an API. The data sent through the API correctly returned the classification of the data as a result. The importance of sage maker is

that the cost of labelling is reduced since the dataset does not need to be labelled when fed to the model.

VII. TOOLS USAGE

7.1. Anaconda

Anaconda is a platform that provides an integrated environment for data mining development and implementation using Python, R, tensorflow etc.

7.2. AWS Sage maker

Amazon AWS Sage Maker is a platform that is professionally operated. It performs the entire machine learning workflow. Using AWS Sage Maker, with far less effort and lower cost, our models get to production faster. Additionally, we can provide API to connect to our project.

7.3. Tensor Flow

It is an open source library used for machine learning applications.

7.3. Google Collab

It is a cloud service which provides python programming platform along with free GPU. Google Collab was used to preprocess the large amount of data in our dataset

VIII. CHALLENGES FACED

8.1 The data for credit cards transactions falls into the sensitive category which made it difficult to find an authentic dataset on the internet.

8.2. Size of the Data set: The size of the available data is really less because of security and confidential reasons.

8.3. Determining the credit card fraud and having a higher accuracy for the same is difficult as the dataset is highly imbalanced. The error cost of misclassifying fraudulent is higher than the error cost of misclassifying legitimate instances.

8.4. Dynamic behavior of the fraudulent: Fraudulent frequently change their behavior after every fraud for fear of not getting caught. Having a Machine Learning model to determine this dynamic behavior is a difficult task.

8.5. Faced a lot of technical issues while deploying a project on AWS Sage maker due to region specific deployments.

IX. LESSONS LEARNT

9.1. Data preprocessing is a very important step in machine learning workflow. Initially only 60% accuracy could be obtained due to lack of proper data preprocessing. After performing proper preprocessing of data it was possible to achieve an accuracy of more than 90%.

X. CONCLUSION & FUTURE WORK

The occurrence of credit card fraud has increased at an alarming rate in recent years since the hackers find a way even around the most sophisticated security checks that the banks have in place. The merchant's risk management level can be improved by building an automated, efficient and accurate risk assessing and monitoring system which is of primary importance to all credit card companies and merchant banks

In this study, four supervised models, one unsupervised model and neural networks were used in order to build and train a credit card fraud detection model. Section V and section VI explained the accuracy obtained using various models. Through this study it was found that the neural networks and the model for logistic regression outperformed the other models in classifying the transactions as fraud or genuine. It was also found that the unsupervised model K-means did not provide great accuracy in spite of passing the data through several preprocessing techniques

This framework can be used by credit card companies to run the transaction through several comparative models and also compare to historical transaction patterns. Thus this study provides issuers with an automated and intelligent model to eliminate credit card fraud

XI. REFERENCES

[1] N. Malini, M. Pushpa "Analysis on credit card fraud identification techniques based on KNN and outlier detection" in 2017 Third International Conference on Advances in Electrical, Electronics, Information,

Communication and Bio-Informatics (AEEICB),27-28 Feb. 2017 IEEE

[2] K. R. Seeja and Masoumeh Aerator “Fraud Miner: A Novel Credit Card Fraud Detection Model Based on Frequent Itemset Mining”, The Scientific World Journal, Volume 2014

[3] Sahil Dhankhad ; Emad Mohammed ; Behrouz Far “Supervised Machine Learning Algorithms for Credit Card Fraudulent Transaction Detection: A Comparative Study” 2018 IEEE International Conference on Information Reuse and Integration (IRI)

[4] John O. Awoyemi ; Adebayo O. Adetunmbi ; Samuel A. Oluwadare “Credit card fraud detection using machine learning techniques: A comparative analysis” 2017 International Conference on Computing Networking and Informatics (ICCNI)

[5] Delamaire, Linda & Abdou, Hussein & Pointon, John. (2009). “Credit card fraud and detection techniques: A review.” Banks and Bank Systems. Volume 4. Issue 2. 2009

[6] Yiğit Kültür, Mehmet Ufuk Çağlayan, "A novel cardholder behavior model for detecting credit card fraud", Application of Information and Communication Technologies (AICT) 2015 9th International Conference on, pp. 148-152, 2015.

XI. APPENDIX

Credit_Card_Fraud-master.zip