Detecting Fraudulent Credit Card Transactions Using Outlier Detection

Surya Teja Marella, K.Karthikeya, Saiteja Myla, M.Mohan Sai , Vamseekrishna Allam

ABSTRACT: Credit card frauds transactions are becoming more frequent day by day and it is becoming more difficult for humans to analyse fraudulent transactions by analysing transaction hence it has become necessary for humans to develop an intelligent system to determine fraudulent transactions. The technique we applied to determine fraudulent transactions are anomaly detection (Outlier detection).
Several intelligent algorithms can be used in this context for anomaly detection (outlier detection), In this paper we implemented Decision Tree algorithm, Random Forest and Neural Networks to determine which algorithm is best fit in terms of time taken and accuracy. We were able to formulate results of 284,407 transactions over a period of two days in September 2013. We were able to identify that three models are almost equal when it comes to accuracy but random forest is more precise.
Index Terms:Anomaly Detection, Credit Fraud Detection, Outlier Detection, Financial Analytics

1 INTRODUCTION

Fraud detection in any sector is extremely important in any sector to prevent crimes, hence it becomes a viable option to make machines perform this tedious and complicated task of fraud detection over humans, This methodology of using machines to determine fraudulent transactions can be applied in many sectors like stock market, Credit card transactions, mobile conversations, balance sheets etc in this paper we were able to apply intelligent algorithms to determine fraudulent transactions in credit card transaction data. If you wish to apply this technology in real life scenario it becomes very important that the technique is both time efficient and is accurate so that nor customers neither officers get false predictions hence we analysed and compared three different techniques to determine the suitability. The technique which we used to determine fraudulent transactions from given transaction data is anomaly detection (Outlier detection) this uses intelligent algorithms to determine the anomaly data from given data set. The intelligent algorithms initially take the given dataset and train the model so as to predict the anomaly data from dataset, Our dataset is highly unbalanced with very less fraudulent transactions and more valid ones hence before applying any technique is important to perform either under sampling and over sampling on data to make extrapolating the data easier. It is also important to determine the relation between various components in the dataset for which correlation matrix can be applied which can be plotted on a heatmap. We also have labels present in dataset which can be helpful to determine the validity of transaction and hence can be used for supervised learning, We can also remove the labels and use unsupervised learning on the given data set, to get the sense of dataset we performed various data visualization methods like bar graph, heatmap, line graph etc. For unsupervised learning we used Decision Tree technique and Random Forest, in both techniques we performed data pre-processing in initial stages and performed data visualization to gather the sense of data and in the end generated classification report to determine the efficiency of the technique. For supervised learning we initially applied data visualization like histogram to gather the sense of data and later applied under sampling in the data pre-processing and line graph, confusion matrix and f1 score to determine the efficiency of the technique. We also used other parameters like number of false positives, false negatives, true positives, true negatives correlation matrix ,f1 score to determine the accuracy of the technique. Ginny Y. Wong et al. [1] predicted protein-ligand Binding Site by making use of Support Vector Machines. Chao Wang et al. [2] made use of Improved Random Forest algorithm to classify the Network traffic. Ana Erika Camargo Cruz et al. [3] made use of regression models to predict faulty codes in software projects. Surya Teja Marella et al.[4] gave analysis of improving data centers by concept of virtualization. Gunasekhar.T et al. [5] provided systematic analysis od load balancing algorithms in cloud computing. Paul.A et al. [6] proposed improved version of random forest algorithm. Zhang.X et al. [7] applied SVM and least square vector machine to predict freight volume. Liu.Y et al. [8] applied logistic regression to predict telecom customer churn data. Vatrapu.R et al.[9] proposed set theory approach for big data analytics. Rind.A et al.[10] proposed visual library for time oriented data. Jabbar.S et al [11] proposed a methodology of real time data fusion for localized big data analytics. Huang.Z et al.[12] applied schema theory data engineering for Big Data Analytics. Lepenioti.K et al.[13] proposed a literature review of predictive analysis. Liang.T et al. [14] proposed a bibliometric study of B.I and big data analytics. Seng J.K.P et al. [15] proposed SC-LDA for integration towards decision analytics. Herbrich et al. [16] proposed true skill a new rating mechanism. Solo.A.M et al.[17] proposed an overview of new interdisciplinary field of Political engineering and computational politics. Qin-Zhang et al.[18] proposed a new modelling and simulation of politics in crisis. Marsland.S et al. [19] proposed an algorithmic perspective of machine learning. Polson.N.G et al. [20] proposed Deep learning as baysian perspective. Haykin.S.S et al. [21] proposed neural network and learning machines. LeCun.Y et al. [22] proposed deep learning. Bengio.Y et al. [23] proposed deep learning architectures. Mahadevan.S et al. [24] proposed probability and reliability of statistical methods in engineering designs. Oberkamph.W.L et al. [25] proposed Verification, validation and predictive capability in computational engineering and physics.

2 PROBLEM DEFINITION

Minimizing the number of false positives and false negatives while dealing with any predictive system is absolute necessity. Hence, It becomes evident that human judgement can be prone to errors in this regards several of the predicting systems are currently being replaced by intelligent algorithms as we observe through our analysis intelligent algorithms are efficient in performing laborious tasks and are less prone to errors hence it is important for us to design these systems which correctly utilizes these intelligent algorithms to minimize false predictions. In this context we took up task of making

machines predict the fraudulent transactions in credit card data as we know that the percentage of fraudulent transactions when compared to valid ones is very low, Therefore it becomes a challenging task for system to predict these low number of fraudulent transactions. The task of predicting such data values from the given data set is called Outlier Detection(Anomaly Detection), We can use several kinds of intelligent algorithms in this respect to predict the outliers in the given data. We have used three different intelligent algorithms to predict fraudulent data from the given data set two are unsupervised in nature(Random Forest and Random Decision Tree) and one Unsupervised(Neural Networks). The system initially gathers the sense of data and then uses these methodologies to predict fraudulent transactions. We have also presented a comprehensive analysis of using each algorithm and compared them with each other on accuracy basis.

3 HARDWARE AND SOFTWARE REQUIREMENTS

Requirements are the minimal configurations of a device and softwares required for the model to work properly and efficiently. Hardware requirements include system configurations and hardware required for the model and software requirements include Operating Systems and applications required to build the model.

3.1 Hardware requirements

Graphics Processing Unit (GPU). Intel Core i3 processor or above.

3.2 Software requirements

Windows 7 or above / Linux. Python 2.7 or above.Jupyter Notebook.

4 ACTUAL WORKING OF OUTLIER PROCESS

Any data analytics project follows these steps in order to do the task of Predictions: Data Collection Gather the Sense of Data Clean the Data Improve your Data to make accurate Predictions Explore the Data using visualizations Using Intelligent algorithms to start predicting your target In context to our project Data Collection: In this process we were able to perform the task of data collections by using the data set from open source which consists of data values of 284,407 transaction in United Kingdom in two days of September 2013. Gathering the Sense of Data**:** PCA has been applied to the data components to hide the identity of user we were also able to observe that the transaction amount has been recorded and class has been added to assist in supervised learning and result analysis, we were also able to observe that the data set was highly imbalanced with only 0.172% of fraudulent transactions. Columns in the data set: Index(['Time', 'V1', 'V2', 'V3', 'V4', 'V5', 'V6', 'V7', 'V8', 'V9', 'V10', 'V11', 'V12', 'V13', 'V14', 'V15', 'V16', 'V17', 'V18', 'V19', 'V20', 'V21', 'V22', 'V23', 'V24', 'V25', 'V26', 'V27', 'V28', 'Amount', 'Class'], dtype='object') Clean the Data: As there is no missing data there in no need of this step.Improve your data to make accurate predictions: As our data is highly imbalanced we used sampling to highlight the features in the data set. We performed under sampling with ratio of 1.

Exploring the data using visualizations:

We used Histogram plots, correlation matrix and heat map to better understand the data.

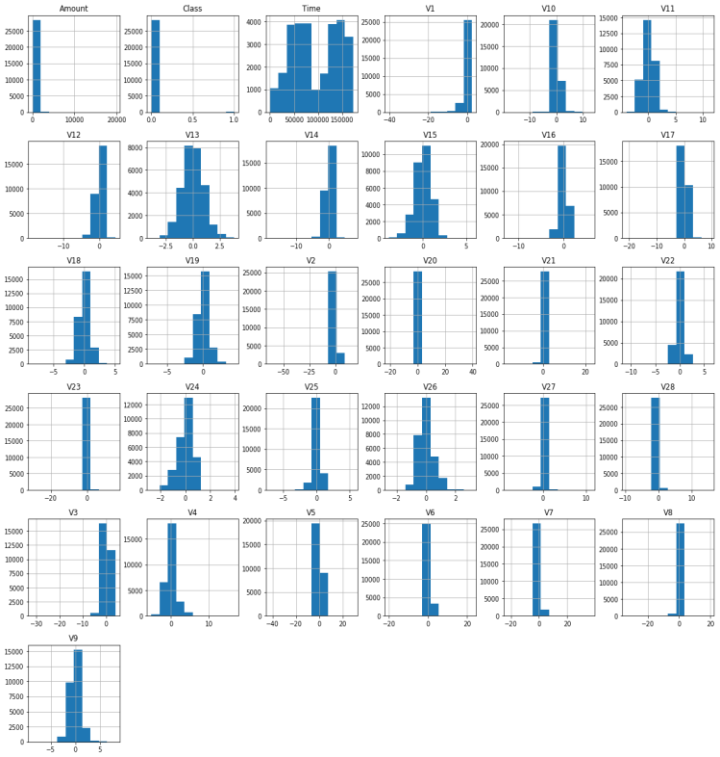

Figure:1.1
Depicts the histogram of each parameter

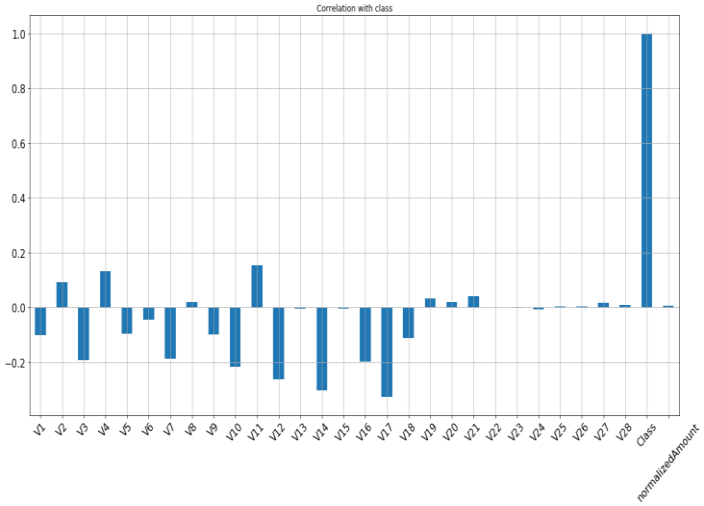

Figure: 1.2
Depicts correlation with class

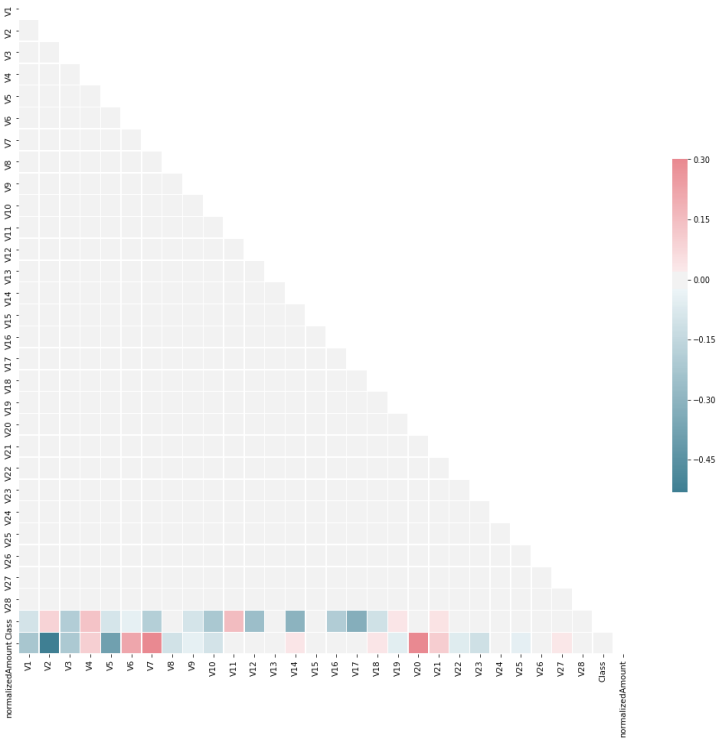

Figure: 1.3
Depicts Heatmap to understand correlation between different parameters

USING INTELLIGENT ALGORITHMS TO START PREDICTING YOUR TARGET:

In this project we decided to use three different algorithms to aid the task of outlier detection: Decision Tree: It is classification technique which utilizes tree like data structure to classify the target as it involves classification it can be categorized into supervised learning. It works on both categorical data and continuous input, it is also able to deal with non linear data. It works similarly to any classification where the model is already trained to tell where the new input data should be sent based on past data, it also utilizes the concept of hierarchy present in any tree structure to make the problem look more easer and act efficiently. There are two types of decision trees continuous ones and categorical ones in this technique we will be using categorical decision trees. In this technique by utilizing decision tree we will be able to predict any fraudulent data present in the credit card transactions data it utilizes training mechanism to initially train the model and next we perform the task of predicting the outliers present in the given dataset. Output:

```
CPU times: user 3.53 s, sys: 0 ns, total: 3.53 s
Wall time: 3.53 s
```

Figure:2.1.1

Depicts that we have performed training of our model
It can be observed that it took almost 4 seconds to train the model.

```
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=100,
                       max_features=None, max_leaf_nodes=100,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, presort=False,
                       random_state=0, splitter='best')
```

Figure:2.1.2

Depicts the characteristics of model after performing the training It can be observed that we limited the maximum leaf nodes to 100.

```
Decision Tree Prediction on Test Set:
-------------------------------------
Accuracy:       0.9993
Precision:      0.8409
Recall:         0.7551
F1 Score:       0.7957
```

Figure:2.1.3

Depicts several statistical performance indicators for Decision trees It can be observed that the model is accurate enough to predict fraudulent transaction but precision and recall are the areas of concern, which can be explained by looking at the graph below.

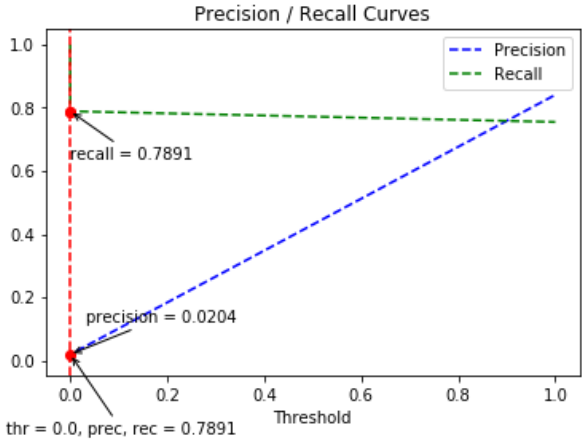

Figure:2.1.4
Depicts a comparative curve of precision vs The Recall

While Precision can be defined as the ratio true positives from predicted ones, Recall is the actual performance of model it is verified data with respect to true data and ration of 0.7891 is not at all appealing this can be said in simpler terms that out of every hundred fraudulent cases detected there are about 21 which are not actually fraud, This can be explained by observing the confusion matrix depicted below.

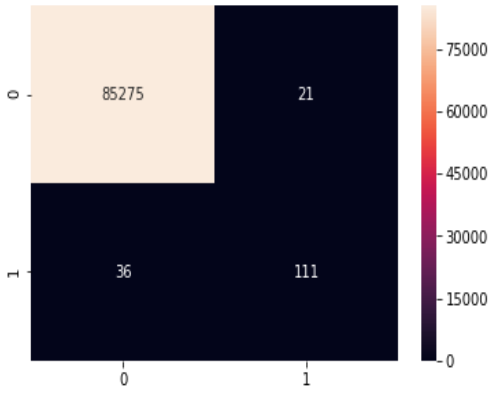

Figure:2.1.5
Depicts the overall performance by using random forest.

The top left depicts true negatives i.e. not fraud depicted as not fraud. The top right depicts False Positives i.e. not fraud as fraud.(discussed earlier) The bottom left depicts False negatives i.e. fraud as not fraud. The bottom right depicts True positives i.e. fraud as fraud Overall we were able to access that system got fooled few times also over estimated few times. To understand clearly we shall observe a line plot between true positives and false positives.

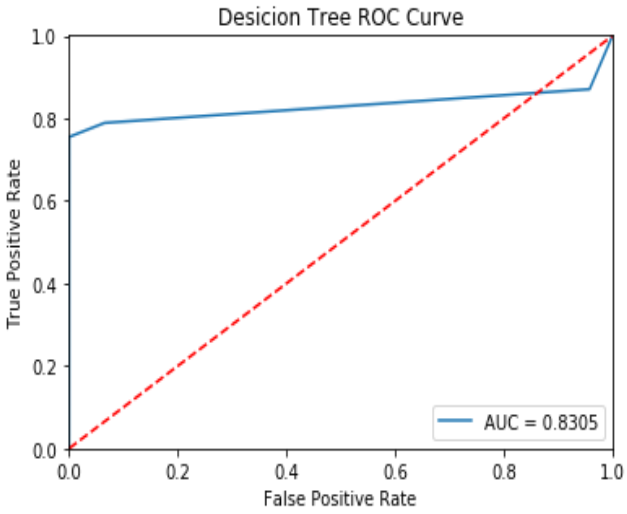

Figure:2.1.6
Depicts the comparative line graph between true and false positives

Random Forest Algorithm:
It works similar to that of a decision tree except we have several random decision trees working together to predict the outlier present in our data. In this technique we shall be using 300 decision trees together.

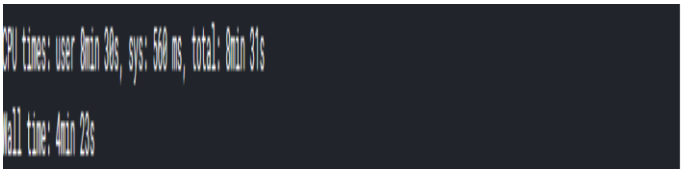

Figure:2.2.1
Depicts the total cpu usage time to train the model

It can be observed that CPU ran for about 8 and half minutes to train random forest of decision trees.

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                       max_depth=100, max_features='auto', max_leaf_nodes=100,
                       min_impurity_decrease=0.0, min_impurity_split=None,
                       min_samples_leaf=1, min_samples_split=2,
                       min_weight_fraction_leaf=0.0, n_estimators=300,
                       n_jobs=-1, oob_score=False, random_state=0, verbose=0,
                       warm_start=False)
```

Figure:2.2.2
Depicts the status of training the random forest algorithm

```
Random Forest Prediction on Test Set:
-------------------------------------
Accuracy:    0.9995
Precision:   0.9417
Recall:      0.7687
F1 Score:    0.8464
```

Figure:2.2.3
Depicts several statistical performance indicators for Random forest

It can be observed that the model is accurate enough to predict fraudulent transaction but precision and recall are the areas of concern, which can be explained by looking at the graph below.

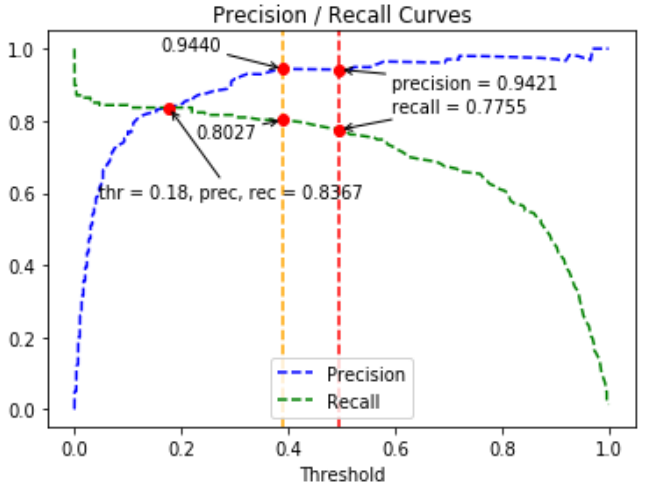

Figure:2.2.4
Depicts precision/ Recall curve for Random forest algorithm

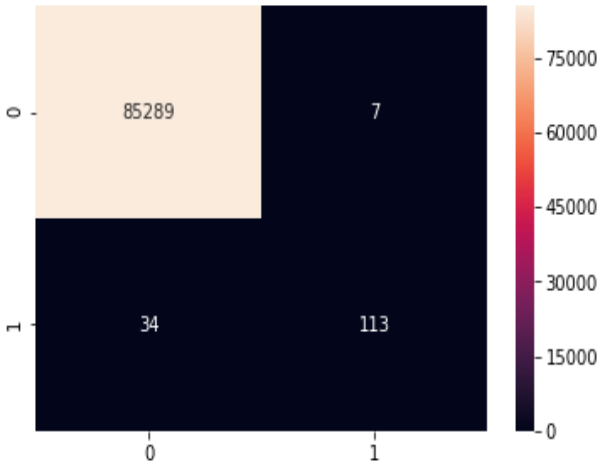

Figure:2.2
Depicts the confusion matrix for Random forest algorithm

The top left depicts true negatives i.e. not fraud depicted as not fraud. The top right depicts False Positives i.e. not fraud as fraud.(discussed earlier) The bottom left depicts False negatives i.e. fraud as not fraud. The bottom right depicts True positives i.e. fraud as fraud. Overall we were able to access that system got fooled few times also over estimated few times.

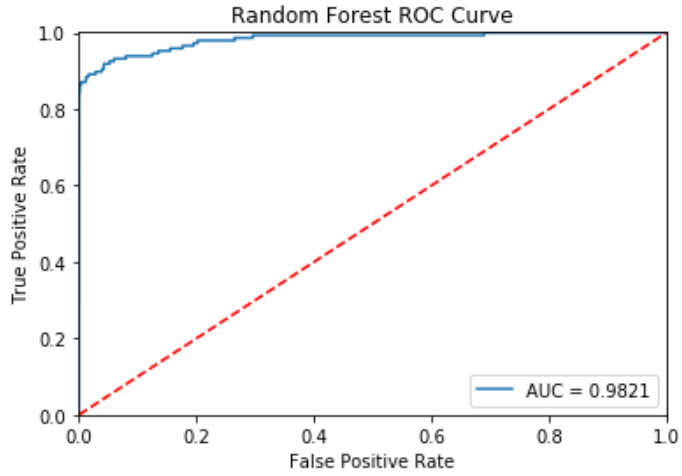

Figure:2.2.6
Depicts ratio for True positives vs False positives

Neural Network Algorithm:
It works in the similar format of a neural net instead we have computer simuated neural networks working to solve the given problem, In this case outlier detection.

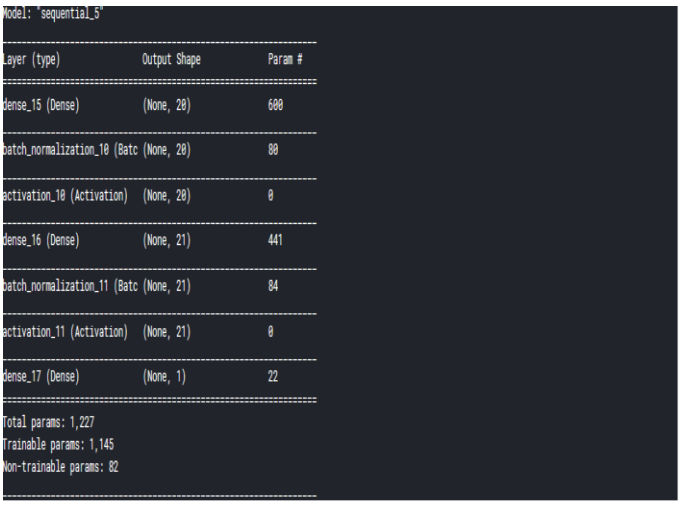

Figure:2.3.1
Depicts training status of neural network technique

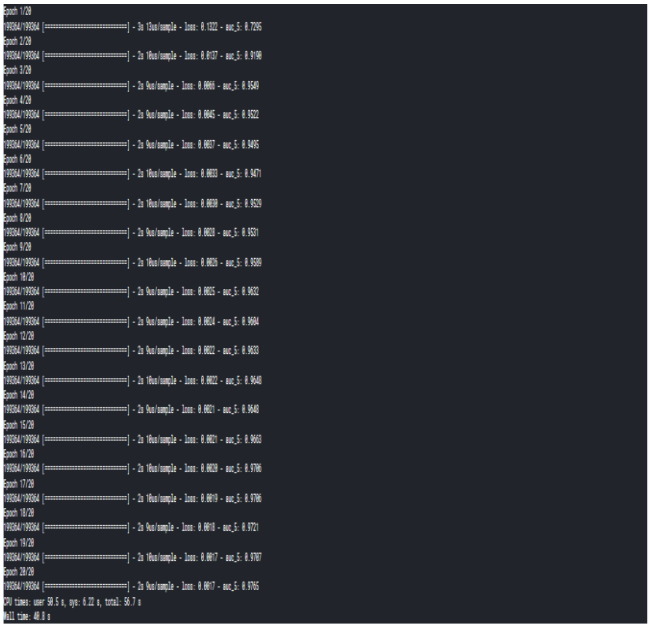

Figure:2.3.2
Depicts the model after training

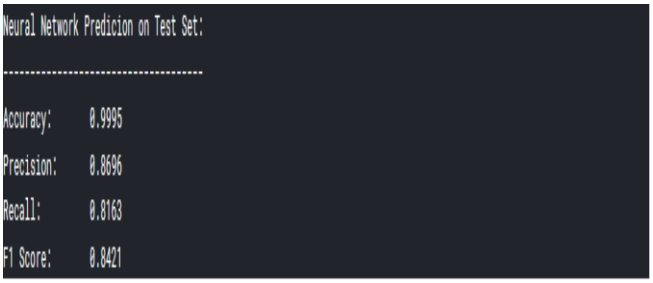

Figure:2.3.3
Depicts several statistical performance indicators for neural network

This shows that the neural network technique is quite accurate and precise but has recall, further explanation is represented in the graph below.

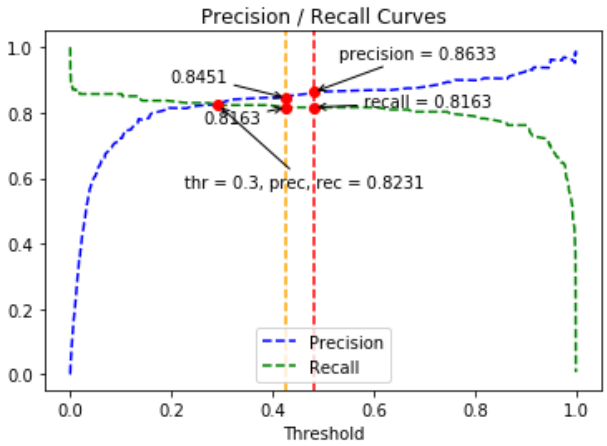

Figure:2.3.4
This graph depicts the ratio of precision vs recall.

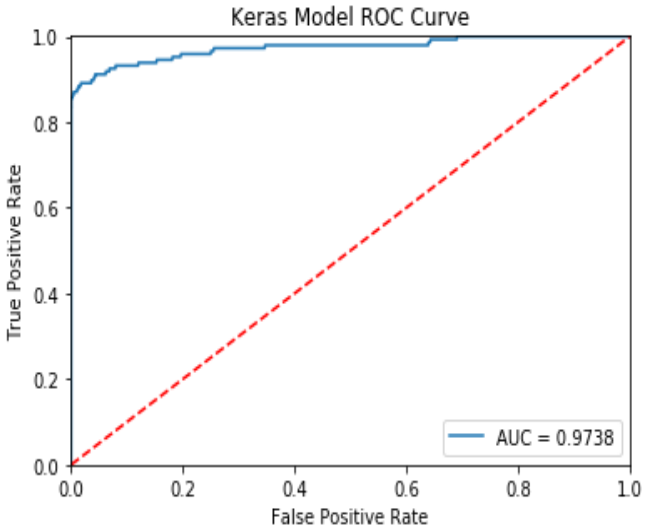

Figure:2.3.5
This graph depicts the ratio between true positives vs false positives

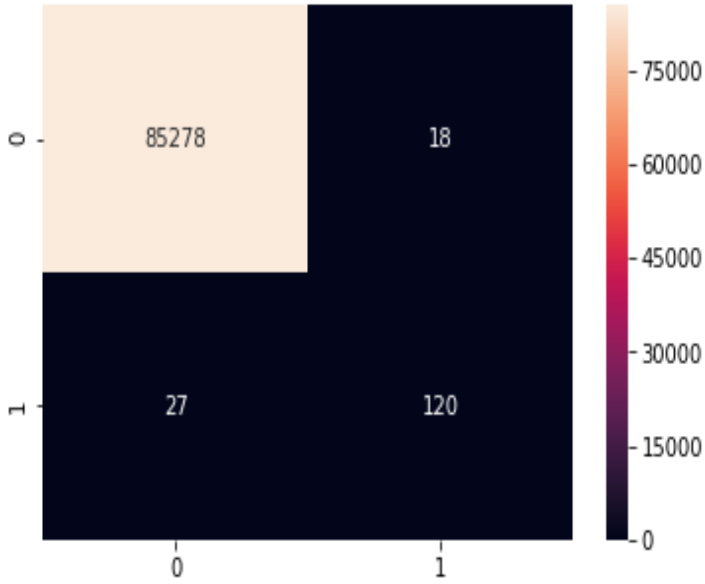

Figure:2.3.5

The top left depicts true negatives i.e. not fraud depicted as not fraud. The top right depicts False Positives i.e. not fraud as fraud.(discussed earlier) The bottom left depicts False negatives i.e. fraud as not fraud. The bottom right depicts True positives i.e. fraud as fraud. Overall we were able to access that system got fooled few times also over estimated few times

5 RESULTS AND INFERENCES

The dataset is pre-processed and the model is trained and tested. To calculate the results, and accuracy, we implement four accuracy scores. 1. Precision: The precision is the ratio T / (TP+FP) where TP is the number of true positives and FP the number of false positives. The precision is intuitively the ability of the classifier not to label as positive a sample that is negative. 2. Recall: The recall is the ratio TP/ (TP+FN) where TP is the number of true positives and FN the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples. 3. F1-Score: The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches its best value at 1 and worst score at 0. 4. Support: The support is the number of occurrences of each class in out puts that are true.

Result Analysis:

	Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
0	Decision Tree	0.999427	0.871212	0.782313	0.824373	0.911394
1	Random Forest	0.999520	0.941667	0.768707	0.846442	0.976211
2	Decision Tree	0.999333	0.840909	0.755102	0.795699	NaN
3	Decision Tree	0.999333	0.840909	0.755102	0.795699	0.830488
4	Random Forest	0.999520	0.941667	0.768707	0.846442	0.982085
5	Decision Tree	0.999333	0.840909	0.755102	0.795699	NaN
6	Keras NN	0.999473	0.869565	0.816327	0.842105	0.973754

Figure:3.1.1
Comparison of various statistical indicators for various techniques

We are able to observe that random forest is most precise and accurate technique among all the techniques but it almost takes a lot of time to train the model in random forest algorithm, neural network is next best technique and is very time efficient, decision tree is least accurate and precise and takes less time than random forest but more time than

implementing through neural network.

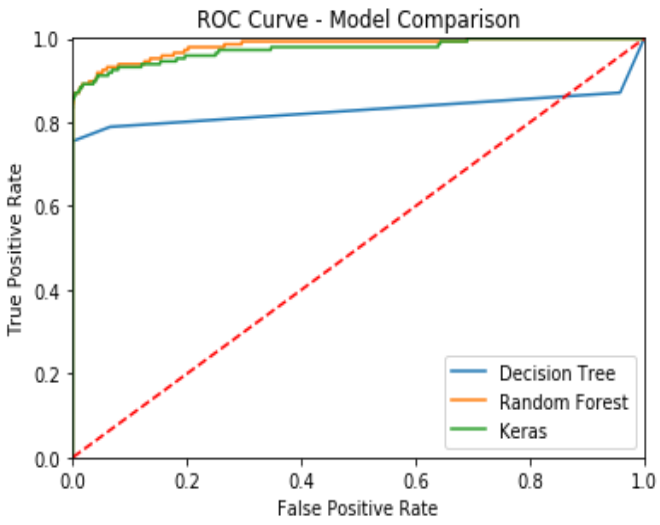

Figure:3.1.2
ROC Curve for various intelligent techniques

We were able to observe that neural network has least false positives while decision tree has most false positive, But neural network is least efficient in determining fraudulent transactions.

6 CONCLUSION

We tried to solve the problem of identifying fraudulent transaction using outlier detection which can be implemented using several machine learning algorithms. In this project we implemented outlier detection using Decision trees, Random Forest, Neural Network techniques. By observing accuracy reports and several statistical visualizations we were able to conclude that decision trees is least accurate while Random Forest has highest accuracy but is least time efficient, In terms of time efficiency and computational resource utilization the neural network is best algorithm.

7 REFERENCES

[1] Ginny Y. Wong, and Frank H. F. Leung, "Predicting Protein-Ligand Binding Site Using Support Vector Machine with Protein Properties," Nov.-Dec. 2013, pp. 1517-1529, vol. 10.

[2] Chao Wang, Tongge Xu , Xi Qin, "Network Traffic Classification with Improved Random Forest" Year: 2015, Volume: 1, Pages: 78-81, DOI Bookmark:10.1109/CIS.2015.27

[3] Ana Erika Camargo Cruz and Koichiro Ochimizu, "Towards logistic regression models for predicting fault-prone code across software projects", Year: 2009, Volume: 1, Pages: 460-463,DOI Bookmark:10.1109/ESEM.2009.5316002

[4] Marella, S.T., Karthikeya, K., Kushwanth, V.S. and Bezawada, A., 2018. Enhancement of Performance and Economy of Data Centers by Virtualization. *International Journal of Simulation--Systems, Science & Technology*, *19*(6).

[5] Bezawada, A., Marella, S.T. and Gunasekhar, T., 2018. A Systematic Analysis of Load Balancing in Cloud Computing. *International Journal of Simulation--Systems, Science & Technology*, *19*(6).

[6] Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R. and Kundu, S., 2018. Improved random forest for classification. *IEEE Transactions on Image Processing*, *27*(8), pp.4012-4024.

[7] Zhang, X., Wang, S. and Zhao, Y., 2011, June. Application of support vector machine and least squares vector machine to freight volume forecast. In *2011 International Conference on Remote Sensing, Environment and Transportation Engineering* (pp. 104-107). IEEE.

[8] Li, P., Li, S., Bi, T. and Liu, Y., 2014. Telecom customer churn prediction method based on cluster stratified sampling logistic regression.

[9] Vatrapu, R., Mukkamala, R.R., Hussain, A. and Flesch, B., 2016. Social set analysis: A set theoretical approach to big data analytics. Ieee Access, 4, pp.2542-2571

[10] Rind, A., Lammarsch, T., Aigner, W., Alsallakh, B. and Miksch, S., 2013. Timebench: A data model and software library for visual analytics of time-oriented data. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), pp.2247-2256.

[11] Jabbar, S., Malik, K.R., Ahmad, M., Aldabbas, O., Asif, M., Khalid, S., Han, K. and Ahmed, S.H., 2018. A methodology of real-time data fusion for localized big data analytics. *IEEE Access*, *6*, pp.24510-24520.

[12] Huang, Z., Li, M., Chousidis, C., Mousavi, A. and Jiang, C., 2017. Schema Theory-Based Data Engineering in Gene Expression Programming for Big Data Analytics. *IEEE Transactions on Evolutionary Computation*, *22*(5), pp.792-804.

[13] Lepenioti, K., Bousdekis, A., Apostolou, D. and Mentzas, G., 2020. Prescriptive analytics: Literature review and research challenges. *International Journal of Information Management*, *50*, pp.57-70.

[14] Liang, T.P. and Liu, Y.H., 2018. Research landscape of business intelligence and big data analytics: A bibliometrics study. *Expert Systems with Applications*, *111*, pp.2-10.

[15] Seng, J.K.P. and Ang, K.L.M., 2017. Big feature data analytics: Split and combine linear discriminant analysis (SC-LDA) for integration towards decision making analytics. *IEEE Access*, *5*, pp.14056-14065.

[16] Herbrich, Ralf, Tom Minka, and Thore Graepel. "Trueskill™: A Bayesian skill rating system." Advances in Neural Information Processing Systems. 2006.

[17] Solo, A.M., 2017, December. An Overview of the New Interdisciplinary Fields of Political Engineering and Computational Politics for the Next Frontier in Politics. In *2017 International Conference on Computational Science and Computational Intelligence (CSCI)* (pp. 1805-1806). IEEE.

[18] Qin-zhang, Y., Jing, Z. and Ying-chao, Z., 2010, January. Modeling and Simulation of International Politics Evolution in Crisis. In *2010 Second International Conference on Computer Modeling and Simulation* (Vol. 1, pp. 324-328). IEEE.

[19] Marsland, S., 2014. *Machine learning: an algorithmic perspective*. Chapman and Hall/CRC.

[20] Polson, N.G. and Sokolov, V., 2017. Deep learning: a Bayesian perspective. *Bayesian Analysis*, *12*(4), pp.1275-1304.

[21] Haykin, S.S., 2009. *Neural networks and learning machines/Simon Haykin*. New York: Prentice Hall,.

[22] LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. *nature*, *521*(7553), p.436.
[23] Bengio, Y., 2009. Learning deep architectures for AI. *Foundations and trends® in Machine Learning*, *2*(1), pp.1-127.
[24] Mahadevan, S. and Haldar, A., 2000. Probability, reliability and statistical method in engineering design. *John Wiley & Sons*.
[25] Oberkampf, W.L., Trucano, T.G. and Hirsch, C., 2004. Verification, validation, and predictive capability in computational engineering and physics. *Applied Mechanics Reviews*, *57*(5), pp.345-384.