COMPARISON OF CREDIT CARD FRAUD DETECTION TECHNIQUES

[1]Saubhagya Vardhan, [2]Dushyant Singh, [3]Dr. (Mrs.) Neha Agarwal
[1]Student, [2]Student, [3]Assistant Professor
[1]IT,
[1]Maharaja Agrasen Institute of Technology, New Delhi, India

Abstract : With the advent of cashless economy and rising participation in online based transactional activities the number of fraudulent cases all over the world is on the rise and causes tremendous losses to the individuals and financial industry. Although, there are many criminal activities occurring in the financial industry, credit card frauds are among the most prevalent and worried by customers. Hence, to counter these frauds data mining along with machine learning is one of the prominent approaches used to prevent the losses caused by these illegal acts. First, data mining techniques were employed to study the patterns and characteristics of suspicious and non-suspicious transactions based on normalized and anomalies data. In addition to this, machine learning (ML) techniques were employed to predict the suspicious and non-suspicious transactions automatically by using classifiers. Therefore, with the combination of machine learning and data mining techniques we were able to identify the genuine and non-genuine transactions by learning the patterns of the data. This paper discusses both the supervised and unsupervised based classification techniques used for credit card fraud detection.

Keywords— Credit Card, Fraud Detection**,** Machine Learning, Data Mining

1.INTRODUCTION

Credit card fraud is defined as the unauthorized usage of card, unusual transaction behaviour, or transactions on an inactive card [1]. In general, there are three categories of credit card fraud namely, conventional frauds (e.g. stolen, fake and counterfeit), online frauds (e.g. false/fake merchant sites), and merchant related frauds (e.g. merchant collusion and triangulation) [2]. To prevent such frauds data mining along with machine learning(ML) is used.

Data Mining is known as the process of gaining interesting, novel and insightful patterns as well as discovering understandable, descriptive and predictive models from large scale of data collections [3, 4]. The ability of data mining techniques to extract fruitful information from large scale of data using statistical and mathematical techniques helps in credit card fraud detection by differentiating the characteristics of common and suspicious credit card transactions. While data mining is focused on discovering the valuable intelligence, machine learning is rooted in learning the intelligence and developing its own model for the purpose of classification, clustering or so on.

Machine Learning(ML) is a technique in computer science according to which a machine imitates human intelligence.
Machine Learning classifiers operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions. There are many
different types of machine learning approaches available with the intentions to solve heterogeneous problems. Due to the nature of this study which was focused on classification, the discussion that follows is based on this topic. Machine learning classification refers to the process of learning to
assign instances to predefined classes. Formally, there are several types of learning such as supervised, semi-supervised, unsupervised, reinforcement, transduction and learning to
learn [5].

Anomaly Detection is defined as the technique used to identify unusual patterns in a dataset.
The machine learning based anomaly detection techniques are density based, clustering based and support vector machine based.

2. ALGORITHMS USED

Local Outlier Factor(LOF) is an anomaly detection algorithm. The local outlier factor is based on the concept of local density, where locality is defined by nearest neighbours, whose distance is used to estimate the density.
By comparing the local density of an object to the local densities of its neighbours, we
can identify regions of similar density, and points that have a substantially lower density than their neighbours. These are considered to be outliers.
The local density is estimated by the typical distance at which a point can be "reached" from its neighbours. The definition of "reachability distance"
used in LOF is an additional measure to produce more stable results within clusters.

Isolation Forest is another anomaly detection algorithm. Isolation Forest explicitly identifies anomalies instead of profiling normal data points. Isolation Forest, like any tree ensemble method, is built on the basis of decision trees. In these trees, partitions are created by first randomly selecting a feature and then selecting a random split value between the minimum and maximum value of the selected feature.

Random Forest Classifier is an ensemble algorithm.
Ensembled algorithms are those which combines more than one algorithms of same or different kind for classifying objects. Random forest classifier creates a set of decision trees from randomly selected subset of training set. It then aggregates the votes from different decision trees to decide the final class of the test object.

Support Vector Machine(SVM) is a supervised learning algorithm. The algorithm learns a soft boundary in order to cluster the normal data using training data and then using the testing data it learns to identify the anomalies outside this soft boundary.

3.PERFORMANCE METRICS USED

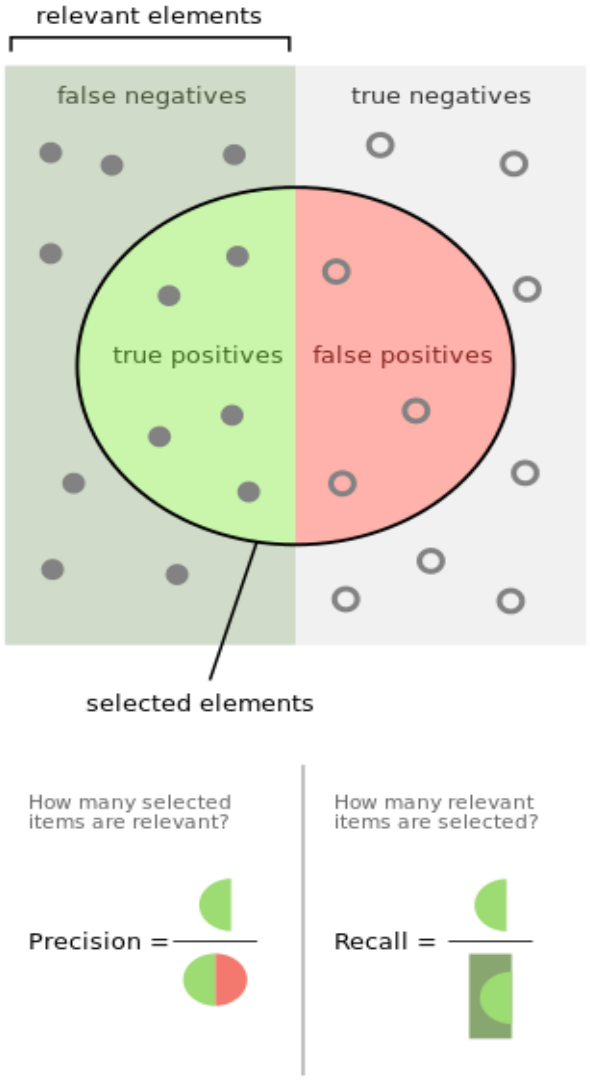
True Positives (TP) – No of valid cases correctly identified as valid cases.
True Negatives (TN) – these are the no of frauds correctly predicted as frauds.
False Positives (FP) – These are the no fraud/ invalid cases being identified as valid cases/true values
False Negatives (FN) – False negatives represent the no of valid cases being identified as frauds

Accuracy – It is performance measurement metrics. It represents the no of correctly identified values over the complete dataset. Accuracy is one of the best performance measurement metric provided you have a symmetric dataset i.e. the no of valid cases are more or less equal to the invalid cases. For an asymmetric dataset however you need different metrics.

Precision – Precision represents the no of correctly predicted posited cases over total no of predicted positive values. Precision = TP/TP+FP

Recall - Also called as sensitivity. Recall is the no of correctly identified positive cases over the total no of valid cases .Recall = TP/TP+FN

F1 score – weighted average taken over precision and recall is F1 score. It incorporates both false positives and false negative values and is a better performance metric than accuracy in cases of uneven /asymmetric datasets. F1 Score = 2*(Recall * Precision) / (Recall + Precision)

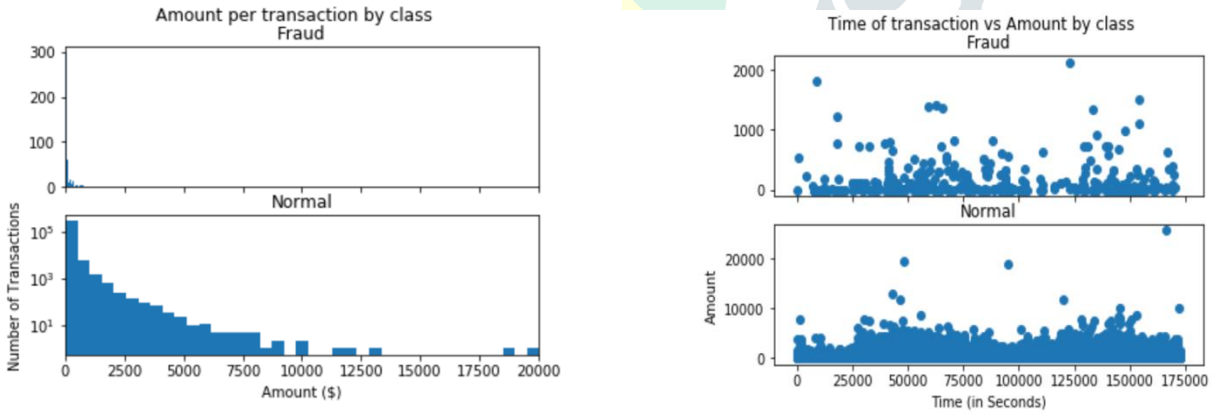


Performance metrics

4.OBSERVATIONS

1. The dataset consists of values from 28 PCA transformed variables.
2. The dataset is highly skewed with only 492 fraud transactions among 2,84,807 transactions in total.
3. The 'time' and 'amount' variables are not transformed data.
4. There are no missing values in the dataset.
5. There is no co-relation between the amount of transaction and fraud transactions.
6. There is no co-relation between the time of transaction and fraud transactions.
7. Local Outlier Factor detected 97 errors.
8. Isolation Forest detected 73 errors.
9. Support Vector Machine detected 8516 errors.
10. Isolation Forest has an accuracy of 99.74%.
11. Local Outlier Factor has an accuracy of 99.65%.
12. Support Vector Machine has an accuracy of 70.09%.
13. Random Forest has an accuracy of 99.94%.
14. The error precision and recall for Random Forest is 74% much higher than Isolation Forest's 27%, Local Outlier Factor's 2% and Support Vector Machine's 0%.

5.OUTPUTS



Dataset parameter analysis

```
Isolation Forest: 73
Accuracy Score :
0.9974368877497279
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.26      0.27      0.26        49

   micro avg       1.00      1.00      1.00     28481
   macro avg       0.63      0.63      0.63     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor: 97
Accuracy Score :
0.9965942207085425
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

   micro avg       1.00      1.00      1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

Output of random forest classifier

```
Support Vector Machine: 8516
Accuracy Score :
0.7009936448860644
Classification Report :
              precision    recall  f1-score   support

           0       1.00      0.70      0.82     28432
           1       0.00      0.37      0.00        49

   micro avg       0.70      0.70      0.70     28481
   macro avg       0.50      0.53      0.41     28481
weighted avg       1.00      0.70      0.82     28481
```

SVM output

```
Random Forest Classifier
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      8533
           1       1.00      0.58      0.74        12

avg / total       1.00      1.00      1.00      8545

0.9994148624926857
[[8533    0]
 [   5    7]]
```

Results of isolation forest and local outlier factor

6.CONCLUSION

After comparing the results of the four algorithms and carefully observing them we can say that Random Forest classifier has the highest accuracy and error precision and recall time so its best suited for credit card fraud detection whereas Support Vector Machine with lowest accuracy and least error precision and recall time is not suited for credit card fraud detection where datasets are highly skewed and is better suited to balanced datasets.

7.REFERENCES

1.https://en.wikipedia.org/wiki/Local_outlier_factor
2.https://medium.com/machine-learning-101/chapter-5-random-forest-classifier-56dc7425c3e1
3.https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e