IT-563 DATA MINING PROJECT

Kaushal Patel (201501219)*

Dhirubhai Ambani Institute of Information & Communication Technology,Gandhinagar, Gujarat 382007, India

Credit Card Fraud Detection

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 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.

I. PROBLEM DEFINITION

The credit card has become the most popular mode of payment for both online as well as regular purchase, in cases of fraud associated with it are also rising. Credit card frauds are increasing day by day regardless of various techniques developed for its detection. Fraudsters are so experts that they generate new ways of committing fraudulent transactions each day which demands constant innovation for its detection techniques. Most of the techniques based on Artificial Intelligence, Fuzzy Logic, Neural Network, Logistic Regression, Nave Bayesian, Machine Learning, Sequence Alignment, Decision tree, Bayesian network, meta learning, Genetic programming etc., these are evolved in detecting various credit card fraudulent transactions. This paper presents a survey of various techniques used in various credit card fraud detection mechanisms.

II. MOTIVATION

The credit card has become the most popular mode of payment for both online as well as regular purchase, in cases of fraud associated with it are also rising. Credit card frauds are increasing day by day regardless of various techniques developed for its detection. Fraudsters are so experts that they generate new ways of committing fraudulent transactions each day which demands constant innovation for its detection techniques. The motivation is to present, compare and analyze recently published detection techniques and findings in credit card fraud detection.

III. REPORT WORK OF RESEARCH PAPER

A. Classification Techniques

Classification used in research paper [1] are:

1. **Naive Bayes:**The process of classifying an instance is done by applying the Bayes rule for each class given the instance. In the fraud detection task, the probability is calculated for each of the two classes (fraudulent and legitimate) and the class associated with the higher probability is the predicted class for the instance.

2. **Decision tree :**A decision tree consists of nodes that forms a tree. Each non-leaf node denotes a test on an attribute, each branch represents the outcome of a test, and each leaf node holds a class label. Given an instance with its features values, the model is able to classify the instance by traversing the decision tree.

3. **K-nearest Neighbors:** The k-Nearest Neighbors (KNN) algorithm is a simple instance-based algorithm that plots all training instances and classify unlabelled instances based on their closest neighbours.

4. **Support vector machines :** SVM use a nonlinear mapping to transform the input data into a multidimensional feature space. After this transformation the SVM finds the best hyper plane inside the feature space. The nonlinear mapping depends on what so called a kernel function.

B. Data set

The dataset used in [1] research paper contains a real-life data of financial transactions of an e- commerce organization. The data set is composed of 100,000 instances and each instance is composed of 20 features. This data set is highly imbalanced with a ratio of approximately 97 : 3 towards legitimate transactions, meaning that 3

*Electronic address: 201501219@daiict.ac.in percent of the transactions are fraud while the other 97 percent are legit.

C. Experiments and results

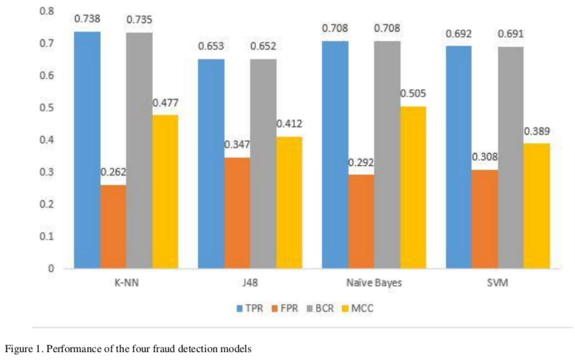

FIG. 1: Performance Metrics vs Various classification methods

The above figures shows the four different performance metrics namely True positive rate, False positive rate , Balanced classification rate, Mathews correlation coefficient for four different classifiers.

IV. EMPIRICAL STUDY OF CLASSIFIERS

A. Data Set and Feature Engineering

The data set used in this experiment contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172 percent of all transactions.

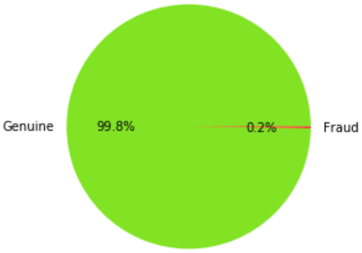

FIG. 2: Data Imbalance

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'.This can be observed from the figure 3 that Time and amount are correlates differently with different features Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

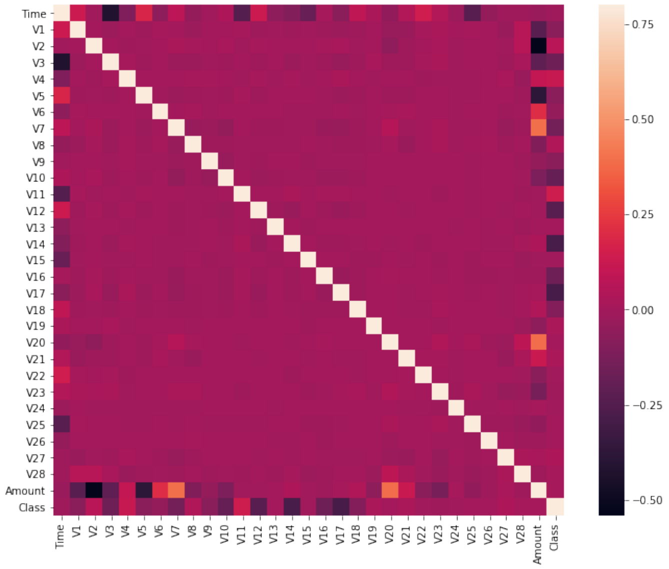

FIG. 3: Correlation of every feature with other feature

B. Classification Techniques

For the experiment five different classification techniques are used:

1. **Random Forest Classifier(RFC)**: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees

2. **Decision Tree Classifier(DTC):**Decision tree is one of the most widely used and practical methods for predictive learning. Its model is very transparent and better explains the class prediction

3. **Naive Bayes Classifier:**In machine learning, naive Bayes classifiers are a family of simple "probabilistic classifiers "based on applying Bayes' theo-

rem with strong (naive) independence assumptions between the features.

4. **Isolation Forest:** Isolation Forest detects data-anomalies using binary trees.

5. **Local Outlier Factor:**In anomaly detection, the local outlier factor (LOF) is an algorithm for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbours

After building the model, we apply it to the respective test data set and we compute the Confusion Matrix, containing starting from up-left cell in clockwise the values: True Positive TP, false Positive FP,True Negative TN and and False Negative FN, and the Performance Parameters like:

1. Accuracy: (TP+TN)/(TP+FP+TN+FN)
2. Sensitivity: TP/(TP+FN)
3. Specificity: TN / (FN+TN)
4. Positive Prediction value: TP/(TP+FP)
5. Negative Prediction Value: TN/(FN+TN).

C. Experiments and results

The five fraud detection models were trained and tested using SciKit library of python. A 10-fold cross validation was used in the process of training and testing the different models. In 10-fold cross validation the data set is divided into 10 subsets; one of them is used as the testing set and the others are used as the training set. This process is repeated taking a different subset as the testing set. The average performance results are then recorded. This methodological approach ensures that all data were represented once as a test data and several times as a training data producing accurate results.

Regarding to the Random forest classifier based model no of the estimators were equal to 98 while the performance to measure feature split used for information index used was the gini index. For Random forest classifier various combination of normalized Time and amount feature were considered.
For Decision Tree C4.5 decision tree algorithm was adopted in this paper to develop the decision tree based model and Gininindex was used to evaluate the splitting of the data based on features. Based on the depth size of the tree various results were observed.
We can observe that the Random Forest classifier and Decision tree classifier outperformed other models in terms of the accuracy,Recall,f1-sore and other performance metrics. C4.5 model (J48 algorithm is an implementation of the C4.5 in python) in terms of some performance metrics outperformed random forest classifier. While RFC outperformed DTC in terms of accuracy and recall. Maybe using other parameters for RFC, DTC and other models various or different settings to the parameters various obsevation can be seen leading to better performance.
Python Libraries used are:

1. Numpy : 1.14.2
2. Pandas : 0.22.0
3. Seaborn : 0.8.1
4. Scipy : 1.0.1
5. sklearn : 0.19.1

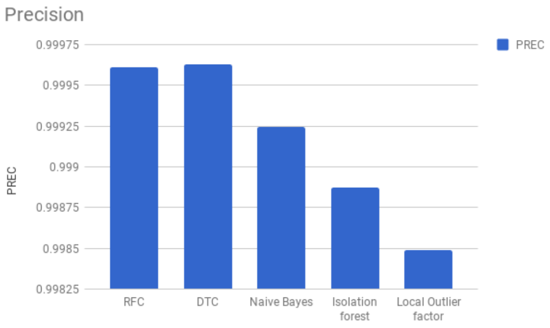

FIG. 4: Precision

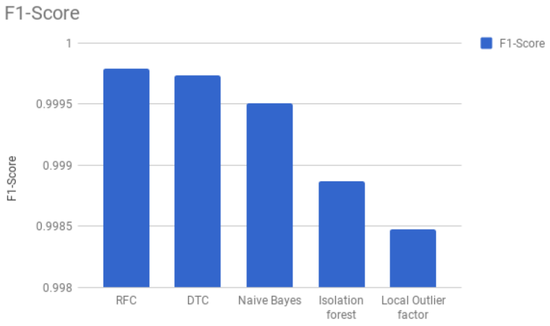

FIG. 5: F1-Score

Here we can observe that when comparing various performance metrics for this five classifiers, the precision, accuracy, F1-score and recall are almost show the same behaviour for the five mentioned classifiers. We can clearly see the behaviour that Random forest classifier and Decision tree classifier gives good accuracy and other performance metrics. While Local outlier factor method shows relatively low performance. This happens because the
$outlierfactor = 0.0015296972254457222$
$Fraud : 87$
$Valid : 56874$
in the data set and as algorithms such as the Isolation

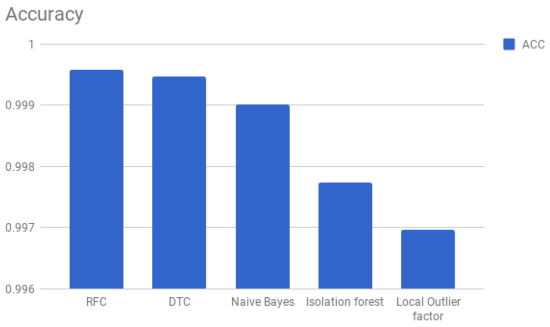

FIG. 6: Accuracy

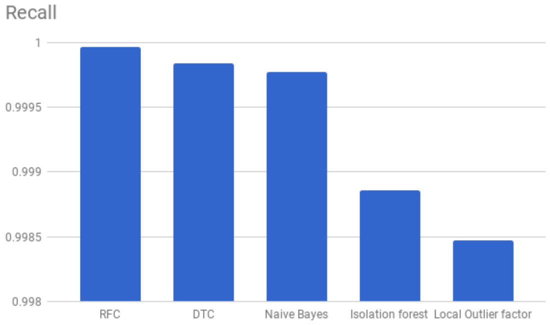

FIG. 7: Recall

tree and local outlier factor method mainly works on the outlier while the Random Forest classifer and Decision tree works based on the no. of transactions that are both legitimate as well as fraud. So this kind of unexpected anomaly for various methods are ovserved

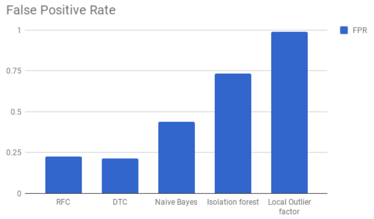

FIG. 8: False positive Rate for various classifiers

False positive in this context is the fraud transactions which are detected legitimate. Thus a classifier must be designed such that the false positive rate is minimum. From the below given figure one can observe that the Random Forest classifier and Decision tree are very effective methods for this data set.

V. CONCLUSIONS

Random Forest Classifier , Naive Bayes , Decision Trees, Isolation Forest and Local Outlier Factor were used in developing five fraud detection models to classify a transaction as fraudulent or legitimate. Four metrics were used in evaluating their performances. The results showed that there is no data mining technique that is universally better than others. Performance improvement could be achieved through developing a fraud detection model using a combination of different data mining techniques (ensemble).

[1] Data Mining Techniques for Credit Card Fraud Detection Empirical Study.pdf

[2] Credit card fraud and detection techniques: a review by Linda Delamaire (UK), Hussein Abdou (UK), John Pointon (UK)

[3] Survey on Credit Card Fraud Detection Techniques:http://www.ijecs.in/issue/v4-i11/25%20ijecs.pdf

[4] Credit Card Fraud Detection using Local Outlier Factor: by International Journal of Pure and Applied Mathematics

[5] https://www.kaggle.com/samkirkiles/credit-card-fraud/data

[6] https://classeval.wordpress.com/introduction/basic-evaluation-measures/