ANOMALY DETECTION AND LOCAL OUTLIER FACTOR FOR CREDIT CARD FRAUD DETECTION

[1] Achal Sai R, [2] Adithi A, [3] Amaresh B Patil, [4] Gokul HN, [5] Dr. Gururaj Murtugudde

[1,2,3,4]Student, [5] Professor,

[1,2,3,4,5]Department of Computer Science & Engineering,

[1,2,3,4,5]Nagarjuna College of Engineering and Technology, Bengaluru – 562164, Karnataka, India.

Abstract: Nowadays, as internet speed has increased and the prices of the mobile have decreased very much in past few years. Also the data prices too are very much affordable to most of the people. This has resulted into the digitization of most of the institutes as it is easy and convenient for the people and also for the authority to maintain the records. So it resulted in most of the banks and other institutes receiving and transferring money through credit card. But with the hackers and other cyber criminals around credit card system is very easy to perform fraud. These credit card fraud creates financial loss for customers and companies and everyday fraudsters find new technique to commit the fraud. The possibilities of the fraud transaction are very less but it is not negligible and even having one fraud transaction is unacceptable because it is crime and we can't neglect even if amount is less as it harms. So this project aims at analyzing various classification techniques using various metrics for judging various classifiers. This model aims at improving fraud detection rather than misclassifying a genuine transaction as fraud. After using the algorithms such as Isolation Forest and Local Outlier Factor, a detailed report is given in this paper.

Index Terms **– Fraud Detection, Imbalanced Datasets, Credit card, Local Outlier Factor, Isolation forest, Amazon web service.**

I. INTRODUCTION

A credit card is a card used for making payment to merchant for goods and service bought by card owners. Usually the bank creates the account and grants the credit to the card holder, by which the cardholder can borrow money and make a payment. The credit card fraud happens usually to obtain goods or service, or to make payment for another account which is usually handled by the criminals. The PCI DSS is data security standard has been created to many businesses card payment for making the payment using credit card securely, even though the crime has not decreased. Credit card Fraud is one the biggest threats in business world today. Credit card theft begins with theft or physical card or important information associated with card holder account, such as card account number and other information necessarily for merchant during transaction. According to report [1] as per the USFTC till the mid of 2000s the theft rate was stable but after the 2008 the rate of theft has gradually increased. In this paper, the dataset is analysed and operation are performed and these datasets are taken from Kaggle [2]. By analysing the datasets and their behaviour, credit card transaction is classified fraud or not.

This paper consists of selecting optimal algorithm for finding fraud pattern through effective comparison of machine learning techniques through an active performance measure for detection of fraudulent credit card transaction. The rest of this paper is presented as follows. Section II presents the related works. Section III presents the methodology. Finally, the conclusion and discussion in Section IV.

II. LITERATURE SURVEY

[1] Mr. Ibtissam Benchaji, Samira Douzi, Bouabid El Ouahidi "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for credit card fraud detection" published in 2018 states that with the growing usage of credit card transactions, financial fraud crimes have also been drastically increased leading to the loss of huge amounts in the finance industry. Having an efficient fraud detection method has become a necessity for all banks in order to minimize such losses. In fact, credit card fraud detection system involves a major challenge: the credit card fraud data sets are highly imbalanced since the number of fraudulent transactions is much smaller than the legitimate ones. Thus, many of traditional classifiers often fail to detect minority class objects for these skewed data sets. This paper aims first: to enhance classified performance of the minority of credit card fraud instances in the imbalanced data set, for that we propose a sampling method based on the K-means clustering and the genetic algorithm. We used K-means algorithm to cluster and group the minority kind of sample, and in each cluster we use the genetic algorithm to gain the new samples and construct an accurate fraud detection classifier. This has advantage of minimized misclassification.

[2] Mr. Hongyu Wang, Ping Zhu, Xueqiang Zou, Sujuan Qin "An Ensemble Learning Framework for Credit Card Fraud Detection based on Training Set Partitioning and Clustering" published on 2018 states that, the popularity of credit card has greatly facilitated the transactions between merchants and cardholders. However, credit card fraud has been derived, which results in losses of billions of euros every year. In recent years, machine learning and data mining technology have been widely used in fraud detection and achieved favorable performances. Most of these studies use the technology of under-sampling to deal with the high imbalance of credit card data. However, it will potentially discard some relevant training samples which will weaken the ability of the classifier. In this paper, we propose an ensemble learning framework based on training set partitioning and clustering. It turns out that the proposed framework not only ensures the integrity of the sample features, but also solves the high imbalance of the dataset. A main feature of our framework is that every base estimator can be trained in parallel. This improves the efficiency of the framework. We

show the effectiveness of our proposed ensemble framework by experimental results on a real credit card transaction dataset has the advantages of processing the categorical and numerical datasets and disadvantages of more outliers.

[3] Yvan Lucas, Pierre-Edouard Portier, Lea Laporte, Sylvie Calabretto, Liyun He-Guelton, Frederic Oble, Michael Granitzer "Dataset shift quantification for credit card fraud detection" published in 2019 states that machine learning and data mining techniques have been used extensively in order to detect credit card frauds. However, purchase behavior and fraudster strategies may change over time. This phenomenon is named dataset shift or concept drift in the domain of fraud detection. In this paper, we present a method to quantify day-by-day the dataset shift in our face-to-face credit card transactions dataset (card holder located in the shop). In practice, we classify the days against each other and measure the efficiency of the classification. The more efficient the classification, the more different the buying behavior between two days, and vice versa. Therefore, we obtain a distance matrix characterizing the dataset shift. After an agglomerative clustering of the distance matrix, we observe that the dataset shift pattern matches the calendar events for this time period (holidays, week-ends, etc.). We then incorporate this dataset shift knowledge in the credit card fraud detection task as a new feature. This leads to a small improvement of the detection and have advantage of having High fraud coverage combined with low fraud alarm rate and disadvantage of standalone solution.

[4] Ankit Mishra, Chaitanya Ghorpade presented paper on "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques" published in 2018 states that Nowadays, as internet speed has increased and the prices of mobile have decreased very much in past few years. Also the data prices too are very much affordable to most of the people. This has resulted into the digitization of most of the institutes as it is easy and convenient for the people and also for the authority to maintain the records. So, it resulted in most of the banks and other institutes receiving and transferring money through credit cards. But with the hackers and other cyber criminals around there is always chances of the frauds in the transactions. The possibility of the fraud transaction is very less but it is not negligible and even having one fraud transaction is unacceptable because it is crime and we can't neglect it even if it is very less as it harms both the customer and credibility of the institute. So this paper aims at analyzing various classification techniques using various metrics for judging various classifiers. This model aims at improving fraud detection rather than misclassifying a genuine transaction as fraud. This paper has the disadvantage of outliers and standalone application and advantage of beneficial for the organizations and for individual users in terms of cost and time efficiency.

[5] Chunzhi Wang, Yichao Wang, Zhiwei Ye, Lingyu Yan, Wencheng Cai, Shang Pan presented "Credit card fraud detection based on whale algorithm optimized BP neural network" published in 2018 states that this paper proposes a credit card fraud detection technology based on whale algorithm optimized BP neural network aiming at solving the problems of slow convergence rate, easy to fall into local optimum, network defects and poor system stability derived from BP neural network. Using whale swarm optimization algorithm to optimize the weight of BP network, we first use WOA algorithm to get an optimal initial value, and then use BP network algorithm to correct the error value, so as to obtain the optimal value. This paper has the advantage of higher percentage accuracy of 91% and beyond in detecting fraudulent transactions as compared to the Neural Network model that recorded 89.6% and disadvantage of standalone application.

III. METHODOLOGY

i. Local Outlier Factor

In anomaly detection, the local outlier factor (LOF) is an algorithm proposed by Markus M. Breuning, Hans-Peter Kriegel, Raymond T.N and Jorg Sander in 2000 for finding anomalous data points by measuring the local deviation of a given data point with respect to its neighbors. LOF shares some concepts with DBSCAN and OPTICS such as the concepts of "core distance" and "reachability distance", which are used for local density estimation. The local outlier factor is based on a concept of a local density, where locality is given by {display style k} k nearest neighbors, whose distance is used to estimate the density. By comparing the local density of an object to the local densities of its neighbors, one can identify regions of similar density, and points that have a substantially lower density than their neighbors. These are considered to be outliers. The local density is estimated by the typical distance at which a point can be "reached" from its neighbors. The definition of "reachability distance" used in LOF is an additional measure to produce more stable results within clusters.

ii. Random Forest Algorithm

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates 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. Random decision forests correct for decision trees' habit of over fitting to their training set. The first algorithm for random decision forests was created by Tin Kam Ho using the random subspace method, which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg. An extension of the algorithm was developed by Leo Breiman and Adele Cutler, who registered "Random Forests" as a trademark (as of 2019, owned by Minitab, Inc.). The extension combines Breiman's "bagging" idea and random selection of features, introduced first by Ho and later independently by Amit and Geman in order to construct a collection of decision trees with controlled variance.

iii. Training and Testing the model for accuracy

Here, the model will be trained using the datasets and tested for finding the accuracy of the model. Optimization will be done to improve the accuracy if needed. In machine learning, a common task is the study and construction of algorithms that can learn from and make predictions on data. Such algorithms work by making data-driven predictions or decisions, through building a mathematical model from input data. The data used to build the final model usually comes from multiple datasets. In particular, three data sets are commonly used in different stages of the creation of the model.

The model is initially fit on a training dataset, that is a set of examples used to fit the parameters (e.g. weights of connections between neurons in artificial neural networks) of the model. The model (e.g. a neural net or a naive Bayes classifier) is trained on the training dataset using a supervised learning method (e.g. gradient descent or stochastic gradient descent). In practice, the training dataset often consist of pairs of an input vector (or scalar) and the corresponding output vector (or scalar), which is commonly denoted as the target (or label). The current model is run with the training dataset and produces a result, which is then compared with the target, for each input vector in the training dataset. Based on the result of the comparison and the specific learning algorithm being used, the parameters of the model are adjusted. The model fitting can include both variable selection and parameter estimation.

Successively, the fitted model is used to predict the responses for the observations in a second dataset called the validation dataset. The validation dataset provides an unbiased evaluation of a model fit on the training dataset while tuning the model's hyper parameters (e.g. the number of hidden units in a neural network). Validation datasets can be used for regularization by early stopping: stop training when the error on the validation dataset increases, as this is a sign of overfitting to the training dataset. This simple procedure is complicated in practice by the fact that the validation dataset's error may fluctuate during training, producing multiple local minima. This complication has led to the creation of many ad-hoc rules for deciding when overfitting has truly begun.

Finally, the test dataset is a dataset used to provide an unbiased evaluation of a final model fit on the training dataset. If the data in the test dataset has never been used in training (for example in cross-validation), the test dataset is also called a holdout dataset.

iv. Cloud based deployment process of the model

Here, the model will be deployed on a cloud server to make the solution accessible across the geographical areas. For the cloud deployment process, we use either of Amazon web service or the Google Cloud. Here the 3rd party application is available across world to use for the client, they can approach the service to use the project.

IV. EXISTING SYSTEM

Credit card fraud detection is most popular area where most of the research is carried out and many algorithms and techniques are used to detect the fraud. Many earlier system models are designed using Markov model and also using neural networks. Even though models are developed using many techniques there may occur few updating features and also few drawbacks.

- Most of them are standalone application, which mean they can't be re-used by existing application.
- Even though maximum number of outliers are removed the results are incorrect many times.
- Accuracy rate of the existing system are low.

V. PROPOSED SYSTEM

Throughout the financial sector, machine learning algorithm are being developed to detect the fraudulent transactions. In this project, that is exact thing what we are going to be doing as well. Using datasets of nearly 28,500 credit card transactions and multiple unsupervised anomaly detection algorithms, we are going to identify the transaction with high probability of being credit card fraud. Furthermore, using metrics such as precision, recall, and F1-scores, we will investigate why the classification accuracy for these algorithms can be misleading. In addition, we will explore the use of data visualization techniques common in data science, such as parameter histograms and correlation matrices, to gain a better understanding of the underlying distribution of data in our data set. Advantages of the proposed system are

- Proven high accuracy
- Memory and time efficient
- Solution made available to public over the cloud in as-a-service model

REFERENCES

[1] Hyder John, Sameena Naaz "Credit Card Fraud Detection using Local Outlier Factor and Isolation Forest", Vol-7, Issue -4, April 2019.

[2] Machine Learning Group, ―Credit Card Fraud Detection, Kaggle,23- Mar-2018.

[3] Mr. Ibtissam Benchaji, Samira Douzi, Bouabid El Ouahidi "Using Genetic Algorithm to Improve Classification of Imbalanced Datasets for credit card fraud detection" published in 2018.

[4] Mr. Hongyu Wang, Ping Zhu, Xueqiang Zou, Sujuan Qin "An Ensemble Learning Framework for Credit Card Fraud Detection based on Training Set Partitioning and Clustering" published on 2018.

[5] Yvan Lucas, Pierre-Edouard Portier, Lea Laporte, Sylvie Calabretto, Liyun He-Guelton, Frederic Oble, Michael Granitzer "Dataset shift quantification for credit card fraud detection" published in 2019.

[6] Ankit Mishra, Chaitanya Ghorpade presented paper on "Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques" published in 2018.

[7] Chunzhi Wang, Yichao Wang, Zhiwei Ye, Lingyu Yan, Wencheng Cai, Shang Pan presented "Credit card fraud detection based on whale algorithm optimized BP neural network" published in 2018.