Credit Card Fraud Detection For Banking Support With Detailed Analysis Using Several Supervised Machine Learning Techniques

Chinnadurai S[1], Venkata Sai Manikanta N[2], Abhiram Chowdary M[3], Venkata Abhilash P[4], Sathish K[5]
[1, 2, 3, 4, 5] Dept of CSE
[1, 2, 3, 4, 5] Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur, TN, INDIA

Abstract- *Nowadays credit card is more popular among the private and public employees. By using the credit card, the users purchase the consumable durable products in online, also transferring the amount from one account to other. Hence, a credit card is the most widely used electronic payment method because of the increasing volume of daily electronic transactions, making it more vulnerable to fraud. The fraudster is detecting the details of the behavior user transaction and doing the illegal activities with the card by phishing, Trojan virus, etc. The fraudulent may threaten the users on their sensitive information.*

Keywords- Machine Learning, Credit Card Fraud Detection.

I. INTRODUCTION

The number of customers for the credit cards (CC) has grown of the last decade. These cards have aggravated the cashless systems of payments as well as alleviated the use of cash credit, which is termed as a short-term continuous loan. The credit cards are known to increase the purchasing power of citizens, and let them meet their daily needs, gadgets, etc. The number of CC (credit card) frauds has increased with the increase in number of credit cards. The unethical use of credit cards by hackers or credit cards users unwilling to pay back the amount are known as the major credit card frauds. The credit card frauds can be detected by evaluating the CC purchasing patterns using the historical data in order to detect the frauds. This data evaluation can help the banks or other organizations offering credit cards to minimize their losses due to the credit card frauds. The historical data evaluation with the current purchasing patterns requires the statistical modelling, which can automatically the fraudulent patterns and alarm the banks about the transactions. This helps the banks for early detection of the frauds, where they can easily eliminate the CC frauds by declining the suspected transactions.

II. SOME MACHINE LEARING METHODS

Machine learning algorithms are often categorized as supervised or unsupervised. Supervised algorithms require a data scientist or data analyst with machine learning skills to provide both input and desired output, in addition to furnishing feedback about the accuracy of predictions during algorithm training. Data scientists determine which variables, or features, the model should analyze and use to develop predictions. Once training is complete, the algorithm will apply what was learned to new data. Unsupervised algorithms do not need to be trained with desired outcome data. Instead, they use an iterative approach called deep learning to review data and arrive at conclusions. Unsupervised learning algorithms -- also called neural networks -- are used for more complex processing tasks than supervised learning systems, including image recognition, speech-to-text and natural language generation. These neural networks work by combing through millions of examples of training data and automatically identifying often subtle correlations between many variables. Once trained, the algorithm can use its bank of associations to interpret new data. These algorithms have only become feasible in the age of big data, as they require massive amounts of training data.

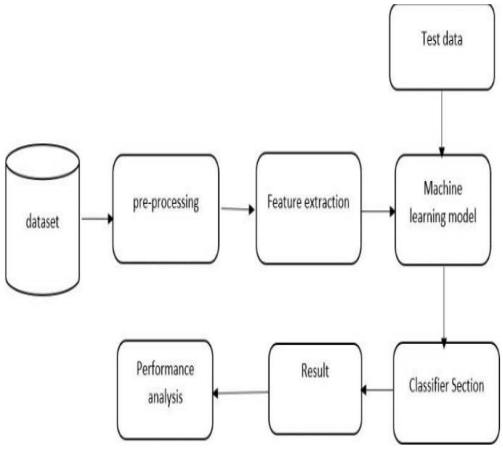
III. RELEATED WORKS

The existing systems use supervised machine learning such as DT, RF, and SVM. The scope of the traditional model is limited to only few cases as they have low accuracy and feasibility. The existing model does not work well terms of performance and precision due to a single algorithm approach and overfitting problem. It Only identifies the fraud, no alert system, if present, the alert takes some time to pass and slower response. We have come across various fraud detection techniques that exists today but none of them were competent enough to detect the fraud at the time it actually took place. They detected the frauds which happened in the past. The setback of all the techniques discussed so fargive accurate results only when performed on a particular dataset and sometimes with some special features only. But

we need to establish a technology that works equally precisely and accurately under all circumstances and with various datasets. Techniques like Logistic Regression works better than SVM when there is a class imbalance and comparatively Random Forest performs better among all three.

IV. PROPOSED SYSTEM

The objectives of the project is to implement machine learning algorithms to detect credit card fraud detection with respect to time and amount of transaction. The procedure which we followed to predict the result are understanding problem statement and data by performing statistical analysis and visualization then checking whether the data is balance or not, In this data set the data is imbalanced, balanced by using oversampling, then scaling the data using standardization and normalization and testing data with different ML algorithms.For any data science project some package are very important such as that is numeric python And pandas and for visualization of the data, matplotlib and seaborn is used which build on matplotlib with some extra features. Logistic Regression an appropriate technique that can be used in predictive analysis when the dependent variable is dyadic or binary. Since the categorization of transactions being fraud is a double-edged variable, this technique can be used. This statistical classification model based on probabilities detects the fraud using logistic curve. Since the value of this logistic curve varies from 0 to 1, it can be used to interpret class membership probabilities. The dataset fed as input to the model is being classified for training and testing the model. Post model training, it is tested for some minimum threshold cut-off value for prediction. Then the most significant variables are selected and the model is tuned accordingly. The accuracy of prediction came out to be 70%. Since the logistic regression, based on some threshold probabilities can divide the plane using a single line and divides dataset points into exactly two regions. Hence, the outlier points are not handled effectively. It uses natural logarithmic function to calculate probability and to show that the results fall under a particular category



V. IMPLEMENTATION AND EXECUTION

1. Data Collection: Data used in this paper is a set of product reviews collected from credit card transactions records. This step is concerned with selecting the subset of all available data that you will be working with. ML problems start with data preferably, lots of data (examples or observations) for which you already know the target answer. Data for which you already know the target answer is called labelled data.

2. Pre-processing: Since the datasets are text data, pre-processing is employed for the machine learning. Pre-processing is the process of three important and common steps as follows:

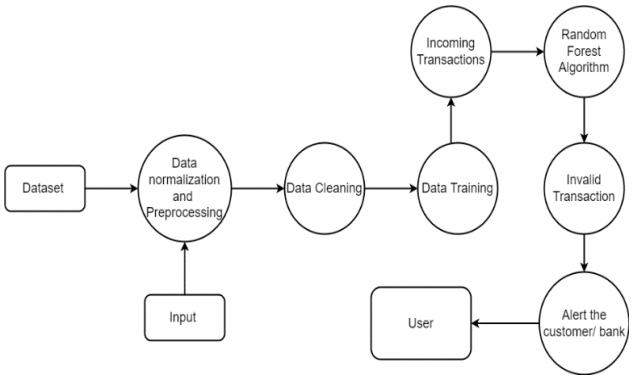
   - Formatting: It is the process of putting the data in a legitimate way that it would be suitable to work with. Format of the data files should be formatted according to the need. Most recommended format is .csv files.
   - Cleaning: Data cleaning is a very important procedure in the path of data science as it constitutes the major part of the work. It includes removing missing data and complexity with naming category and so on. For most of the data scientists, Data Cleaning continues of 80% of work.
   - Sampling: This is the technique of analyzing the subsets from whole large datasets, which could provide a better result and help in understanding the behavior and pattern of data in an integrated way

3. Data visualization Data Visualisation is the method of representing the data in a graphical and pictorial way, data scientists depict a story by the results they derive from analysing and visualising the data. The best tool used is

Tableau which has many features to play around with data and fetch wonderful results

4. Feature extraction Feature extraction is the process of studying the behavior and pattern of the analyzed data and draw the features for further testing and training. Finally, our models are trained using the Classifier algorithm. We use classify module on Natural Language Toolkit library on Python. We use the labelled dataset gathered. The rest of our labelled data will be used to evaluate the models. Some machine learning algorithms were used to classify pre-processed data. The chosen classifiers were Randomforest. These algorithms are very popular in text classification tasks.

5. Result evaluation – In this section, we evaluate our Result and also define the evaluation criteria to calculate the performances of our classification models. Model Evaluation is an essential part of the model development process. It helps to find the best model that represents our data and how well the selected model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can effortlessly generate overoptimistically and over fitted models. To avoid overfitting, evaluation methods such as hold out and cross-validations are used to test to evaluate model performance. The result will be in the visualized form. Representation of classified data in the form of graphs. Accuracy is well-defined as the proportion of precise predictions for the test data. It can be calculated easily by mathematical calculation i.e. dividing the number of correct predictions by the number of total predictions



VI. CONCLUSION

The paper successfully proposes that artificial intelligence can detect frauds in more efficient and cost-effective ways. As per logistics regression values, the system was able to build a classifier in order to detect fraudulent credit card transactions. The success is evident when the model is compared to some other well known classifiers of frauds like voting classifier and the support vector machine classifier. The proposed classifier is easier to use and shows more accurate results. Besides that, the proposed classifier has been able to answer the intended research questions: How can AI build a system of detecting fraud? How can it also deal with imbalanced data effectively? How clean can the data be made using this AI system before putting it to use for detection? How can the AI system detect frauds by adapting to the changing behavior of the user.

VII. ACKNOWLEDGEMENT

It is with immense pleasure that I present my first venture in the field of real applications of computing in the form of a project work. First of all, I am indebted to the Almighty for his choicest blessing showered on me in completing this endeavor. I express my sincere thanks to Department of CSE, Dhanalakshmi Srinivasan Engineering College, Perambalur.

REFERENCES

[1] Dahee Choi, KyunghoLee."Machine Learning Based Approach to Financial Fraud Detection Process in Mobile Payment System." IEEE 2018

[2] Vijayshree B. Nipane et al. "Fraudulent Detection in Credit Card System Using SVM & Decision Tree" International Journal of Scientific Development and Research, Volume 1, Issue 5 (2016)

[3] RuttalaSailusha et al. "Credit Card Fraud Detection Using Machine Learning" 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). 2020.

[4] AnuruddhaThennakoon et al. "Real-time Credit Card Fraud Detection Using Machine Learning". 2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE 2019

[5] Samidha Khatri et al "Supervised Machine learning algorithms for Credit Card Fraud Detection: A comparison." 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE 2020

[6] Emmanuel Illebri et al. 2.6 "Performance Evaluation of Machine Learning Methods for Credit Card Fraud Detection Using SMOTE and AdaBoost", IEEE 2021

[7] Ebenezer Esenogho et al "A Neural Network Ensemble With Feature Engineering for Improved Credit Card Fraud Detection." IEEE 2022

[8] Kuldeep Randhawa et al "Credit Card Fraud Detection Using AdaBoost and Majority Voting." IEEE 2018
[9] Wen-Fang Yu et all "Research on Credit Card Fraud Detection Model Based on Distance Sum.", 2019 International Joint Conference on Artificial Intelligence, IEEE 2019
[10] Parth Roy et al "Comprehensive Analysis of Fraud Detection of Credit Cards through Machine Learning.", 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE 2021