Students Achievements:

Paper Publication:

Sl.No	Students Name	Title of the Paper Published
1	Lahari Bale	*A NOVEL APPROACH IN CREDIT CARD FRAUD DETECTION SYSTEM USING MACHINE LEARNING TECHNIQUES*
2	Suhas Jain	
3	Pranavi K	
4	S. Gowtham	*Text Detection and Language Identification on Natural Scene Images using Faster R-CNN*

A NOVEL APPROACH IN CREDIT CARD FRAUD DETECTION SYSTEM USING MACHINE LEARNING TECHNIQUES

Suhas Jain G. M.[1], N. Rakesh[2], Pranavi K[3], Lahari Bale[4]
[1,2,3] Department of Artificial Intelligence and Machine Learning
BMS Institute of Technology and Management
Bengaluru, Karnataka.
{1by19ai056@bmsit.in[1],n_rakesh@bmsit.in[2], 1by19ai025@bmsit.in[3], 1by19ai026@bmsit.in[4]}

2021 International Conference on Forensics, Analytics, Big Data, Security (FABS) | 978-1-6654-2005-1/21/$31.00 ©2021 IEEE | DOI: 10.1109/FABS52071.2021.9702672

Abstract: **With the rapid expansion of daily life, the use of credit cards for online purchases is steadily increasing and credit card fraud is on the rise. Nowadays, in the social distancing environment, due to covid-1, 9online shopping has become important. Credit card credentials are used to make online payments, and then deduct money which does not involve any contact and makes people's life difficult. Because of this, finding the most effective method of detecting scams in online systems is essential. To prevent customers from being charged for goods they have not purchased, credit card companies must be able to identify fraudulent credit card transactions. Therefore, there are several theories either completed or proceeding to detect these kinds of frauds. This study is an approach to identify non-legitimate transactions using semi-supervised machine learning models by explaining how to deal with imbalanced datasets, using a wide variety of models to better understand which ones work better.**

Keywords—Data Science. Semi-Supervised Classification, Credit Card Fraud.

I. INTRODUCTION

Recently, online purchases using credit cards have increased drastically, people are not generally aware of a probable fraudulent transaction that could happen to them. Credit card security is determined by the card's physical characteristics and the privacy of the card number.As a result of globalization and the growth of Internet-based commerce, worldwide credit card purchases have increased. In addition to the rapid increase in credit card purchases, another important factor contributing to the increase in fraud is credit card fraud. The term credit card fraud is a broad term to refer to theft and fraud committed as a source of fraudulent funding in a particular interaction using a credit card. Theft and fraud are committed in a given transaction using a payment card as a fraudulent source of funds. A vast range of methods to conduct theft are used by Credit Card Fraudsters. To successfully combat credit card fraud, it is important to have a basic understanding of the process of detecting credit card fraud. Due to numerous credit card fraud monitoring and avoidance mechanisms, credit card fraud has stabilized a lot over the years. However, cardholders use fake transactions to scam bank cash. External card fraud, on the other hand, is primarily expressed in the use ofstolen fraudulent, stolen credit cards to consume or obtain cash in concealed ways, such as buying valuable, limited amounts of products or items that are easy to sell in cash. This project will specifically explore &analyze the development of a machine learning-based fraud detection system.

II. OBSERVATION

Fraud detection by MasterCard is a serious drawback according to Bhattacharyya.In observed studies, the utilization of information mining approaches for detecting credit card fraud is comparatively low, most likely due to a lack of readily available information [1]. By using authority and mobility to simulate synthetic data, they can remove customer privacy and security restrictions associated with real data when financial fraud is detected. To do this, researchers and the general public need to create a simple set of synthetic financial data. [2].The current systems used across the business sector are used to detect fraud in different ways.Each system detects fraud differently.Using the application, most users agree that the overall system has proved to be effective and meaningful for fraud detection [3].Most security weaknesses were attributed to credit card fraud. In this study, we have explored different solutions to the same problems and identified customers and fraud. [4]. Ashen et al investigated the effectiveness of the credit card fraud detection model. The authors propose three methods of classification, namely decision trees, neural networks, and logistic regression. Logistic regression and neural networks performed really well while compared to decision tree [5].Y.Sahin, E. Duman(2011) Excerpts from studies using artificial neural networks and regression classification explain that ANN classifiers perform better than LR classifiers in problem solving. In this scenario, the distribution of training datasets is biased, and the effectiveness of all models is reduced in capturing fraudulent transactions [6]. Known as unsupervised classification methods, these methods help detect anomalous behavior within a system and uncover transactions that may be fraudulent [7]-[8].

III. PROPOSED METHODOLOGY

In credit card transactions, various fraudulent activity detections have been implemented so far. We use different techniques. In September 2013, European cardholders made over 2,84,000 credit card transactions using the dataset we are using. In this dataset, frauds make up 0.172% of all transactions, an extremely unbalanced number. The input variables are all numeric and the result of a PCA

transformation. We will use various techniques to find fraudulent behavior intransactions.

A. Data Visualization:
We will use some data visualization tools in our method to get some insights about our dataset. We have used matplotlib and seaborn for this purpose.

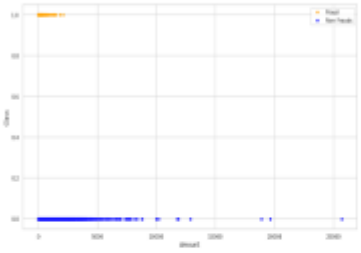

Fig.1. Data Distribution

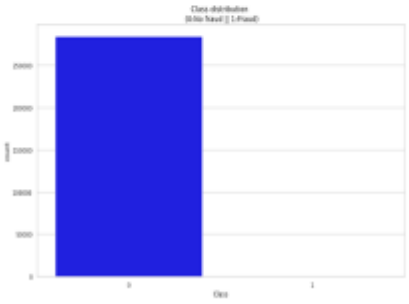

Fig.2.Data distribution

By seeing these two distribution diagrams we can say that our dataset is highly imbalanced. We have a very high number of non-fraudulent transactions whereas fraudulent transactions are just countable such that we cannot even see the distribution in the diagrams.

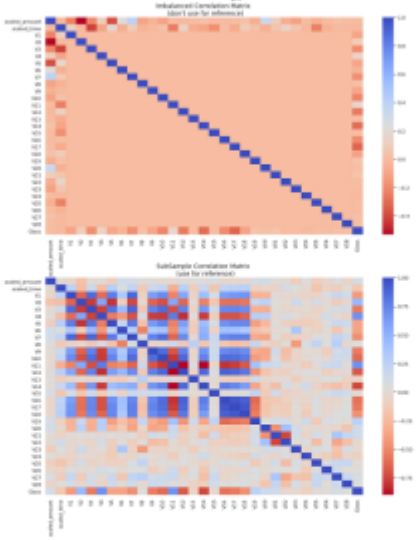

Fig 3. Correlation Matrix

B. Data Preprocessing:
a) Distribution:The dataset consists of 2,84,807 credit card transactions, out of which only 492 transactions are fraudulent. Our dataset then has a significantly skewed distribution. We will randomly select 492 transactions out of fraud transactions in order to have a 50:50 ratio in our dataset.

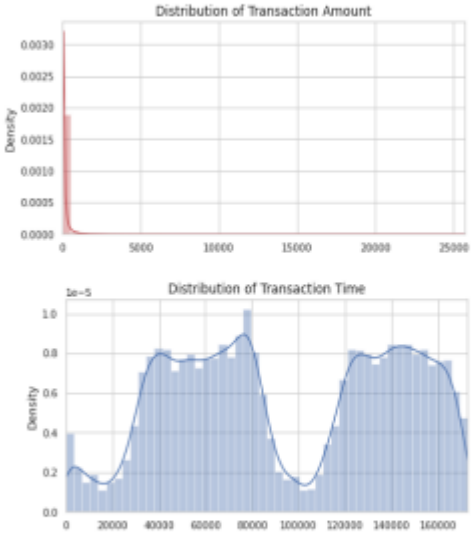

Fig 4. Data Distribution

Above we have drawn the graphs for the distribution of Transaction Amount and Transaction Time classes.

b) Anomaly detection:In order to avoid overfitting we need to remove all the outliers in our dataset. We will use anomaly detection techniques for removing outliers. Once we have all the classes that are highly correlated with the dependent variable, we will remove the extreme outliers. Here are some of the features before removing outliers and after removing outliers. We will draw the positively correlated classes using Boxplot.

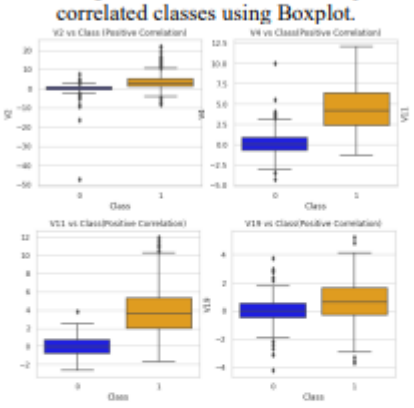

Fig 5. Boxplot of positively related classes

978-1-6654-2005-1/21/$31.00 ©2021 IEEE

We will draw the boxplot for negatively related classes just them again after outliers removal.

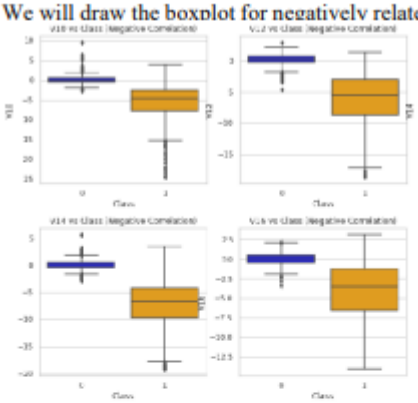

Fig 6.Boxplot of negatively related classes

Then plot the distribution of all highly correlated values.

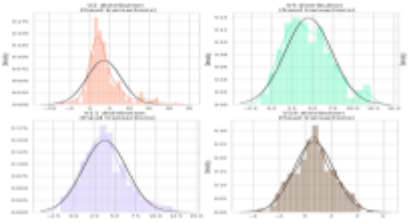

Fig 7. Distribution of Positively Correlated Classes

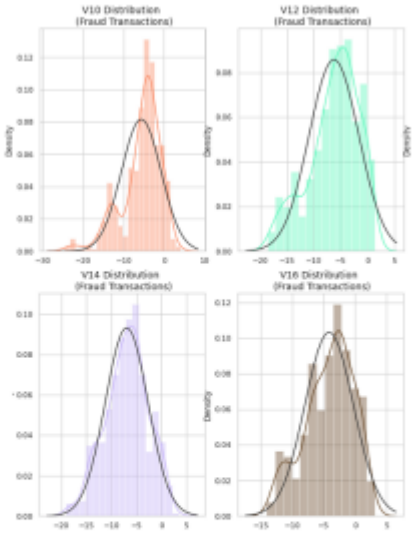

Fig 8.Distribution of Positively Correlated Classes

We will then decide the threshold for Anomaly Detection based on the distributions of these classes. We will visualize

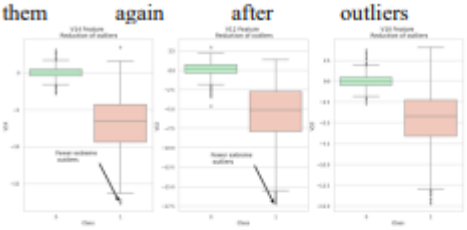

Fig.9.Distribution after anomaly detection

c) Dimensionality reduction: The dataset consists of more than 30 input variables. Datasets with many variables will be reduced using dimensionality reduction techniques. We will perform dimensionality reduction so as to capture its essence for our data. Using PCA,t-SNE, Truncated SVD techniques for dimensionality reduction and we will select one method at the end. Below are the results.

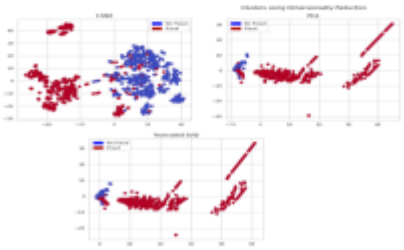

Fig 10.Visualization after dimensionality reduction

Wehave further used,a) Logistic Regression: Additionally, We used a classification algorithm, Logistic Regression, which is an algorithm for predicting binary values (1 / 0, Yes / No, False / True) from a set of independent variables. Predicts the probability of an event occurring as a function of a dependent variable when the resulting variable is categorical. Your logistic regression then becomes a linear regression.

b) XGBoost: Gradient boosted trees can also be implemented with XGBoost. It's an open-source program that's popular and efficient. Gradient boosting Combine estimates from a set of simpler and weaker models in an attempt to accurately predict target variables. The gradient boosted trees algorithm is implemented by the open-source software XGBoost. The supervised learning algorithm combines the Controlled learning algorithms include estimating arrays of weaker and simpler models to predict target variables with higher accuracy. A regression tree is usually used as a weak learner in gradient boosting, and a regression treemaps its input data to the leaf containing the continuous score. By combining a convex loss function (predicting the target output in terms of the predicted output) with a penalty term for model complexity, XGBoost minimizes a regularized (L1 and L2) objective function. In each iteration of training, additional trees are added that predict the remnants of the previous tree. These new trees are then combined with the previous tree to make the final

prediction. Gradient boosting reduces losses when adding a new model by using a gradient descent algorithm.

.

IV. RESULTS

Metrics for performance: A confusion matrix serves as a basic measure for performance.The confusion matrix consists of two by two matrix tables with four outcomes produced by the binary classifier.The confusion matrix provides a variety of measures including sensitivity, specificity, accuracy, and error rate. The accuracy of the prediction is calculated as the sum of two correct predictions (P+Q) divided by the total number of datasets (R+S).Essentially, it is (1-error rate).

$$A=P+Q/R+S \quad (1)$$

Where,
A=Accuracy
P=True Positives
Q=True Negatives
R=False Positives
S=false Negatives

We will find the ROC score and cross-validation score for both models in order to validate the model.

TABLE 1. CCROSS-VALIDATIONSCORE VS ROC_AUC

	Logistic Regression	XGBoost
Cross Validation score	0.9416	0.9376
ROC_AUC Score	0.9400	0.9355

And the results for both Logistic regression and XGBoost models are represented below using a confusion matrix.

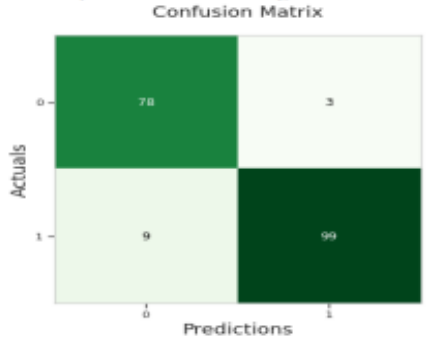

Fig 11. Result of Logistic Regression

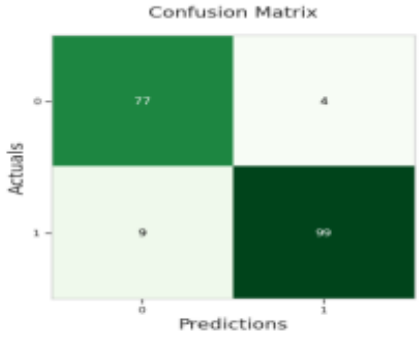

Fig 12. Result of XGBoost

We have got pretty much the same accuracy from both models. Both of the classification models have performed well, we have very few false positives and false negatives false positives must be reduced as much as possible. Logistic regression has an accuracy score of 93.65, while XGBoost has an accuracy score of 93.12. XGBoost appears to be more accurate than Logistic Regression based on these results.

V. CONCLUSION

Credit card fraud is dishonest. Machine learning can help improve fraud detection results in conjunction with the logistic regression algorithm, xgboost.While we have used only 984 transactions out of 2,84,000, we have lost a lot of information.On the whole dataset, one can try implementing one method . Another drawback is that we cannot determine the names of fraud and non-fraud transactions for the given dataset using machine learning.The project can be further developed by finding a way to address this issue using various methods.

REFERENCES

[1] *S.* Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Syst., vol. 50, no. 3, pp. 602613, 2011.

[2] E. A. Lopez-Rojas and S. Axelsson, "A review of computer simulation for fraud detection research in financial datasets," in 2016 Future Technologies Conference (FTC), 2016, pp. 932–935.

[3] D. Al-Jumeily, A. Hussain, A. MacDermott, G. Seeckts, and J. Lunn, "Methods and techniques to support the development of fraud detection system," in 2015 International Conference on Systems, Signals and Image Processing (IWSSIP), 2015, pp. 224–227.

[4] Ayushi Agrawal, Shiv Kumar, and Amit Kumar Mishra, "Implementation of Novel Approach for Credit Card Fraud Detection," in 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), 2015, pp. 1–4.

[5] A. Shen, R. Tong, Y. Deng, "Application of classification models on credit card fraud detection", Service Systems and Service Management 2007 International Conference, pp. 1-4, 2007.

[6] Y. Sahin, E. Duman, "Detecting credit card fraud by ANN and logistic regression", Innovations in Intelligent Systems and Applications (INISTA) 2011 International Symposium, pp. 315-319, 2011.

[7] F. T. Liu, K. M. Ting, and Z. H. Zhou, "Isolation forest," in Proc. the Eighth IEEE International Conference on Data Mining, 2008, pp. 413-422.

[8] C. S. Hemalatha, V. Vaidehi, and R. Lakshmi, "Minimal infrequent pattern-based approach for mining outliers in data streams," Expert Systems with Applications, vol. 42, no. 4, pp. 1998-2012, March 2015