Credit Card Fraud Detection Identify Using Machine Learning and Data Science

[1]Sariki Pradeep, [2]V.Veerendra Subhash
[1,2]Dept. of Computer Science & Engineering, KIET, Kakinada, AP, India

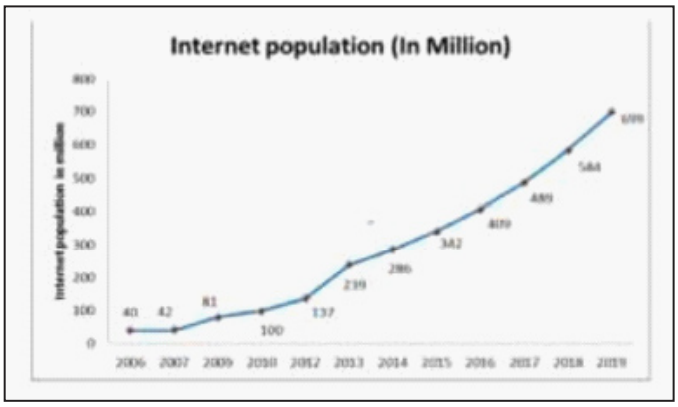
Abstract
Credit card fraud detection is presently the most frequently occurring problem in the present world. This is due to the rise in both online transactions and e-commerce platforms. Credit card fraud generally happens when the card was stolen for any of the unauthorized purposes or even when the fraudster uses the credit card information for his use. In the present world, we are facing a lot of credit card problems. To detect the fraudulent activities the credit card fraud detection system was introduced. This project aims to focus mainly on machine learning algorithms. The algorithms used are random forest algorithm ,linear regression , XGBoost ,KNearest, Support vector classifier, Linear Discriminant Analysis, GaussianNB algorithm. The results of the algorithms are based on accuracy, precision, recall, and F1-score. The ROC curve is plotted based on the confusion matrix. Algorithms are compared and the algorithm that has the greatest accuracy, precision, recall, and F1-score is considered as the best algorithm that is used to detect the fraud.

I. Introduction

Credit card fraud is a significant threat in the BFSI sector. This credit card fraud detection system studies and analyzes user behavior patterns and uses location scanning techniques to identify any unusual patterns. One of The user patterns includes important user behavior like spending habits, usage patterns, etc. The system uses geographic location for identity verification. In case it detects any unusual pattern, the user will be required to undergo the verification. The fraud detection system stores thepast transaction data of each user. Based on this data, it calculates the standard user behavior patterns for individual users, and any deviation from those normal patterns becomes a trigger for the system. In the instance of any unusual activity, the system will not only raise alerts, but it will also block the user after three invalid attempts.

'Fraud' in credit card transactions is unauthorized and unwanted usage of an account by someone other than the owner of that account. Necessary prevention measures can be taken to stop this abuse and the behaviour of such fraudulent practices can be studied to minimize it and protect against similar occurrences in the future. In other words, Credit Card Fraud can be defined as a case where a person uses someone else's credit card for personal reasons while the owner and the card issuing authorities are unaware of the fact that the card is being used. Fraud detection involves monitoring the activities of populations of users in order to estimate, perceive or avoid objectionable behaviour, which consist of fraud, intrusion, and defaulting. This is a very relevant problem that demands the attention of communities such as machine learning and data science where the solution to this problem can be automated. This problem is particularly challenging from the perspective of learning, as it is characterized by various factors such as class imbalance. The number of valid transactions far outnumber fraudulent ones. Also, the transaction patterns often change their statistical properties over the course of time.

These are not the only challenges in the implementation of a real-world fraud detection system, however. In real world examples, the massive stream of payment requests is quickly scanned by automatic tools that determine which transactions to authorize. Machine learning algorithms are employed to analyse all the authorized transactions and report the suspicious ones. These reports are investigated by professionals who contact the cardholders to confirm if the transaction was genuine or fraudulent. The investigators provide a feedback to the automated system which is used to train and update the algorithm to eventually improve the fraud-detection performance over time.



II. Literature Survey

New methods for credit card fraud detection with a lot of research methods and several fraud detection techniques with a special interest in the neural networks, data mining, and distributed data mining. Many other techniques are used to detect such credit card fraud. When done the literature survey on various methods of credit card fraud detection, we can conclude that to detect credit card fraud there are many other approaches in Machine Learning itself. The research on credit card fraud detection uses both Machine Learning and Deep Learning algorithms.

In this section, we enhance the work done in two different points:

(i). the methods that are readily available for fraud detection
(ii). The techniques that are available to handle the imbalanced data.

To handle the imbalanced data some of the techniques are available. They are

(a). classification methods
(b). sampling methods
(c). resembling techniques.

Here are some of the Machine Learning algorithms that are used for credit fraud detection are support vector machine(SVM), decision trees, logistic regression, gradient boosting, K-nearest neighbor, etc.

In 2019, Yashvi Jain, NamrataTiwari, ShripriyaDubey, Sarikajain have researched various techniques[10] for credit cards fraud detection such as support vector machines(SVM), artificial neural networks(ANN), Bayesian Networks, Hidden Markov Model, K-Nearest Neighbours (KNN) Fuzzy Logic system and Decision Trees. In their paper, they have observed that the algorithms

k-nearest neighbor, decision trees, and the SVM give a medium level accuracy. The Fuzzy Logic and Logistic Regression give the lowest accuracy among all the other algorithms. Neural Networks, naive bayes, fuzzy systems, and KNN offer a high detention rate. The Logistic Regression, SVM, decision trees offer a high detection rate at the medium level. There are two algorithms namely ANN and the Naïve Bayesian Networks which perform better at all parameters. These are very much expensive to train. There is a major drawback in all the algorithms. The drawback is that these algorithms don't give the same result in all types of environments. They give better results with one type of datasets and poor results with another type of dataset. Algorithms like KNN and SVM give excellent results with small datasets and algorithms like logistic regression and fuzzy logic systems give good accuracy with raw and unsampled data.

In 2019, HetaNaik, PrashastiKanikar, has done their research on various algorithms like Naïve Bayes, Logistic Regression, J48, and Adaboost. Naïve Bayes on among the classification algorithm. This algorithm depends upon Bayes theorem. Bayes's theorem finds the probability of an event that is occurring is given. The Logistic regression algorithm is similar to the linear regression algorithm. The linear regression is used for the prediction or forecasting the values. The logistic regression is mostly used for the classification task. The J48 algorithm is used to generate a decision tree and is used for the classification problem. The J48 is the extension of the ID3 (Iterative Dichotomieser). J48 is one of the most widely used and extensively analyzed areas in Machine Learning. This algorithm mainly works on constant and categorical variables. Adaboost is one of the most widely used machine learning algorithms and is mainly developed for binary classification. The algorithm is mainly used to boost the performance of the decision tree. This is also mainly used for the classification of the regression. The Adaboost algorithm is fraud cases to classify the transactions which are fraud and non-fraud. From their work they have concluded that the highest accuracy is obtained for both the Adaboost and Logistic Regression. As they have the same accuracy the time factor is considered to choose the best algorithm. By considering the time factor they concluded that the Adaboost algorithm works well to detect credit card fraud.

In 2019 Sahayasakila V, D.KavyaMonisha,
Aishwarya, Sikhakolli VenkatavisalakshiswshaiYasaswi have explained the Twain important algorithmic techniques which are the Whale Optimization Techniques (WOA) and SMOTE (Synthetic Minority Oversampling Techniques). They mainly aimed to improve the convergence speed and to solve the data imbalance problem. The class imbalance problem is overcome using the SMOTE technique and the WOA technique. The SMOTE technique discriminates all the transactions which are synthesized are again resampled to check the data accuracy and are optimized using the WOA technique. The algorithm also improves the convergence speed, reliability, and efficiency of the system.

In 2018 NavanushuKhare and SaadYunusSait have explained their work on decision trees, random forest, SVM, and logistic regression. They have taken the highly skewed dataset and worked on such type of dataset. The performance evaluation is based on accuracy, sensitivity, specificity, and precision. The results indicate that the accuracy for the Logistic Regression is 97.7%, for Decision Trees is 95.5%, for Random Forest is 98.6%, for SVM classifier is 97.5%. They have concluded that the Random Forest algorithm has the highest accuracy among the other algorithms and is considered as the best algorithm to detect the fraud. They also concluded that the SVM algorithm has a data imbalance problem and does not give better results to detect credit card fraud.

Methodology: System development method is a process through which a product will get completed or a product gets rid from any problem. Software development process is described as a number of phases, procedures and steps that gives the complete software. It follows series of steps which is used for product progress.

III. Model Phases

- **Requirement Analysis:** This phase is concerned about collection of requirements of the system. This process involves generating document and requirement review.
- **System Design:** Keeping the requirements in mind the system specifications are
- translated in to a software representation. In this phase the designer emphasizes on algorithm, data structure, software architecture etc
- **Coding:** In this phase programmer starts his coding in order to give a full sketch of product. In other words, system specifications are only converted in to machine readable compute code.
- **Implementation:** The implementation phase involves the actual coding or programming of the software. The output of this phase is typically the library, executables, user manuals and additional software documentation.
- **Testing:** In this phase all programs (models) are integrated and tested to ensure that the complete system meets the software requirements. The testing is concerned with verification and validation.
- **Maintenance:** The maintenance phase is the longest phase in which the software is updated to fulfill the changing customer need, adapt to accommodate change in the external environment, correct errors and oversights previously undetected in the testing phase, enhance the efficiency of the software.

IV. Reason For Chossing Waterfall Model For Development Process

- Clear project objectives.
- Stable project requirements.
- Progress of system is measurable.
- Strict sign-off requirements.
- Helps you to be perfect.
- Logic of software development is clearly understood.
- Production of a formal specification.
- Better resource allocation.
- Improves quality, the emphasis on requirements and design before writing a single line of code ensures minimal wastage of time and effort and reduces the risk of schedule slippage.
- Less human resources required as once one phase is finished those people can start working on to the next phase.

V. Result

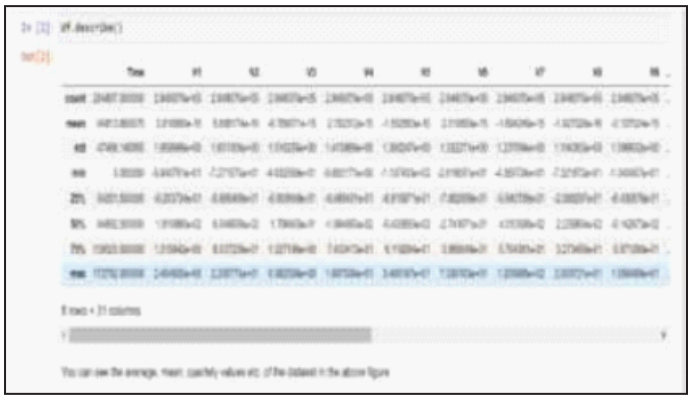
Stage 1. Data gathering

Early data analysis techniques were oriented extracting quantitative and statistical data characteristics this technique facilitate useful data interpretations and can help to get better insights into the processes behind the data .Traditional data can indirectly lead us to knowledge created by humans. Data analysis system has to be equipped with substantial amount of background knowledge and be able to perform reasoning tasks. Let us take for example totally 10 transactions are transacted from account. The input variables

are converted into numerical values by PCA. In this 0.0172 fraud transactions are detected. The class 'Time 'represents the difference in the seconds elapsed between the particular transaction and the first transaction. The class 'Amount ' represents the money transaction that had occurred. Another important feature 'Class' shows whether the transaction is fraudulent or not. The number indication 1 shows that it is a fraud transaction and 0 indicates the non-fraud transactions.

Stage 2. Exploratory Data analysis

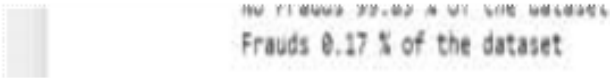
Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns,to spot anomalies,to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.
It is a good practice to understand the data first and try to gather as many insights from it. EDA is all about making sense of data in hand,before getting them dirty with it.
The describe() function in pandas is very handy in getting various summary statistics.This function returns the count, mean, standard deviation, minimum and maximum values and the quantiles of the data.



Dataset comprises of 284807 observations and 31 characteristics.

It is also a good practice to know the columns and their corresponding data types,along with finding whether they contain null values or not.

- Data has only float and integer values.
- No variable column has null/missing values.



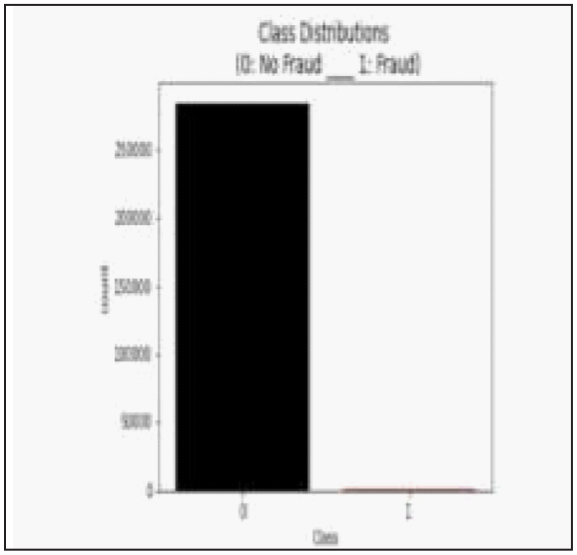
As You can see the dataset is heavily imbalanced What I mean by being imbalanced is that only about 0.17 percent of the target feature is 'fraud' so as a result if we train our model without modifying the dataset our model will generalize the pattern that there are always less no. of frauds that will occur and we do not want that. It will be heavily biased.

For better understanding see the below figure.

We will deal with this problem later. As of now let us gather more insight about the dataset 'time' and 'amount' feature over the instances of the dataset.

As we can see amount feature is highly skewed

VI. Data Preprocessing



The above graphs shows the distribution of the

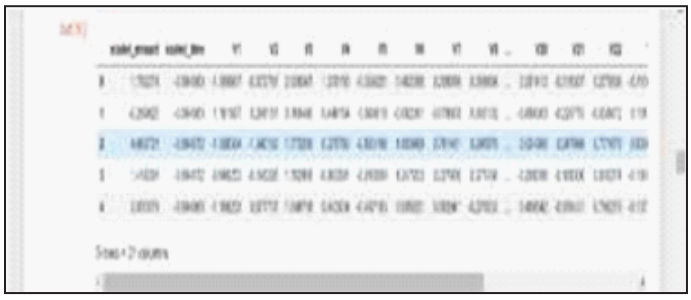
Now we will first scale the columns comprise of Time and Amount. Time and amount should be scaled as the other columns. On the other hand, we need to also create a sub sample of the dataframe in order to have an equal amount of Fraud and Non-Fraud cases, helping our algorithms better understand patterns that determines whether a transaction is a fraud or not.

Subsample is a sample of the dataset where we have equal number of target variables.

We are creating a subsample to overcome the problem of overfittting.

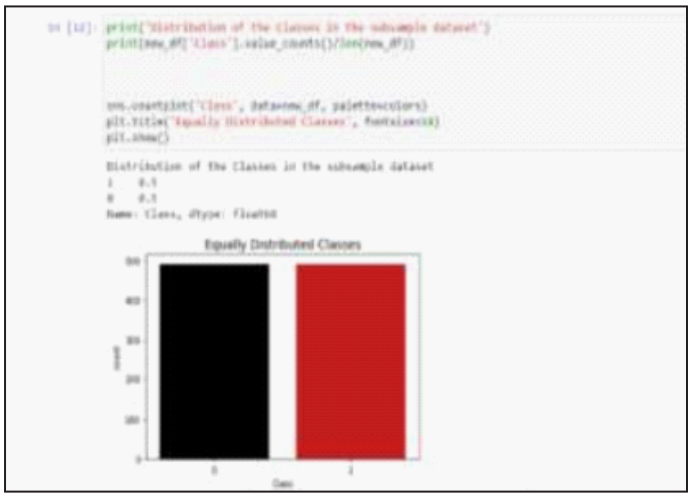
Since most of our data has already been scaled we should scale the columns that are left to scale (Amount and Time).

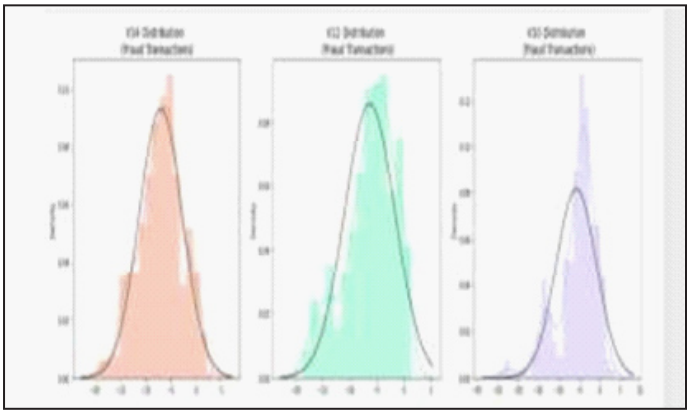
We will use RobustScaler as it is less prone to outliers.

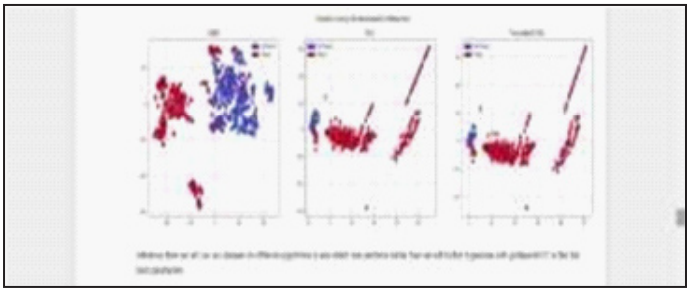


As we can see above the requried scaling variables are scaled Before proceeding with the Random Under Sampling technique we have to separate the orginaldataframe. Why? for testing purposes, remember although we are splitting the data when implementing Random UnderSampling or Over Sampling techniques, we want to test our models on the original testing set not on the testing set created by either of these techniques. The main goal is to fit the model either with the dataframes that were undersample and oversample (in order for our models to detect the patterns), and test it on the original testing set. Random Under Sampling is basically consists of removing data in order to have a more balanced dataset

and thus avoiding our models to overfitting.







Output:-

1. 1.True Positive Rate, which can be defined as the number of fraudulent transactions that are even classified by the system as fraudulent.
2. 2.True Negative Rate, which can be defined as the number of legitimate transactions that are even classified as legitimate by the system.
3. 3. False Positive Rate, which can be defined as a number of the legal transactions which are wrongly classified as fraud.
4. 4. False Negative Rate is defined as the transactions that are fraud but are wrongly classified as legal.

Conclusion

Machine learning algorithms are used for credit card fraud detection. The power of machine learning is used to detect credit cards frauds and the performance of different machine learning algorithms is compared. Three machine learning algorithms, Decision Tree, XGBoost , Linear regression, KNearest , Support vector classifier ,etc are applied on a data set have the data of 284808 credit cards. The performance of XGBoost, KNearest and logistic regression algorithms are found accuracy of 93% percent. And Random overs sampling and Smote techniques accuracy is 97% .The performance of other algorithms is minimum. We conclude that Smote and oversampling techniques yield accurate predictions .

Reference:

[1] ARMEL and D. ZAIDOUNI, "Fraud Detection Using Apache Spark," 2019 5th International Conference on Optimization and Applications (ICOA), Kenitra, Morocco, 2019, pp. 1-6. doi: 10.1109/ICOA.2019.8727610

[2] Z. Kazemi and H. Zarrabi, "Using deep networks for fraud detection in the credit card transactions," 2017 IEEE 4th International Conference on Knowledge-Based Engineering and Innovation (KBEI), Tehran, 2017, pp. 0630-0633. doi: 10.1109/KBEI.2017.8324876

[3] A. Charleonnan, "Credit card fraud detection using RUS and MRN algorithms," 2016 Management and Innovation Technology International Conference (MITicon), Bang-San, 2016, pp. MIT-73- MIT-76. doi: 10.1109/MITICON.2016.8025244

[4] M. Kavitha and M. Suriakala, "Hybrid Multi-Level Credit Card Fraud Detection System by Bagging Multiple Boosted Trees (BMBT)," 2017 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), Coimbatore, 2017, pp. 1-5. doi: 10.1109/ICCIC.2017.8524161

[5] M. Kavitha and M. Suriakala, "Real time credit card fraud detection on huge imbalanced data using meta-classifiers," 2017 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2017, pp. 881-887. doi: 10.1109/ICICI.2017.8365263

[6] M. F. Zeager, A. Sridhar, N. Fogal, S. Adams, D. E. Brown and P. A. Beling, "Adversarial learning in credit card fraud detection," 2017 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, 2017, pp. 112-116. doi: 10.1109/SIEDS.2017.7937699

[7] Ugo Fiore, Alfredo De Santis, Francesca Perla, Paolo Zanetti, Francesco Palmieri, Using generative adversarial networks for improving classification effectiveness in credit card fraud detection, Information Sciences, Volume 479, 2019, Pages 448-455, ISSN 0020- 0255, https://doi.org/10.1016/j.ins.2017.12.030.

[8] C. Wang, Y. Wang, Z. Ye, L. Yan, W. Cai and S. Pan, "Credit Card Fraud Detection Based on Whale Algorithm Optimized BP Neural Network," 2018 13th International Conference on Computer Science & Education (ICCSE), Colombo, 2018, pp. 1-4. doi: 10.1109/ICCSE.2018.8468855

[9] K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions," 2017 International Conference on Intelligent Computing and Control (I2C2), Coimbatore, 2017, pp. 1-5. doi: 10.1109/I2C2.2017.8321781

[10] F. Ghobadi and M. Rohani, "Cost sensitive modeling of credit card fraud using neural network strategy," 2016 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS), Tehran, 2016, pp. 1-5. doi: 10.1109/ICSPIS.2016.7869880

[11] A. Agrawal, S. Kumar and A. K. Mishra, "Implementation of Novel Approach for Credit Card Fraud Detection," 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, 2015, pp. 1-4.

[12] FabrizioCarcillo, Yann-Aël Le Borgne, Olivier Caelen,

YacineKessaci, FrédéricOblé, GianlucaBontempi, Combining unsupervised and supervised learning in credit card fraud detection, Information Sciences, 2019, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2019.05.042.

[13] H. Wang, P. Zhu, X. Zou and S. Qin, "An Ensemble Learning Framework for Credit Card Fraud Detection Based on Training Set Partitioning and Clustering," 2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDC