Python Machine Learning for Identifying Credit Card Fraud

JUSTIN JOHN
Master of Computer Applications.
Amal Jyothi College of Engineering.
Kanjirapally, India
justinjohn2022@mca.ajce.in

Mrs.NIMMY FRANCIS
Assistant professor in Computer Applications.
(of Affiliation)
Amal Jyothi College of Engineering.
Kanjirapally, India
nimmyfrancis@amaljyothi.ac.in

***Abstract*— Credit card payments are wont to buy the bulk of invoices and transactions created online. This can be way more convenient than carrying cash. Typically, paying cards are used in these transactions. If some other person uses your MasterCard while not your consent. Deceitful purchases are getting more and more common as more people use credit cards. The challenges connected with these types of transactions can be solved using machine learning and its algorithms. This observation is provided to show how to use system investigations to detect credit card fraud and version records. Credit card fraud can be detected by merging older transaction data with records of transactions that have been determined to be tampered with; this version is that then accustomed verifies if whether or not or not the current dealings are genuine. Our objective is to sight all dishonorable transactions 100% of the time while lowering the number of false positives. The common pattern of classification is the identification of MasterCard fraud. We tend to focus on record review and production additionally to applying various techniques to anomaly identification, like native outliers and isolated forest procedures, wherever PCAs modify credit card transaction records.**

Keywords—Machine Learning, Logistic Regression,

I. INTRODUCTION

When someone uses another person's credit card for personal advantage without notifying the cardholder or issuer, this is referred to as credit card fraud. With the rise and acceleration of e-commerce, credit cards are becoming more popular for online purchases, leading to credit card fraud. In the age of digitalization, credit card fraud needs to be identified as a prerequisite. Fraud detection involves monitoring and detecting the behavior of different users to evaluate, detect, prevent, or analyze unwanted behavior. To effectively detect credit card fraud, you need to understand the different technologies, algorithms, and types used to detect credit card fraud. Algorithms can tell the difference between illicit and lawful transactions. Analysts need access to datasets and knowledge of illicit transactions to detect fraud. They examine the records and classify all transactions. Fraud detection includes monitoring user behavior to evaluate, detect, or eliminate improper behavior such as fraud, intrusion, or failure. It uses machine learning algorithms to analyze all allowed transactions and report suspicious activity. Professional liaison experts will examine these reports to see if the transaction is genuine or fraudulent. Analysts provide feedback to automated systems used to train and update algorithms to improve fraud detection over time.

II. *MACHINE LEARNING ALGORITHMS*

Machine learning is a technique of data analysis that automates the generation of analytical models. This is linked to artificial intelligence, which is based on the idea that robots can learn from past data, recognize patterns, and make decisions with little human interaction. With the introduction of new computing technologies, the current state of machine learning is vastly different from what it was when it was first introduced. Pattern recognition was used to construct machine learning, with the notion that it could be done without being programmed for specific tasks. The repetitious part of machine learning is the most significant aspect: as data models interact with new data, they adapt independently. Previous calculations have been learned to produce reliable and repetitive results. Machine learning has recently gained new momentum. Machine learning is a study of artificial intelligence with computer science, focusing on the data and algorithms to follow the ways humans learn, improving its accuracy. The bank field is using advanced machine learning algorithms to make predictions based on past datasets. Because of the availability of many datasets related to different transaction types are available the machine learning algorithms could be used efficiently to reduce credit card fraud. Our study has identified two types of machine learning algorithms that can be utilized to predict facts –

1. Supervised learning
2. Unsupervised Learning

In Supervised Learning, an algorithm is designed to learn how to translate an input to a certain output. This is accomplished using labeled datasets accumulated throughout time. The algorithm is learned successfully if the mapping is done correctly. If the mapping is not done correctly then required alterations can be done to the algorithm to correctly learn. Trained data models created with Supervised Learning algorithms may be used to forecast fresh data that will be acquired in the future. Supervised Learning provides expertise to the algorithm, which may then be used to predict the outcomes of fresh, previously unseen data. Experience

DOI: 10.5281/zenodo.6367468
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

aids in optimizing the algorithm's performance. The 2 major varieties of supervised learning are regression and classification. An algorithmic rule that learns from labeled datasets predicts a continuous-valued output for a new dataset. It's used once the range is desired as an output. Linear regression and logistical regression are two samples of regression ways. The kind of learning within which the algorithm maps new data to one} of the 2 classes in our dataset is termed classification. The classes may be either 1 or 0, also as 'Yes' or 'No'. Classification algorithms embody call Trees, Naive Bayes Classifiers, Support Vector Machines, and others. The information in unsupervised learning is created from values with no labe1s, and also the output isn't pre-determined. On the premise of self-learning, the model makes predictions. The fundamental goal of this mode k is to predict, classify, detect, segment, and organize information. Analysis, recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics are just a few examples of machine learning applications.

a) Decision tree

One of the foremost used machine learning algorithms is that the decision tree. It's a classification model that supported supervised learning. It's easy to understand and place into action. It's a tree-like structure with an internal node for the properties of the dataset, branches for decision rules, and a leaf node for the conclusion. The selection or check is based on the characteristics of the dataset.

b) Logistic Regression

Under the machine learning approach, logistic regression is used to simulate the likelihood of a given class or event occurring, such as pass or fail. It forecasts a categorical dependent variable's outcome. 0 and 1 are the most typical results of logistic regression. In logistic regression, the sigmoid function is used to determine the likelihood of valid and invalid classes.

Sigmoid function, P $1/1+e-(a+bx)$

a) Random Forest

It's a data classification and prediction approach based on machine learning. A forest with a high number of decision trees is generated by the random forest algorithm. A huge number of trees leads to a high level of accuracy in detection. This method may be applied to any dataset. It constructs a swarm of decision trees by training and reporting the mode of the classes or the regression of discrete trees.

b) Support Vector Machine

The Support Vector Machine (SVM) is a machine learning approach for determining regression and classification issues. SVM is used to represent each information set in n-dimensional space. The hyperplane that separates 2 categories is used to classify the data.

III. Data Set Description

Features present in the dataset

1. Time

The particular time of transaction is present in seconds.

2. V1, V2...V28

Transaction details or features about the particular transaction but the credit card details are sensitive so they cannot be exposed. The dataset provider converted all the features through the principal component analysis method and converted the dataset into numerical values. And the dataset is used for the data analysis.

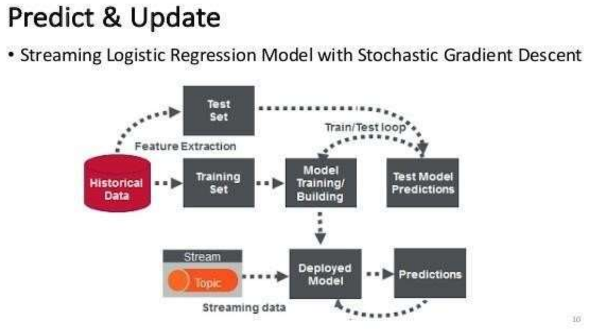
3. Amount

Amount in dataset transaction.

4. Class

The particular transaction is legit or fraudulent with label 0 or 1 Label 0 represents a normal transaction or legit transaction and 1 represents a fraudulent or fake transaction.

IV. Flow Diagram



Predict & Update

- Streaming Logistic Regression Model with Stochastic Gradient Descent

DOI: 10.5281/zenodo.6367468
ISBN: 978-93-5607-317-3@2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

V. Implementation

About dataset

Credit card transactions created by European cardholders in the Gregorian calendar month 2013 are enclosed within the dataset. we've 492 scams during this dataset, out of 284,807 transactions in the previous 2 days. info} is severely skewed, with fraud accounting for under 0.172 % of all transactions. It only has numerical input variables that have undergone a PCA transformation. we tend to be unable to present the first options and additional background information concerning the data because of confidentiality concerns. options V1, V2 … V28 is the most elements derived by PCA; the only properties that aren't affected by PCA are 'Time' and 'Amount.' The 'Time' parameter keeps track of how many seconds have passed between each transaction and the collection's initial transaction. The feature 'Amount' represents the transaction Amount and may be utilized for cost-sensitive learning based on examples. When there is a fraud, the response variable 'Class' has a value of 1 and when there isn't, it has a value of 0.

Machine-Learning-Based Methodologies

A quick overview of popular machine learning-based anomaly detection algorithms is provided here.

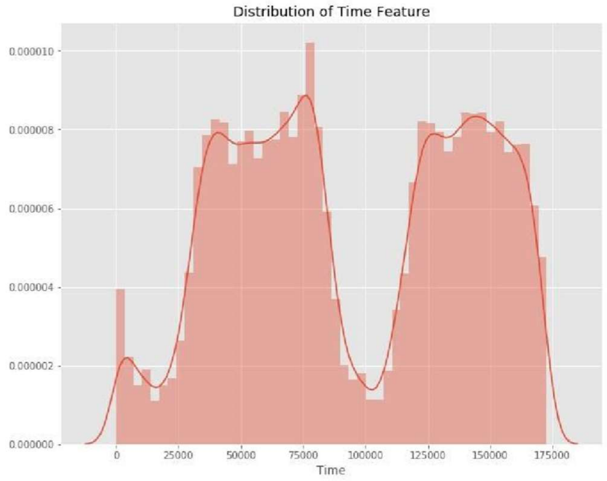
a. Anomaly Detection Using Density the k-nearest neighbor technique is used to discover anomalies based on density. Assumption: Normal data points are produced in a populated area, but anomalies are far away. Depending on the type of data, a score is used to evaluate the closest group of data points, which might be Euclidian distance or another metric (categorical or numerical). They may be divided into two types of algorithms:

K-nearest neighbor: KNN is a non-parametric lazy learning approach for classifying data based on distance metrics like Euclidian, Manhattan, Minkowski, and Hamming distance. Relative density of data: This is most commonly referred to as the local outlier factor (LOF). This concept is based on a distance metric called reachability distance b. Clustering-Based Anomaly Detection Clustering is a widely used concept in the field of unsupervised learning. Assumption: Data points with comparable distances from local centroids are more likely to belong to similar groups or clusters. The clustering algorithm K-means is commonly utilized. It generates a total of 'k' clusters of data points. Anomalies might be defined as data examples that do not fit into one of these categories.

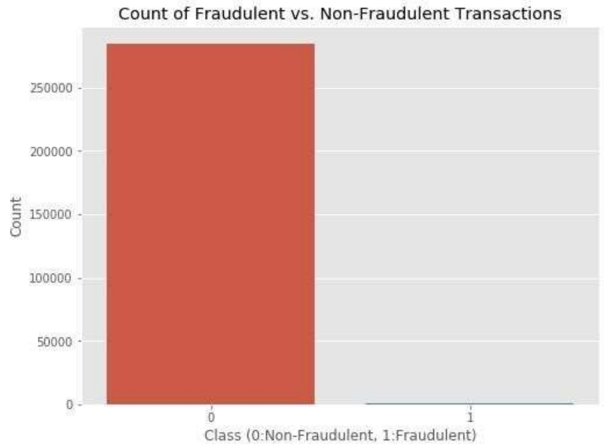
c. Support Vector Machine-Based Anomaly Detection • Another successful tool for finding abnormalities is a support vector machine.

- Although supervised learning is the most typical application of an SVM, there are also variations (such as OneClassCVM) that may be utilized for unsupervised anomaly detection (in which training data are not labeled).
- Using the training set, the system learns a soft boundary to cluster the normal data instances and then utilizes the testing set to fine-tune itself to find anomalies that fall beyond the taught region.
- Depending on the specific scenario, an anomaly detector's result might be numeric scalar values for domain-specific threshold filtering or textual labels (such as binary/multi labels). In this Jupiter notebook, we are going to take credit card fraud detection as a case study for fully comprehending this topic utilizing the Anomaly Detection Techniques listed below,
- Isolation Forest Anomaly Detection Algorithm.
- Density-Based Anomaly Detection (Local Outlier Factor) Algorithm.
- Support Vector Machine Anomaly Detection Algorithm

VI. Data Analysis



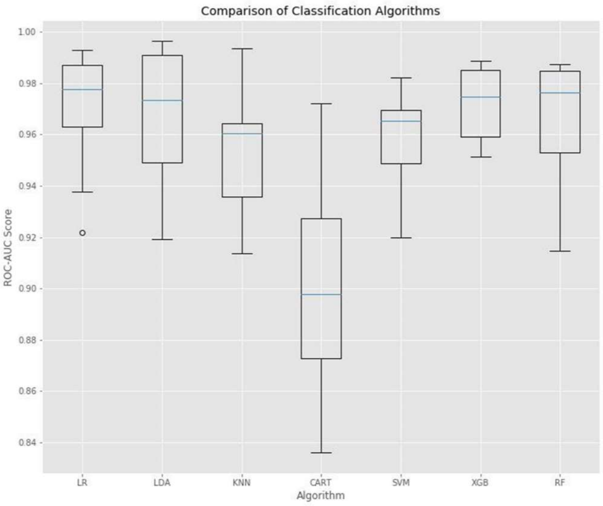
The majority of transactions are, as one might assume, nonfraudulent. In actuality, just 0.17 % of the transactions in this data set were fraudulent, making 99.83 % of them nonfraudulent. The following visualization underlines this significant contrast.



Classifications Algorithms
After that, we'll get to the part you've undoubtedly been looking forward to the most: training machine learning algorithms. To assess the efficacy of our algorithms, I split our balanced data set into two sections and did an 80/20 traintest split. I utilized the well-known resampling approach of k fold cross-validation to avoid overfitting. Simply put, you divide your training data into k parts (folds), fit your model on k-1 folds, and then make predictions for the kth hold-out fold. After that, you repeat the process for each fold, averaging the results. To get a better feeling of which algorithm would perform best on our data, some most popular classification algorithms:

- Logistic Regression
- Linear Discriminant Analysis
- K Nearest Neighbors (KNN)
- Classification Trees
- Support Vector Classifier
- Random Forest Classifier
- XGBoost Classifier

The results of this spot-checking can be visualized as follows:



VII. RESULT

Checking whether or not Credit card usage is legitimate. Because of the category imbalance ratio, we tend to advise measurement accuracy exploitation the realm underneath the Precision-Recall Curve (AUPRC). The accuracy of the confusion matrix is sorry within the case of uneven classification. The formula verifies the number of false positives and compares them to important data. This can be wont to determine the exactness and accuracy of the algorithm. The information we used for speedier testing accounted for 10% of the entire dataset. Finally, the entire dataset is utilized, and each result is presented. These findings are provided in the output, on with the classification report for every algorithm, with class zero indicating that the dealings were judged legitimate and sophistication one indicating that the transaction was found fraudulent.

VIII. CONCLUSION

Fraud detection is a complicated problem that needs extensive planning before applying machine learning techniques. Nonetheless, it is a solid use of data science and machine learning, as it ensures that the customer's money is secure and not readily tampered with. Future work will include a comprehensive tuning of the Random Forest algorithm I talked about earlier. Having a data set with nonanonymized features would make this much more interesting because displaying the feature importance would allow one to determine which individual characteristics are most essential for identifying fraudulent transactions.

IX. REFERENCES

1. L.J.P. van der Maaten and G.E. Hinton, Visualizing High-
Dimensional Data Using t-SNE (2014), Journal of Machine Learning Research
2. Machine Learning Group — ULB, Credit Card Fraud Detection (2018), Kaggle

3.ScienceDirect--
https://www.sciencedirect.com/science/article/pii/S18770 50 92030065X#:~:
text=Credit%20Card%20Fraud%20Detectio n%20using%20Deep%20Learning%20based%20on,Enco de r%20and%20Restricted%20Boltzmann%20Machine.

4. ResearchGate-
https://www.researchgate.net/publication/336800562_Cre dit _Card_Fraud_Detection_using_Machine_Learning_and_ Dat a_Science

DOI: 10.5281/zenodo.6367468