Credit Card Fraud Detection Through Anomaly Detection

Author: Rutuja Kawade

Co-Author / Corresponding Author: Jimoh Abdulganiyu

1.1 INTRODUCTION

According to V. Caronmani et.al., (2019), “A credit card fraud transaction detection system may be a method used for determining the fraudulent transactions that happen every once during a while.” This comprises primarily 2 major algorithms and uses anomaly detection as a way to classify the fraudulent transactions. The Local Outlier Factor (LoF) defines the varied parameters that need to be utilized in determining the standards for fraudulent transactions. It then checks upon the various transactions for the varied parameters present within the given LoF. This factor then gives each transaction a score supported by the varied transactions that have or will have taken place. These scores can range from 0 - 1. Each transaction is thus given a score which is predicated on the varied parameters given within the LoF. The second part of the project is isolation forest algorithms which are an algorithm that isolates the transaction which features a high rate of anomaly detected in them. Thus, these transactions are isolated then checked with various parameters to be labeled as either fraudulent or real transactions. The algorithm also uses charts to see for spikes within the average transaction. We also use a way like data visualization to point out the output in additional understandable ways which can include histograms, graphs, and matrices. Through these two algorithms and with help of knowledge visualization techniques we will detect the fraudulent transactions from the right transactions and acquire leads at pace Since these algorithms are far more time-efficient than other machine learning algorithms during this sort of task.

According to John Richard et.al.,(2017) ”An Anomaly are some things that deviate from what's normal or expected. In Mastercard transactions, most of the fraudulent cases deviate from the behavior of a traditional transaction.” It's subjective to mention what normal transaction behavior is but there are different types of anomaly detection techniques to seek out this behavior. One among the foremost common approaches to seek out fraudulent transactions was to randomly select some transactions and ask an auditor to audit it. This approach was quite inaccurate since the relation between the number of fraudulent transactions and normal transactions is on the brink of 0.1%.

According to S.P. Manirajwe et.al.,(2019) “Aim to leverage machine learning to detect and stop frauds and make fraud fighters more efficient and effective. Commonly, there's a supervised and unsupervised approach.” Mastercard companies must be ready to identify fraudulent Mastercard transactions so that customers aren't charged for items that they didn't purchase. Such problems are often tackled with Data Science and its importance, along with side Machine Learning, can't be overstated. This project intends, for instance, the modeling of a data set using machine learning with Mastercard Fraud Detection. The Mastercard Fraud Detection Problem includes modeling past Mastercard transactions with the info of those that clothed to be a fraud. This model then wants to recognize whether a new transaction is fraudulent or not. Our objective here is to detect 100% of the fraudulent transactions while minimizing the wrong fraud classifications. Mastercard Fraud Detection may be a typical sample of classification. During this process, we've focused on analyzing and pre-processing data sets also because the deployment of multiple anomaly detection algorithms like the Local Outlier Factor and Isolation Forest algorithm on the PCA transformed MasterCard Transaction data.

1.2 BACKGROUND OF STUDY

Credit Card Fraud detection can be identified using Isolation Forest Algorithm, Principal Component Analysis, and Local Outlier Factor. Here the overall concept is that the fraudulent transitions behave differently when compared to the normal transactions, thus give rise to some anomaly. For a given dataset, Isolation Forest creates an ensemble of Binary Trees. Because of their existence, anomalies have the shortest path in the trees than usual examples. With a very limited number of trees, Isolation Forest converges quickly and subsampling helps us to achieve good results while being computer-efficient. If a point is regarded centred on its surrounding neighbourhood as an outlier, it is a local outlier. An outlier given the density of the neighbourhood will be defined by LOF. When the density of the information is not the same across the dataset, LOF performs well. The background of the algorithm used involves preprocessing the data wherein the important part is data visualization. For which I have used data science libraries like numpy, pandas and seaborn. The Local Outlier Factor is based on K^{th} nearest neighbor, Reachability distance, and Local reachability density. For more precision and utilizations, devices like TensorFlow and pytorch were used.

1.3 STATEMENT OF PROBLEM

The Mastercard Fraud Detection Problem includes modeling past Mastercard transactions with the knowledge of those that clothed to be a fraud. This model is then wont to identify whether a replacement transaction is fraudulent or not. My aim here is to detect 100% of the fraudulent transactions while minimizing the wrong fraud classifications.

Consistent with V. Caronmani et.al., (2019), “Fraud detection may be a set of activities that are taken to stop money or property from being obtained through pretenses.” Fraud is often committed in several ways and lots of industries. the bulk of detection methods combine a spread of fraud detection datasets to make a connected

overview of both valid and non-valid payment data to form a choice. This decision must consider IP address, geolocation, device identification, “BIN” data, global latitude/longitude, historic transaction patterns, and therefore the actual transaction information. In practice, this suggests that merchants and issuers deploy analytically based responses that use internal and external data to use a group of business rules or analytical algorithms to detect fraud.

Credit Card Fraud Detection with Machine Learning may be a process of knowledge investigation by a knowledge Science team and therefore the development of a model that will provide the simplest leads to revealing and preventing fraudulent transactions. this is often achieved through bringing together all meaningful features of card users’ transactions, like Date, User Zone, Product Category, Amount, Provider, Client’s Behavioral Patterns, etc. the knowledge is then run through a subtly trained model that finds patterns and rules so that it can classify whether a transaction is fraudulent or is legitimate. Now you recognize what's fraud protection, let’s check out the foremost common sorts of threats. data processing Implies classifying, grouping, and segmenting of knowledge to look at many transactions to seek out patterns and detect fraud. Pattern Recognition: Implies detecting the classes, clusters, and patterns of suspicious behavior. Machine Learning here represents the selection of a model/set of models that best fit a particular business problem. For instance, the neural networks approach helps automatically identify the characteristics most frequently found in fraudulent transactions; this method is best if you've got tons of transaction samples.

1.4 MOTIVATION FOR STUDY

In today's world, it's often heard about fraud Mastercard owners, who wisely begin all the balances from the bank. I feel I can work on this to prevent these cases. Mastercard fraud costs consumers and therefore the financial company billions of dollars annually, and fraudsters continuously attempt to find new rules and tactics to

commit illegal actions. Thus, fraud detection systems became essential for banks and financial institutions, to attenuate their losses. The Importance of Developing a Fraud Detection and Prevention System. In most companies, fraud is identified only after it occurs. If they're unable to stop it during a timely fashion, however, fraud detection is the best bet for eradicating it from the environment and preventing a recurrence.

1.5 AIM AND OBJECTIVES OF THE STUDY

1.5.1 AIM

This work aims to detect credit card fraud through anomaly detection using machine learning algorithms like local outlier factor, isolation forest and principal component analysis, with the very best possible accuracy.

1.5.2 OBJECTIVES OF THE STUDY

To realize the stated aim, the subsequent objectives shall be morally followed: The model used must be simple and fast enough to detect the anomaly and classify it as a fraudulent transaction as quickly as possible.

1. The objective of these systems is to proactively detect any activity or event that is an outlier and is susceptible to fraud in the credit card translation.
2. For protecting the privacy of the user the dimensionality of the info is often reduced.
3. A more trustworthy source must be taken which double-checks the info, a minimum of for training the model.
4. We can make the model simple and interpretable so that when the scammer adapts thereto with just a few tweaks we will have a replacement model up and running to deploy.

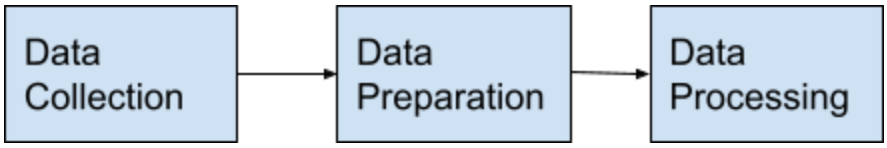
1.6 METHODOLOGY FOR THE STUDY

For successful implementation of this study, the following procedure will be followed:

i)Data Collection: The data for this study is collected from the Kaggle datasets. The URL of the dataset is: https://www.kaggle.com/mlg-ulb/creditcardfraud

ii)Data Preparation and Representation: The gathered information will be cleaned, spoken to in proper configurations, and utilized as a piece of information as needs are. The informational index is exceptionally slanted, comprising 492 fakes in an aggregate of 284,807 perceptions. This brought about just 0.172% misrepresentation cases. This slanted set is supported by the low number of deceitful exchanges. The dataset comprises mathematical qualities from the 28 'Principal Component Analysis (PCA)' changed highlights, in particular, V1 to V28. Moreover, there is no metadata about the first highlights given, so pre-examination or highlight study wasn't possible. The 'Time' and 'Sum' highlights are not changed information. There is no missing an incentive in the dataset.

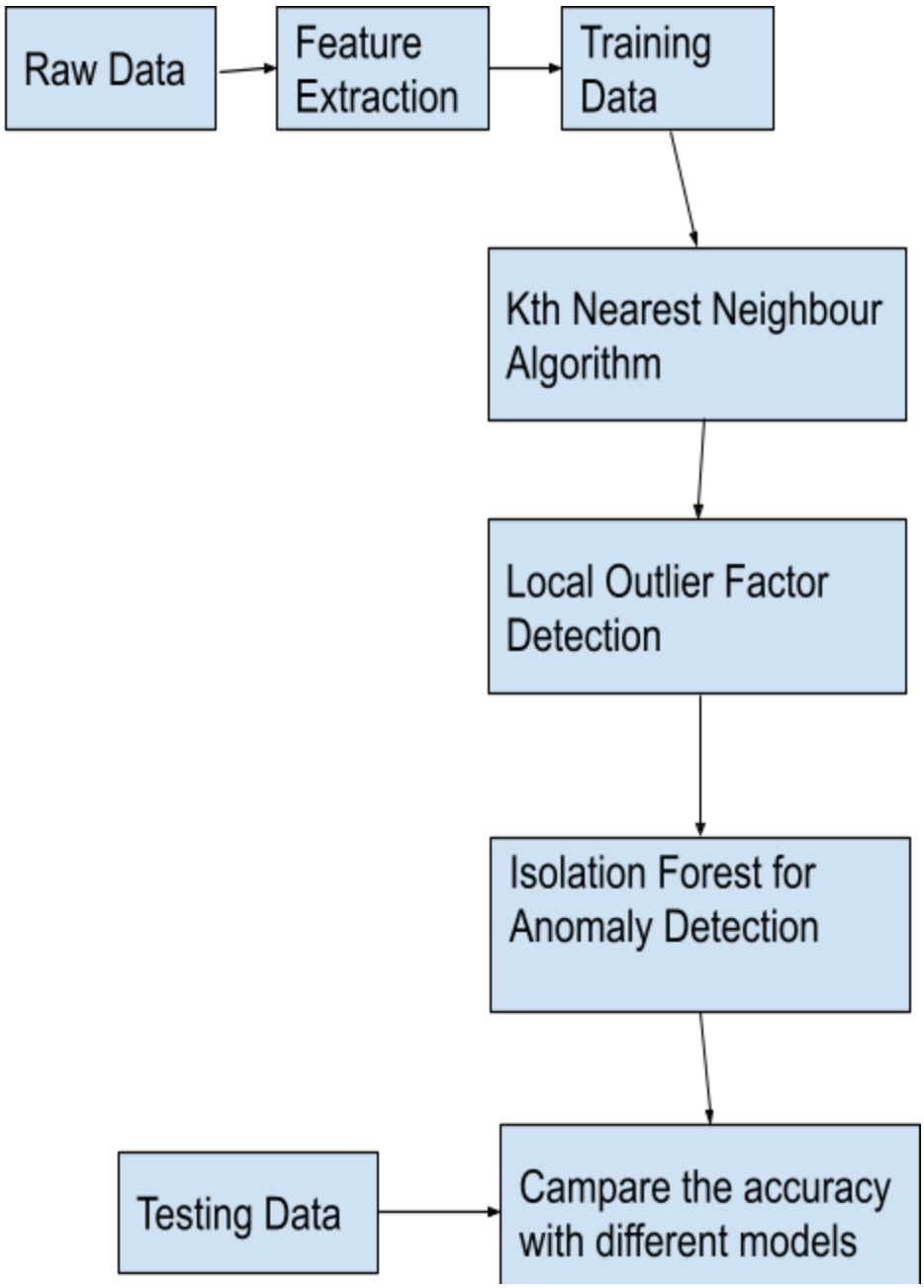
iii)Data Processing: Appropriate machine learning models like PCA, local outlier factor, Isolation Factor, along with some statistical and mathematical equations like gaussian distribution shall be used on the data to gain the most accurate results.



1.7 PROPOSED MODEL

1.7.1 Conceptual Model

The following figure represents the conceptual model for this study.



1.8 SCOPE OF THE STUDY

Anomaly detection algorithm has two stages training and testing: Training stages involve building Isolation forest, and testing stages involve passing each data point through each tree to calculate the average number of edges required to reach an external node. We begin by building several decision trees by selecting an attribute randomly. Then we choose a split value from the maximum and minimum values of that randomly selected attribute in an unpredictable way. Ideally, each terminating node of the tree contains one observation from the data set, which isolates the sample. We presume that if one finding in our data set is identical to another, it would require further random splits to isolate the finding precisely, as opposed to isolating an outlier. As we created multiple decision trees, which sum as an isolation forest, for each observation, we calculate the path length. The amount of splitting needed to distinguish the observation is equivalent to the path length from the root node to the leaf node. Then this path length is averaged over a forest of a decision tree, which serves as a scale for the anomaly and further use for determining the final anomaly score. Less the path length, the more likely it is to be anomalous.

1.9 MATHEMATICAL MODEL

Isolation Forest:

$$s(x,n) = 2^{-\frac{E(h(x))}{c(n)}}$$

where *h(x)* is the path length of observation *x*, *c(n)* is the average path length of unsuccessful search in a Binary Search Tree and *n* is the number of external nodes. Each observation is given an anomaly score and the following decision can be made on its basis:

- A score close to 1 indicates anomalies
- Score much smaller than 0.5 indicates normal observations
- If all scores are close to 0.5 then the entire sample does not seem to have clearly distinct anomalies

Local Outlier Factor Detection:
To find the reach distance between two points, we use:

$$reachdist_k(o, o') = max[dist_k(o), dist(o, o')]$$

From here, we find the local reachable distance using the following formula:

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o, o')}$$

Where *k* is a user-specified parameter that controls the smoothing effect.

1.10 SIGNIFICANCE OF THE STUDY

On a worldwide scale, Mastercard handling extortion has hit $32.320 trillion altogether, with $21.84 billion lost in the US in particular. This information represents a wide range of changes (on the web and face to face), including exchanges at POS, ATMs, and those made sure about by PINs. Here is another kicker – proactive bank cheats counteraction and recognition isn't just about making sure about your advantages and maintaining a strategic distance from misfortunes. Actualizing a strong extortion recognition framework will bring about extra advantages. Specifically, the earlier examination on misrepresentation recognition focused on utilizing private

information assortments to make measurable techniques and abusing the authority of data mining to discover disguised patterns or peculiarities in financial data. In any case, inferable from the nonappearance of generally available data, it is hard to assess and differentiate a portion of these procedures. Lundin et al. proposed a strategy for producing engineered information for checking misrepresentation location as a rule, with no particular fixation on financial data. Age gauges consistently start with data assemblage, joined by an information examination that recognizes client accounts. Finally, with the personalities determined, a definitive stage is client, assailant, and framework displaying. The system isn't so particular regarding the activity checked. We should contribute, notwithstanding, that the technique is iterative and the discoveries help survey and approve the estimation of the manufactured information created.

1.11 DEFINITION OF TERM

Extortion recognition is a lot of exercises that attempted to keep cash or property from being acquired through affectations. Misrepresentation discovery is applied to numerous businesses, for example, banking or protection. In banking, extortion may incorporate producing checks or utilizing taken Visas. Different types of extortion may include overstating misfortunes or causing a mishap with the sole purpose of the payout.

With a boundless and rising number of ways somebody can submit extortion, recognition can be hard to achieve. Exercises, for example, revamping, cutting back, moving to new data frameworks, or experiencing a network safety penetrate could debilitate an association's capacity to recognize misrepresentation. This implies methods, for example, continuous observing for cheats is suggested. Associations should search for extortion in budgetary exchanges, area, gadgets utilized, started meetings and verification frameworks. Extortion is regularly a demonstration that includes many rehashed strategies; making looking for designs an overall concentration

for misrepresentation location. This exploration paper manages Visa extortion recognition, information examiners can forestall protection misrepresentation by making calculations to distinguish examples and inconsistencies.

2.1 RELATED WORKS

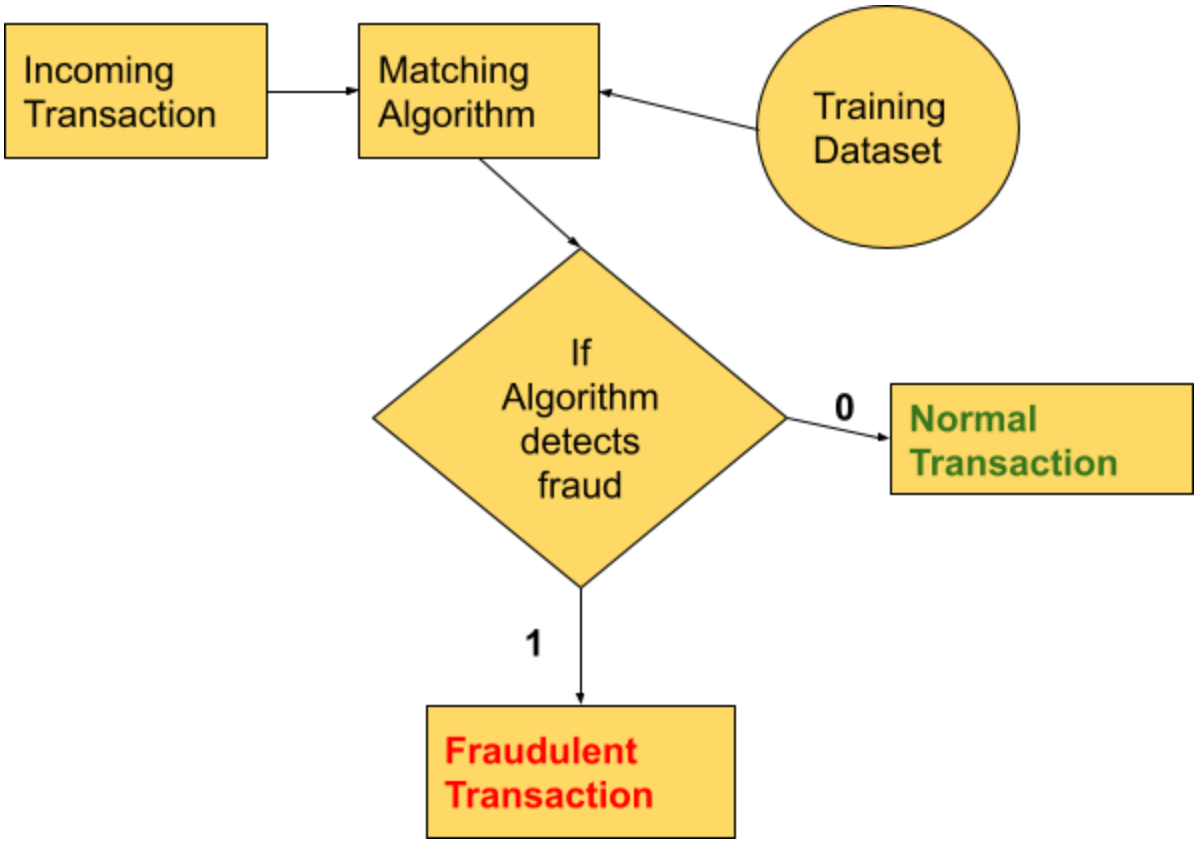
There are distinctive regulated and solo learning calculations utilized for extortion recognition in the charge card. Some significant ones are portrayed beneath. The creator [1] has proposed a paper where they have first clarified the correct exhibition estimates which are utilized for misrepresentation distinguishing proof. The creators have organized a novel learning method that can settle idea float, confirmation dormancy, and class unevenness issues. The paper likewise demonstrated the impact of the above issues in obvious Visa exchanges. Here in paper [2], creators introduced two kinds of classifiers utilizing arbitrary backwoods which are utilized to prepare the conduct highlights of exchanges. The creators have thought about the two irregular ways and have investigated their presentation on misrepresentation recognizable proof in Visa. In paper [3] creators introduced an FDS for Mastercard utilizing Artificial Neural Network and Logistic Regression. The framework used to screen every exchange independently utilizing the classifier and afterward classifier would produce the score for every exchange and name this exchange as legitimate or illicit exchange. A choice tree technique was proposed in the paper [4]. The strategy diminished generally misclassification costs and chose to part property at every hub. The creator additionally analyzed the choice tree technique for extortion recognizable proof with different models and demonstrated that this methodology performs well-utilizing execution measures like exactness and authentic positive rate. The creator [5] built up an FDS for charge card exchange utilizing support vector machines and choice trees. This investigation manufactured seven elective models that were made utilizing support vector machines and choice trees. The creator additionally analyzed this classifiers' execution utilizing execution measure exactness. The investigation additionally demonstrated that as the

size of the preparing dataset expands the quantity of extortion recognized by SVM is not as much as misrepresentation distinguished by choice tree strategy. Here in [6], the creator introduced an extortion discovery framework utilizing a Naive Bayes K-Nearest Neighbors technique. The principle point of the proposed framework was to improve precision. Gullible Bayes Classifier predicts probabilities of misrepresentation in exchange while the KNN classifier predicts how close to the unclear example information is to kth preparing dataset. The creator looked at both this classifier and indicated that both work distinctively for the given dataset. The greater part of the prescient model utilized for distinguishing extortion in Visa exchange faces the issue of idea float. The creator [7] introduced two FDS dependent on sliding window and troupe learning and indicated that classifiers should be prepared independently utilizing criticism and postponed tests. The result of the two was than amassed to improve the ready accuracy in FDS. Hence the creator demonstrated that to explain the issue of idea float, the input and deferred tests are to be taken care of independently.

3.1 METHODOLOGY

Credit card fraud is increasing considerably with the development of modern technology and the global superhighways of communication. Credit card fraud costs consumers and the financial company billions of dollars annually, and fraudsters continuously try to find new rules and tactics to commit illegal actions. Thus, fraud detection systems have become essential for banks and financial institutions, to minimize their losses. The algorithms used for this research are Local Outlier Factor (LOF), Isolation Forest,and SVM. These techniques can be used alone or in collaboration using ensemble or meta-learning techniques to build classifiers. After several trials and comparisons; we introduced the bagging classifier based on decision three, as the best classifier to construct the fraud detection model. The performance evaluation is performed on real life credit card transactions dataset to demonstrate the benefit of the bagging ensemble algorithm.

3.2 SYSTEM ARCHITECT OF THE MODEL



CHAPTER THREE

METHODOLOGY AND MODEL

3.3 SYSTEM ARCHITECT OF THE MODEL

In this model you explain in details the system architecture of the model related to your research work, which consists of:

i)Data Collection: The data for this study is collected from the Kaggle datasets. The URL of the dataset is: https://www.kaggle.com/mlg-ulb/creditcardfraud

ii)Data Preprocessing: The gathered information will be cleaned, spoken to in proper configurations, and utilized as a piece of information as needs are. The informational index is exceptionally slanted, comprising 492 fakes in an aggregate of 284,807 perceptions. This brought about just 0.172% misrepresentation cases. This slanted set is supported by the low number of deceitful exchanges. The dataset comprises mathematical qualities from the 28 'Principal Component Analysis (PCA)' changed highlights, in particular, V1 to V28. Moreover, there is no metadata about the first highlights given, so pre-examination or highlight study wasn't possible. The 'Time' and 'Sum' highlights are not changed information. There is no missing incentive in the dataset.

iii) Data Splitting: The data is split into training and testing data of which 30% data is testing data and the remaining 70% is training data.

iv) Algorithm/Approach: Appropriate machine learning models like PCA, local outlier factor, Isolation Factor, along with some statistical and mathematical equations like Gaussian distribution shall be used on the data to gain the most accurate results.

v) Prediction: The data was thoroughly fitted and the credit card fraud detection could be accurately tracked by the recall, precision, and F1 score.

vi) Model Evaluate: The models are run on a credit card fraud dataset and the accuracy of the analytical model is evaluated with help of a confusion matrix. The confusion matrix tells how the tuples in training and testing models are correctly classified. The models are evaluated based on parameters such as precession, recall, accuracy, and F1 score. The detailed description of each parameter is shown in the confusion matrix. The precession is the proportion of the predicted non-fraudulent transactions that are actually predicted as good and the recall is the proportion of actual non-fraudulent transaction that are correctly recall, as good.

Next 5 Task 9 End of Task

1. Implementation

Algorithms Used:

a) **Isolation Forest Algorithm:**

One of the new techniques to distinguish irregularities is called Isolation Forests. The calculation depends on the way that anomalies are data points that are few and different. Because of these properties, anomalies are susceptible to isolation.

Isolation is a more effective approach to identify irregularities or anolomolies than the usually utilized methods like distance measures. Additionally, this calculation works with O(1) time complexity and a little memory space is enough. It builds a decent performing model with few random forest trees utilizing little sub-trees of fixed size. This calculation functions admirably on both, small and huge datasets.

This method isolates observations by selecting a feature in a casual way and then randomly selecting a split value that ranges from the maximum and minimum values of the selected feature. This is an efficient way because isolating anomaly observations is easier as only a few conditions are required to separate those cases from the normal observations. Whereas, isolating normal observations requires more checks on conditions. Therefore, the anomaly score can be found out as the number of conditions required to separate a given observation.

The separation is done by first creating isolation trees, or random decision trees. Then, the score is calculated as the path length to isolate the observation.

b) Local Outlier Factor(LOF) Algorithm

The LOF algorithm is an unsupervised outlier detection method which computes the local density deviation of a given data point with respect to its neighbors. The outlier samples have low density than their neighbors.

The number of neighbors considered, (parameter n_neighbors) is typically chosen 1) greater than the minimum number of objects a cluster has to contain, so that other objects can be local outliers relative to this cluster, and 2) smaller than the maximum number of close by objects that can potentially be local outliers.Here I have taken n_neighbors=20 appears to work well in general.

c) Support Vector Machine(SVM)

Support Vector Machine is a supervised algorithm which can be used for both classification or regression problems. Mostly used in classification problems. Here I have used it for fraud or non-fraud transactions classification. Here, we

plot each data item as a point in n-dimensional space (where n is the number of features) where the value of each feature is the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the two classes.

2. Discussion

- Isolation Forest detected 43 errors, Local Outlier Factor detecting 57 errors and SVM detecting 3907 errors.
- Isolation Forest has a 99.74% more accurate than LOF of 99.65% and SVM of 70.09%
- When comparing error precision & recall for 3 models , the Isolation Forest performed much better than the LOF and SVM.
- Overall the Isolation Forest Method was best in determining the fraud cases.

3. Limitation

The highest accuracy obtained was from Isolation Forest but it detected the minimum frauds. The lowest accuracy was from SVM where the maximum frauds were detected, this can be considered as false negatives.

4. Recommendation/Further Studies

We can also improve on this accuracy by increasing the sample size or use deep learning algorithms however at the cost of computational expense.We can also use complex anomaly detection models to get better accuracy in determining more fraudulent cases.

5. Conclusion

After a comprehensive analysis and research about different algorithms to detect credit card fraud detection, I conclude that the Isolation Forest Algorithms proved to be most efficient.

REFERENCES:

[1] Jalinus, N., Nabawi, R. A., & Mardin, A. (2017). The Seven Steps of Project-Based Learning Model to Enhance Productive Competences of Vocational Students. In 1st International Conference on Technology and Vocational Teacher (ICTVT 2017). Atlantis Press. Advances in Social Science, Education, and Humanities research (Vol. 102, pp. 251-256).

[2] Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi and Gianluca Botempi, ‖Credit card Fraud Detection: A realistic Modeling and a Novel Learning Strategy‖, IEEE Trans. on Neural Network and Learning System, vol.29, No.8, August 2018.

[3] Shiyang Xuan, Guanjun Liu, Zhenchuan Li, Lutao Zheng,Shuo Wang, Jiang,‖Random Forest for credit card fraud detection‖, Int.conf.on Networking, Sensing and control,2018.

[4] Y. Sahin, and Duman,E.,(2011) —Detecting credit card fraud by ANN and logistic regression.‖ In Innovations in Intelligent Systems and Applications(INISTA),2011 international Symposium on (pp.315-319).IEEE

[5] Y. Sahin, S. Bulkan, and E. Duman, —A cost-sensitive decision tree approach for fraud detection,‖ Expert Syst. Appl., vol. 40, no. 15,pp. 5916–5923, 2013

[6] Sahin Y. and Duman E. (2011),‖Detecting Credit Card Fraud by Decision Trees and Support Vector Machines‖, International Multi-Conference Of Engineers and Computer Scientists(IMECS 2011),Mar 16-18,Hong Kong,Vol.1,pp.1-6

[7] Sai Kiran, Jypti Guru, Rishabh Kumar, Naveen Kumar, Deepak Katariya,‖Credit card fraud detection using Naïve Bayes model based and KNN classifier‖, Int. Journal of Adv. Research , Ideas and Innovations in Technology,vol.4,2018.

[8] V. Ceronmani Sharmila, Kiran Kumar R., Sundaram R., Samyuktha D., Harish R. 2019. Credit Card Fraud Detection Using Anomaly Techniques: IEEE Xplore 2019 1st

International Conference on Innovations in Information and Communication Technology (ICIICT), 2019.

[9] S.P. Maniraj, Aditya Saini, Swarna Deep Sarkar, Credit Card Fraud Detection using Machine Learning and Data Science:in International Journal of Engineering and Technical Research 08(09), September 2019.

[10] John Richard D. Kho, Larry A. Vea: Credit Card Fraud Detection Based on the Transaction Behavior published by Proc. of the 2017 IEEE Region 10 Conference (TENCON), Malaysia, November 5-8, 2017.