**Credit Card Fraud Detection**

Hetanshi Kothari

Department of Information Technology

Devang Patel Institute of Advance Technology and Research

Changa,India

18dit029@charusat.edu.in

Tanushree Kurup

Department of Information Technology

Devang Patel Institute of Advance Technology and Research

Changa,India

18dit030@charusat.edu.in

Panth Shah

Department of Information

Technology

Devang Patel Institute of Advance Technology and Research

Changa,India

18dit068@charusat.edu.in

**Abstract:** **As the age of digitization reaches our day-to-day lives, people are becoming more interested in online banking, online shopping, and online transaction systems, the number of fraudulent credit card transactions has risen. In our minds, the word "fraud" still conjures up images of credit card fraud. And, as a result of the massive increase in credit card purchases in recent years, credit card fraud has risen dramatically. As a result, fraud detection should include surveillance of the person's/spending customer's attitude in order to determine, prevent, and detect unwanted behavior. Since credit cards are the most common method of payment for both online and offline purchases. Credit card fraud is on the rise. Fraud detection is concerned not only with capturing fraudulent activities, but also with discovering them as quickly as possible, since fraud costs millions of dollars in business losses and is increasing over time, having a significant impact on the global economy. In the detection of credit card fraud, machine learning algorithms have played a significant role. Traditional classification algorithms.do not perform well in detection of credit card fraud due to the unbalanced nature of real-life datasets. We use the Isolation Forest Algorithm in the proposed scheme. The Local Outlier Factor is used to identify fraudulent transactions and their accuracy.**

**Keywords— Credit Card fraud. Detection, Machine learning, Imbalanced. Data**

**INTRODUCTION**

The word credit card fraud refers to any act of theft or fraud involving a payment card that occurs as a result of the owner's physical loss of the card, or the card being stolen by fraudsters, or as a result of techniques such as phishing, skimmer, identity theft, and so on. This form of financial fraud can have a significant impact on a country's business, organizational, and government sectors. In today's age of internet technology, where credit card transactions have become the most convenient method of transaction, whether online or offline, the rate of fraudulent transactions has increased. Internal and external fraud are the two forms of credit card fraud that may occur, as previously mentioned. The first type depicts a scenario in which information is leaked between the cardholder and the bank by the use of a false identity, while the second type depicts fraud that occurs when a stolen or missing credit card falls into the hands of fraudsters. Credit. Card fraud detection strategies have been developed by researchers over the years to solve this problem. As a result, effective credit card fraud has become a reality. Prior to applying any ML techniques, successful pre-processing of the dataset is needed. The unbalanced existence of credit card data has been considered in this paper, as well as an isolation forest and. Outlier factor is a local factor. The most significant benefit of using this strategy for credit card fraud detection is that it can easily deal with. the massive size of the training.data, Online purchases have become an integral and essential part of our lives in recent years. It's important for credit card companies to be able to spot fraudulent credit card purchases so that consumers aren't charged for things they didn't buy. The number of fraudulent transactions is increasingly growing as the volume of transactions increases. Machine.Learning and its algorithms can be used to solve such problems. The aim of this project is to demonstrate how to model a data set using machine learning. Detection of Credit Card Fraud. Modeling past credit card purchases with the data of those that turned out to be fraud is part of the Credit Card Fraud Detection Issue. The model is then used to determine whether or not a new transaction is fraudulent. Our goal is to identify 100% of fraudulent transactions while reducing the number of incorrect fraud classifications. The detection of credit card fraud is a standard example of classification. We've concentrated on analysing and pre-processing data sets, as well as deploying multiple anomaly detection algorithms including  Isolation forest algorithm and the Local Outlier Factor . Credit Card.Transaction data was transformed by PCA.

1. **RELATED WORK**

A significant number of.research works have been done for credit card fraud detection. The techniques developed can be categorized into two sections, as discussed below:

*A*.*Machine Learning based techniques*

Techniques for detecting credit card fraud have been presented. A list of challenges that one might face is summarised in the report. During the identification of credit card fraud. The method proposed in this paper employs the most up-to-date machine learning algorithms to detect outliers, or unusual behaviors. The full architecture diagram can be represented as follows when viewed in detail on a larger scale along with real life elements: To begin, we got our dataset from.Kaggle, a data.analysis website that provides datasets. There are 31 columns in this dataset, with 28 of them labelled as v1-v28 to protect confidential data. Time, Number, and an are represented in the other columns. It's all about the class. The time difference between the first and subsequent transactions is shown by time. Sum refers to the total amount of money exchanged. A legitimate transaction is represented by class 0, while a fraudulent transaction is represented by class 1. To search for anomalies in the dataset and to visually comprehend it, we plot different graphs: According to our results, the number of fraudulent transactions is significantly lower than the number of legitimate transactions. It can be seen that the lowest number of transactions were made at night and the largest number were made during the day. The majority of transactions are small, and only a few come close to exceeding the maximum transaction limit. We plot a histogram for each column after reviewing this dataset. This is done to provide a graphical representation of the dataset that can be used to ensure that no values are missing from the dataset. This is done to ensure that no missing value imputation is required and that the machine learning algorithms can process the dataset efficiently We plot a heatmap after this analysis to get a colored representation of the data and to investigate the association between our predicting variables and the class variable. The dataset has been formatted and analyzed at this stage. The number and time columns are also standardized. To ensure that the evaluation is equal, the Class column is omitted. A series of algorithms from modules processes the data. The following outlier detection modules are added to this data after it has been fitted into a model: The Isolation Forest Algorithm and the Local Outlier Factor These algorithms are part of the sklearn library. Ensemble-based.methods and functions for the classification, regression, and outlier detection are included in the sklearn package's ensemble module. The NumPy, SciPy, and Matplotlib modules were used to build this free and open-source Python library. which provides a number of simple and effective tools for data analysis and machine learning. It comes with a variety of grouping, clustering, and regression algorithms and is designed to work with numerical and scientific libraries.

•  Point Anomalies:  A point anomaly occurs when one object may be detected as an anomaly as compared to other objects. This is the most basic anomaly, and it's seen in a lot of studies.

• Contextual Anomalies :If an object is anomalous in a certain context. Only in this circumstance. It's a contextual anomaly.

•Collective  Anomalies:If some linked objects can be observed as an anomaly against other objects. Only a set of objects can be anomalous in this situation, not a single entity.

Anomaly detection can be carried out using a variety of methods, including:Supervised Anomaly Detection. A setup in which the data is labelled in training and test data sets, allowing for the training and application of a simple classifier. This case is similar to conventional pattern recognition, with the exception of groups, which are usually highly unbalanced. Not all classification methods are appropriate for this task. Some decision trees, for example, struggle to deal with unbalanced data. Artificial Neural Networks (ANN) or Vector Machines (SVM) can perform better. This configuration, however, is irrelevant since we need to be aware of all irregularities and correctly mark data. Anomalies aren't always established ahead of time in certain situations. or as a result of novelties discovered during the research process. Semi Supervised Anomaly.Detection is a term used to describe the detection of anomalies. We gather knowledge from. training results at first when. we do not have any knowledge at all. This setup often employs training and evaluation datasets, with the training data consisting solely of standard data free of anomalies. The concept is that a model of the average class has already been taught, and deviations can be identified by departing from that model. This method of classification is also known as "one-class" classification. There are well-known methods. SVMs and autoencoders of one class. In general, any density estimation method, such as Gaussian Mixture approaches or Kernel Density Estimation, can be used to model the probability density function of the normal groups.

Unsupervised Anomaly Detection A setup in which we don't know what is standard and what isn't in the data. It is the most adaptable configuration that doesn't require labels. In addition, there is no distinction between a training and a test dataset. Unsupervised anomaly detection approaches score data solely based on natural features of the dataset, according to the definition. Distances or densities are often used to determine what is natural and what is an outlier.

*B. Isolation Forest*

The isolation forest is an unsupervised algorithm for anomaly detection that operates on the concept of isolating anomalies. It directly isolates anomalous points in the dataset rather than attempting to construct a model of normal instances. It's a fast algorithm with a small memory footprint. It locates anomalies in data by isolating outliers. The isolation forest occurs as a result of an unsupervised machine learning algorithm. Isolation forest is based on the decision tree algorithm's theorem. The algorithm in a decision tree for predicting the class of a given dataset begins from the root node of the tree or the next node, the algorithm. Again compares.the.attribute value with the other sub-nodes and movesfurther. It repeats the loop until it reaches the tree's leaf node. The recursion principle underpins the operations of Isolation forest. By randomly selecting a feature and then randomly selecting a split value for the feature, this algorithm recursively creates partitions on datasets. In comparison to the so-called normal data points in the dataset, the anomalies can require less random partitions to be isolated. As a result, the. anomalies will be. the points that have a short life span.

*C.Local Outlier Factor*

The unsupervised outlier detection tool is the local outlier factor. It calculates each sample's anomaly ranking. It calculates the local density variance of a given sample in relation to its neighbors. The anomaly score is determined by how different the sample is from the surrounding neighborhood. The Local Outlier Factor may also be considered for unsupervised outlier identification. It generates an anomaly score to reflect data points that are considered outliers in the data. This. is done by measuring the local density difference of a datapoint relative to the data points near it. Estimating distances between data points that are neighbors determines local density (k-nearest neighbors). As a result, local density can be determined for each data point. We can see which data points have similar densities and which have a lower density than its neighbors by comparing them. The ones with the lowest densities are referred to as outliers. To begin with,.k-distances are distances between points.that are calculated for each point in order to de termine their  k-nearest neighbors. The. second closest point to\ the point is said to be the second closest neighbor to the point. Here's a picture that shows the k-distances between different neighbours in a point cluster

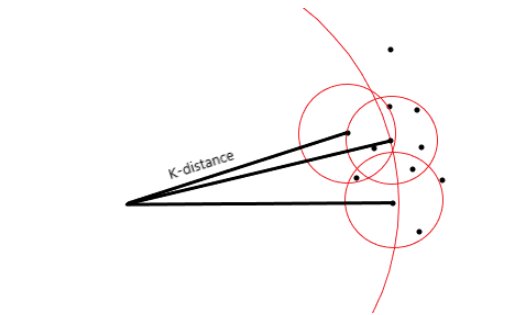
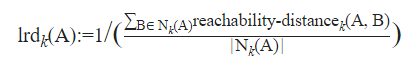


Fig 1.1

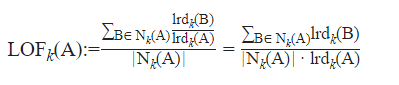
The reachability distance is calculated using this distance. It is defined as the sum of the distance between two points and that point's k-distance. Consider the following equation, in which B is the centre point and A is a point near it.

https://media.geeksforgeeks.org/wp-content/uploads/20200810072132/Screenshot342.png

reachability distances to any or all of a degree area unit's k-nearest neighbours measured to determine the time's native Reachability Density (LRD). The native reachability density can be determined by taking the inverse of the sum of all the reachability distances of all the k-nearest neighbouring points around a degree. The greater the distance between the lines, the smaller the difference, and thus the higher the density, so the opposite is used in the equation.



The quantitative relation of the common of the lrds of k set of neighbours of a degree and also the lrd of that time is used to calculate the native outlier problem (LOR). The LOR equation is as follows:



Identifying outliers can be difficult at times. An outlier is a point that is a short distance from a dense cluster, while an inlier is a point that is a longer distance from a broader spaced cluster. Outliers in local areas are identified using LOR, so this problem is no longer an issue. The method used in LOF can be used to solve problems of detecting outliers in a variety of fields, including geographic data, video streams, and so on. A different dissimilarity function can also be implemented using the LOF. It also outperforms a number of other anamoly detection algorithms.

**II.PROCESS**

We are completing this fraudulent transaction detection activity in following three phases:

1. Data Exploration

Steps: a) Load dataset

b) Preprocess dataset

c) Perform graphing

d) Display dataset

2) Data Preprocessing

Steps:

a) Load dataset

b) Remove Null values

c) Split dataset

d) Move to training phase

3) Data Classification

Steps: a) Train the dataset

b) Develop classifier

c) Isolation Forest

d) Perform Classification

The first phase, also known as the Data discovery phase, entails loading the dataset. Data exploration, including data analysis, is a method where we use visual exploration to understand what is in the dataset and what its characteristics are. We used a data set from the Kaggle website that contained a variety of parameters that were reduced using the PCA dimensionality reduction method, including number, class, time, and others. The dataset is explored and represented to produce descriptive statistics for the given series object that summaries the central tendency, dispersion, and shape of the dataset's distribution. Null values are not used in any of the calculations. The histogram is created to display this represented information as a result of victimization. Data Preprocessing is the second step. To increase the dataset's efficiency, it reloads it and removes all of the null values and garbage values. We must divide the dataset into training and testing phases during this process. We mainly work on the training process here, defining class 0 transactions as genuine and class 1 transactions as fake. In order to improve the quality of the data, false and legitimate entries are given at random during the dataset training process. As a result, more realistic data is obtained. A correlation matrix is given to summaries data, to be used as an input into a more advanced analysis, and to be used as a diagnostic for advanced analysis. Data classification is the third and final phase. It's just a matter of inputting a training data set of pre-labeled classes for the algorithm to learn from. The model is then applied to a new dataset in which the classes are not specified, and the model predicts the class to which it is similar using the training set's learning. Both algorithms must be used to produce a useful result, with words like accuracy, recall, f1-score, and help being used to determine the outcome.

**III.EXPERIMENTAL SETUP**

Since accuracy is an easy measure to apply and generalises to more than just binary labels, many classification tasks use simple assessment metrics like Accuracy to compare performance between models. However, there is one significant flaw in precision. It is presumed that each class has an equal representation of instances, and consistency is a misleading factor for distorted datasets like ours. It does not provide precise results. As a result, in our case, accuracy isn't a good indicator of performance. To identify transactions as fraud or non-fraud, we'll need some additional. Correctness criteria, such as: • Precision • Recall • F1-score • Support All of these correctness standards are focused on the Actual and Predict. classes. One of the hottest topics in data mining is outlier detection, also known as anomaly detection. Isolation Forest(iForest) and Local Outlier Factor (LOF) are two well-known outlier detection algorithms. iForest, on the other hand, is only responsive to global outliers and is ineffective at coping with local outliers. While LOF is effective at detecting local outliers, it has a high time complexity. To identify transactions as fraud or non-fraud, we'll need some additional.correctness criteria, such as: • Precision • Recall • F1-score • Support All of these correctness standards are focused on the Actual and Predict. classes. One of the hottest topics in data mining is outlier detection, also known as anomaly detection. Isolation Forest(iForest) and Local Outlier Factor (LOF) are two well-known outlier detection algorithms. iForest, on the other hand, is only responsive to global outliers and is ineffective at coping with local outliers. While LOF is effective at detecting local outliers, it has a high time complexity. While LOF is effective at detecting local outliers, it has a high time complexity. A two-layer progressive ensemble approach for outlier detection is proposed to solve the shortcomings of iForest and LOF. It has a low time complexity and can reliably detect outliers in complex datasets. This approach uses a low-complexity version of iForest to quickly search the dataset, prune the data that appears to be regular, and generate an outlier candidate collection. The outlier coefficient is used to design a pruning threshold setting method that is based on the outlier degree of data, in order to improve pruning accuracy even further.

**IV.RESULTS**

Here we calculate the mean, count, max and other information of the data.

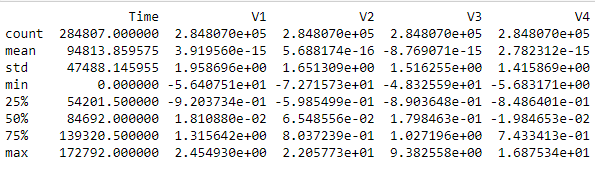


Fig 4.1

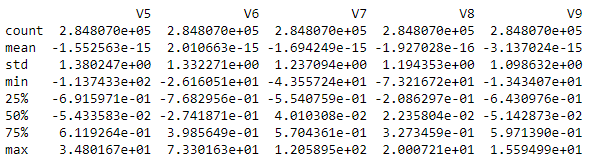


Fig 4.2

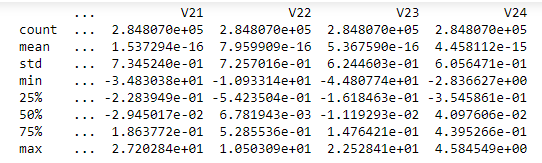


Fig 4.3

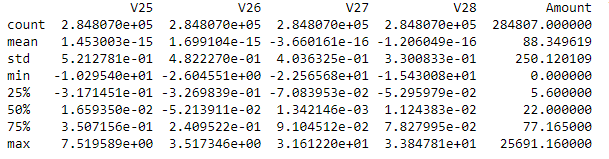


Fig 4.4

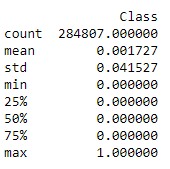


Fig 4.5

The project employs histograms to make it easier to distinguish between fraudulent and legitimate transactions. For this, we can use the matplotlib package. We can also adjust the plot's size accordingly.

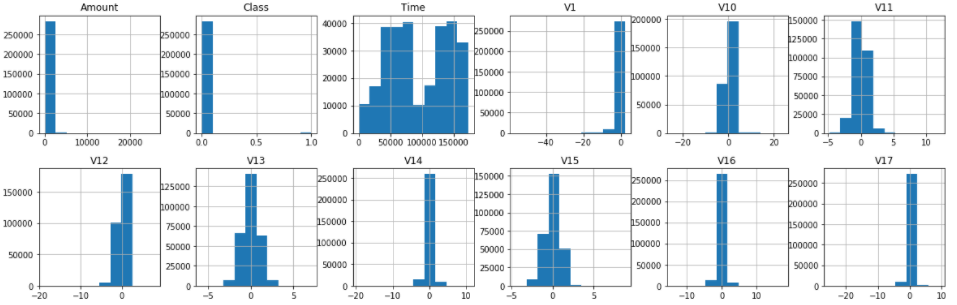


Fig 4.6

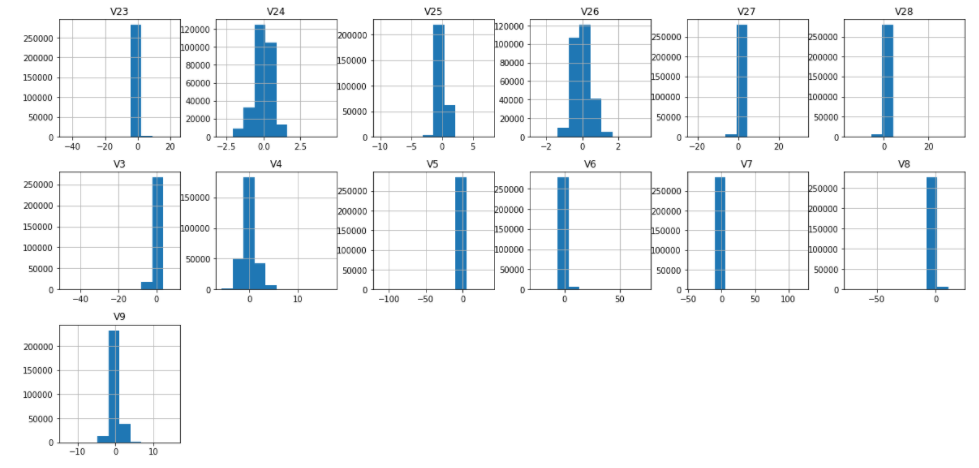
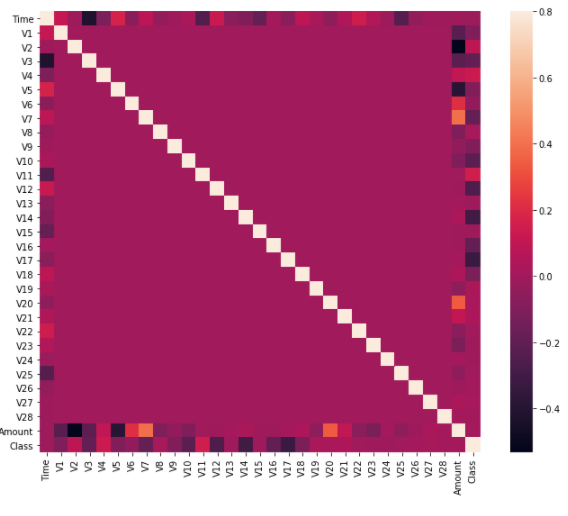


Fig 4.7

The above output shows that it created the bar chart for every attribute within the dataset. Histograms cluster the info in bins and is that the quickest way to get plan regarding the distribution of every attribute in dataset.

The correlation matrix is a heat map that is used to see whether there is a relationship between various parameters and variables in our dataset.



The above graph was generated with pyplot and uses seaborn and SNS heatmap. It gives our basic correlation matrix a visual appearance and makes analysis easier. At both X-Y axes, with a range of -0.75 to +0.50, all 31 parameters V1-V29, class, and number are present.

**V.CONCLUSION AND SCOPE**

Local outlier and isolation forest techniques are suggested in this paper for successful detection of credit card fraud. This research looked into the unbalanced existence of credit card data. The experimental findings demonstrated the effectiveness of the proposed model in effectively treating imbalanced cases in the detection of credit card fraud. With the increased use of credit cards for purchases, the chances of credit card fraud are skyrocketing. This paper presents an analysis of credit card fraud detection on a publicly accessible dataset using Machine Learning algorithms such as Local outlier factor and Isolation Forest. The proposed framework is written in the Python programming language. For high-dimensional datasets, the proposed model has not been validated. To be applied in high-dimensional datasets, the proposed model can be expanded by integrating with different data.cleaning techniques such as sampling or feature.selection (or extraction) algorithms. This paper does not look into the unbalanced nature of detection methods and their impact on results. Another area of potential research is a thorough examination of the problem in relation to the computational efficiency of credit.card fraud detection techniques. The code prints the number of false positives it found and compares it to the real numbers. This is used to measure the algorithm's accuracy score and precision. The percentage of data we used for faster testing was 10% of the total dataset. At the end, the whole dataset is used, and all reports are written..These results, as well as the classification report for each algorithm, are presented in the output, where class 0 indicates that the transaction was determined to be legitimate and class 1 indicates that the transaction was determined to be fraudulent. To rule out false positives, this result was compared to the class values. Credit card fraud is unquestionably a form of criminal deception. This article summarised recent studies in this field and identified the most common types of fraud, as well as how to identify them. This paper also includes a detailed description of how machine learning can be used to improve fraud detection results, as well as the algorithm, pseudocode, explanation, and experimentation results. Since the whole dataset is made up of just two days' worth of transaction information, it's just a small portion of the data that could be made accessible if this project were to be used commercially. Since the software is based on machine learning algorithms, it can only get more efficient over time as more data is fed into it. Although we didn't achieve our target of 100 percent accuracy in fraud detection, we did create a system that can get very close to it given enough time and data. As with every project of this nature, there is space for improvement. Because of the design of this project, several algorithms can be combined as modules and their results combined to improve the accuracy of the final result. More algorithms can be added to this model to boost it even more. These algorithms' performance, however, must be in the same format as the others. The modules are simple to add once that condition is met, as shown in the code. The project gains a lot of modularity and flexibility as a result of this. The dataset contains more space for change. As previously shown, the accuracy of the algorithms improves as the dataset size grows. As a result, more data will undoubtedly improve the model's accuracy in detecting frauds and reduce the number of false positives.

**VI.ACKNOWLEDGEMENT**

We'd like to take this opportunity to express our heartfelt appreciation and warm regards to our advisor for their outstanding guidance, monitoring, and relentless encouragement during the thesis. The blessings, assistance, and encouragement that they provide from time to time will take us a long way in the life path that we are about to embark on. Every successful project is built on the continuous motivation, goodwill, and support of those who surround it. We would like to take this opportunity to express our gratitude to the many people who have contributed their valuable time, full support, and cooperation to the project's growth. We are grateful to our Principal, Dr. Amit Ganatra, Head of the Department, Dr. Amit Nayak, and Project Guide, Prof. Mohammed Bohra, for their assistance during the study and development period. It is because of them that we have been motivated to work hard and implement new technologies. They created a favourable atmosphere for us, and without them, we would not have been able to achieve our target.

**VII. REFERENCES**

[1]https://www.sciencedirect.com/science/article/pii/S187705092030065X

[2]<https://www.researchgate.net/publication/336800562_Credit_Card_Fraud_Detection_using_Machine_Learning_and_Data_Science1>

[3]<https://www.irjet.net/archives/V7/i3/IRJET-V7I3710.pdf>

[4]<https://www.acadpubl.eu/jsi/2018-118-7-9/articles/7/31.pdf>

[5] <https://ieeexplore.ieee.org/document/8741421>

[6] <https://ieeexplore.ieee.org/document/8123782>

[7] <https://dl.acm.org/doi/abs/10.1145/3338840.3355641>

[8] Altyeb Altaher Taha, Sharaf Jameel Malebary.

"An Intelligent Approach to Credit Card Fraud

Detection Using an Optimized Light Gradient

Boosting Machine", IEEE Access, 2020

[9] Leila Seyedhossein, Mahmoud Reza Hashemi.

"Mining information from credit card time

series for timelier fraud detection" 2010 5th

[10] Smita Agrawal, Atul Patel. "SAG Cluster: An

unsupervised graph clustering based on

collaborative similarity for community

detection in complex networks".