Credit Card Fraud Detection

Project report submitted in partial fulfilment of the requirement for the degree of Bachelor of Technology in

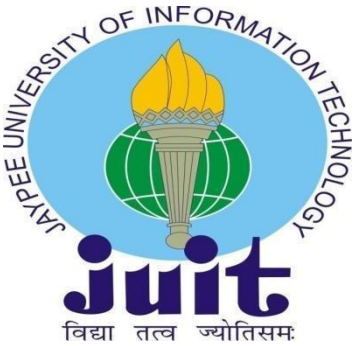
Computer Science and Engineering

By

Kautilya Sharma 181218

UNDER THE SUPERVISION OF

Dr. Jagpreet Sidhu



Department of Computer Science & Engineering and Information Technology

Jaypee University of Information Technology, Waknaghat, 173234, Himachal Pradesh, INDIA

TABLE OF CONTENT

I

DECLARATION

I hereby declare that this project has been done by me under the supervision of (Dr. Jagpreet Sidhu, Assistant Professor S.G.) the Jaypee University of Information Technology. I also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:
Dr. Jagpreet Sidhu
Assistant Professor S.G.
Department of Computer Science & Engineering and Information Technology
Jaypee University of Information Technology

Submitted by:
Kautilya Sharma 181218
Computer Science & Engineering Department
Jaypee University of Information Technology

II

CERTIFICATE

This is to certify that the work which is being presented in the project report titled **"**Credit Card Fraud Detection" is in partial fulfilment of the requirements for the award of the degree of B.Tech in Computer Science And Engineering and submitted to the Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat is an authentic record of work carried out by Kautilya Sharma during the period from January 2022 to May 2022 under the supervision of Dr. Jagpreet Sidhu, Department of Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat.

Kautilya Sharma 181218

The above statement made is correct to the best of my knowledge.

Dr. Jagpreet Sidhu
Assistant Professor S.G.
Computer Science & Engineering and Information Technology
Jaypee University of Information Technology, Waknaghat

Ⅲ

ACKNOWLEDGEMENT

Firstly, I express my heartiest thanks and gratefulness to Almighty God for his divine blessing that makes it possible to complete the project work successfully.

I am grateful and wish my profound indebtedness to supervisor Dr. Jagpreet Sidhu, Assistant Professor S.G., Department of CSE Jaypee University of Information Technology, Wakhnaghat. Deep Knowledge & keen interest of my supervisor in the field of Machine Learning helped me to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

I would like to express my heartiest gratitude to Dr. Jagpreet Sidhu**,** Department of CSE, for his kind help to finish my project.

I would also generously welcome each of those individuals who have helped me straightforwardly or in a roundabout way to make this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking.

Finally, I must acknowledge with due respect the constant support and patients of my parents.

Kautilya Sharma

IV

ABSTRACT

“Financial fraud is becoming more prevalent in the financial business, with far-reaching implications. Data mining was required to detect credit card fraud in internet transactions. To detect Credit Card Frauds, it is a data mining task and very hard coz of two fundamental issues: firstly, because the portraits of legit and fraudulent behaviour vary often, and secondly, the datasets of credit card frauds are highly skewed. The sampling strategy used in the dataset, the variables chosen, and the technique(s) used for detection purposes all have an influence on the accuracy in detecting credit card frauds. The performances of nave bayes, KNN, and the logistic regression on significantly skewed data of credit card frauds is determined in this paper. A credit card transaction dataset of 284,807 transactions was given by European cardholders. A hybrid method of under-sampling and oversampling is utilised to work with the skewed data. The 3 techniques are forced upon the unprocessed and preprocessed data. “Fraud detection is a fixed of activities that aretaken to prevent cash or belongings from being acquired thru fake pretenses.”Fraud may be committed in exceptional ways and in lots of industries. Credit card frauds are clean and friendly objectives. E-commerce and plenty of different on-line websites have extended the net price modes,
increasing the chance for on line frauds. Increase in fraud quotes, researchers started out using distinct device gaining knowledge of strategies to detect and examine frauds in on line transactions. Credit card fraud commonly happens when the card become stolen for any of the unauthorized functions or maybe whilst the fraudster makes use of the credit card facts for his use. Lots of mony are lost due to credit card fraud every yr. There is a lack of studies studies on reading real-international credit score card data thanks to confidentiality problems. In this paper, device getting to know algorithms are used to locate credit card fraud. To evaluate the model efficacy, a publicly to be had credit card facts set is used. The System prediction degree & accuracy of fraud detection isn't one hundred precent correct ,So there's a risk of having fraud. Then, a real-international credit card facts set from a monetary group is analyzed. In addition, noise is brought to the records samples to in addition verify the robustness of the algorithms. The experimental outcomes positively imply that the majority voting method achieves excellent accuracy charges in
detecting fraud cases in credit cards.”

Chapter 01:- INTRODUCTION

1.1 Introduction

“Fraud is defined as the intentional deception or concealment of a substantial truth with the purpose of influencing someone to act on it to their harm (as defined by the American Institute of Certified Public Accountants).Fraud is something you've either witnessed firsthand or heard about from a friend. You've probably heard stories about people having their identities stolen. That's a ruse of some type.
People who steal other people's identities use their fictitious identities to make transactions using the personal information of victim, such as their bank account and credit card. Many victims of identity theft have lost tens of thousands of dollars in addition to their excellent credit ratings.
Fraud is a severe offence that must not be taken lightly. Fraud is a punishable offence in most regions, with varying degrees of punishment.”

Despite the fact that there are just two main forms of fraud, there are several subtypes. Here's a rundown of the most common:

- Fake invoices, for example, are one example of mail fraud.
- Insurance fraud, such as claiming more than is required, is a serious problem.
- Tax evasion, such as failing to record your income accurately.
- Check fraud is when someone writes a bogus check.
- Fraudulent transactions on the internet, such as the sale of counterfeit goods
- Fake websites are a type of website misdirection.
- Charity fraud happens when charities fail to transfer funds to the people they claim to assist.

- Pyramid schemes, such as purchasing a money-making package only to discover that you must now offer the same package to others.

- Scams involving work-from-home opportunities, such as adverts that encourage you to pay for further information.

- Credit Card Fraud means using a person's credit card without his knowledge by means of withdrawing funds or purchase of goods.

The most prevalent methods of card theft include stealing the card before it is received by the owner, obtaining card details from the owner via phone calls, sending inappropriate links to the owner's mobile phone in order to obtain card details, and appropriating missing cards.

Because today almost every transaction could be done online using only the credit card info, this has become a major problem in the current day.
In 2017, 16.7 million individuals died as a result of illegal card operations.

Furthermore, credit card fraud claims climbed by 40% in 2017 compared to the previous year, according to the Federal Trade Commission (FTC). California had almost 13,000 instances and Florida with 8k , the two states having the highest per capita rates of this type of crime.

The most prevalent methods of card theft include stealing the card before it is received by the owner, obtaining card details from the owner via phone calls, sending inappropriate links to the owner's mobile phone in order to obtain card details, and appropriating missing cards.

The following are the primary concerns that we have:

- Is it simple to track down the individual and his current location?
- Is it feasible for the owners to receive a refund?
- How are ordinary people affected by technological advancements?

“A credit score card is a skinny accessible plastic card that carriesidentification data which include a signature or photo,and authorizes the character named on it to rate purchases or offerings to his account - prices for which he will be billed periodically. Today, the records on the cardboard is read through automatic teller machines (ATMs), keep readers, financial institution and is likewise utilized in on line net banking device.
They have a unique card variety that is of maximum significance. Its safety is based on the physical protection of the plastic card as well as the privacy of the credit score card wide variety. There is a fast boom within the number of credit score card transactions which has brought about a sizeable upward thrust in fraudulent sports. Credit card fraud is a wide-ranging time period for theft and fraud committed the usage of a credit score card as a fraudulent supply of finances in a given transaction. Generally, the statistical techniques and many information mining algorithms are used to remedy this fraud detection hassle. Most of the credit card fraud detection structures are based on artificial intelligence, Meta learning and pattern matching. The Genetic algorithms are evolutionary algorithms which aim to reap the better answers in eliminating the fraud. A excessive significance is given to broaden green and relaxed electronic payment gadget to stumble on whether a transaction is fraudulent or not. In this paper, we will awareness on credit card fraud and its detection measures.A credit card fraud takes place when one individual uses other individuals’ card for their non-public use without the know-how of its proprietor. When such kind of instances takes vicinity by using fraudsters, it is used until its complete to be had limit is depleted. Thus, we want a solution which minimizes the full to be had restriction on the credit score card that's greater outstanding to frauds. And, a Genetic set of rules generates better solutions as time progresses. The entire emphasis is given on developing efficient and at ease digital fee system for detecting the fraudulent”

How Do Fraudsters Obtain Credit Card Information?

“Credit card fraud is mainly triggered by the cardholder's carelessness with his data or a security breach on a website. Some instances are as follows:
-Unknown individuals receive a consumer's credit card number.
-Someone else uses a card that has been lost or stolen.
-Criminals steal mail from the intended receiver and utilise it for their own purposes.
-Employees at businesses copy the owner's cards or card numbers.
-Making a credit card that isn't real or counterfeit credit card.”

- Cloning and skimming - Cloning entails using software to duplicate the card's details and transferring them to another card. The equipment used to replicate the device is known as skimming.

- Smishing - The scammers send a text message containing a URL that they ask you to click. When we click on such a link, malicious material is downloaded to our mobile device. It sends all of the information from our phone.

- Vishing — Fraudsters impersonate bank employees and phone the cardholder to obtain information such as the card number, CVV, OTP, and expiration date. Banks never ask for such personal information over the phone.

- Identity Theft - Creating a fictitious identity in the likeness of another individual in order to get and use their credit card.

Important Precautions to Prevent Credit Card Fraud

- Never lend your card to anyone else.

- Never give up your banking information on any internet portals or social media sites.

- Nowhere should the credit card number or pin number be written.

- Do not give out card numbers or OTPs over the phone.
- Do not forget about the card or the receipts.

- In the event of a lost card, a change of address, or any suspicious cards, contact the card company right away.

1.2 Problem Statement

“The Credit Card Fraud Detection Problem includes modeling beyond credit card transactions with the know-how of the ones that became out to be a fraud. This model is then used to pick out whether a new transaction is fraudulent or no longer”.

“Our aim right here is to hit upon 100% of the fraudulent transactions while minimizing the wrong fraud classifications”.

“The Credit Card Fraud Detection Problem consists of modeling pvevious credit rating card transactions with the info of those that grew to grow to be out to be fraud. This model is then used to discover whether a brand new transaction is fraudulent or no longer. Our intention right right here is to find out one hundred% of the fraudulent transactions while minimizing the incorrect fraud classifications”.

What does credit card fraud detetcion includes?

“The Credit Card Fraud Detection Problem includes modeling beyond credit card transactions with the expertise of those that grew to come to be out to be fraud.
This model is then utilized in figuring out if any new transaction is fraudulent or now not”.

1.3 Objectives

“The purpose of this assignment is to predict whether or not a credit score card transaction is fraudulent or not, based totally on the transaction quantity, location and other transaction related records. It ambitions to song down credit card transaction information, that's executed by way of detecting anomalies within the transaction information. Credit card fraud detection is usually applied using an algorithm that detects any anomalies inside the transaction statistics and notifies the cardholder (as a precautionary degree) and the bank approximately any suspicious transaction.

The key objective of our credit card fraud detection gadget is to perceive suspicious activities and file them to an analyst at the same time as letting normal transactions be mechanically processed.

For years, monetary establishments have been entrusting this assignment to rule-based systems that rent rule units written with the aid of professionals. But now they increasingly turn to a device getting to know method, as it can deliver huge improvements to the process.

1. Higher accuracy of fraud detection. Compared to rule-primarily based solutions, system gaining knowledge of equipment have better precision and go back greater relevant results as they do not forget more than one additional factors. This is due to the fact ML technology can take into account many greater data points, along with the tiniest info of conduct patterns associated with a specific account.

2. Less manual paintings wanted for added verification. Enhanced accuracy leads reduces the load on analysts. “People are not able to test all transactions manually, although we are talking approximately a small financial institution,” Alexander Konduforov, statistics science competence leader at AltexSoft, explains. “ML-driven systems filter, kind of speaking, ninety nine. percentage of normal patterns leaving best 0.1 percentage of activities to be confirmed by means of specialists.”

3. “Fewer false declines. False declines or false positives manifest while a system identifies a valid transaction as suspicious and wrongly cancels it”.

4. “Ability to become aware of newer styles and adapting to modifications. Unlike rule-primarily based structures, ML algos are aligned to a constantly converting surroundings and financial situations. They allow analysts to become aware of new

suspicious patterns and to make new guidelines to prevent new styles of scams".

1.4 Methodology

"Credit Card Fraud Detection using Machine Learning"

There's a Data Science team that studies data and builds a model for detecting and preventing the illegal/fraud transactions. It is accomplished by combining all the pertinent aspects of cardholder transactions, including the date, user zone, product category, amount, supplier, and the client's behavioural tendencies, among others. The data is then loaded into a machine learning model that has been trained to look for patterns and rules in order to determine if a transaction is fraudulent.

"Introduction To Machine Learning"

"Machine Learning knowledge of is a department of artificial intelligence (AI). Machine studying's purpose is to recognize the shape of facts and in shape it into models that humans can apprehend and utilise."

"Machine Learning's a subfield of pc technology that differs from conventional computational strategies. In classical computing, algorithms are collections of explicitly certain commands used by computers to calculate or solve issues.Machine studying, however, permits computer systems to study from records inputs after which use statistical analysis to provide outputs which can be within a given variety. Machine getting to know makes it simpler for computer systems to develop fashions from pattern records and automate choice-making techniques based totally on facts inputs as a result."

"Everyone who utilises technology nowadays has profited from ML. Face recognition era can be utilized by social media networks to help customers tag and share pix of pals. Optical individual recognition (OCR) generation converts textual content pictures into movable kind. Using gadget mastering, advice systems endorse what movies or TV series to observe next based totally at the person's possibilities. Consumers might also soon have the ability to buy self-driving automobiles that navigate the usage of gadget gaining knowledge of."

“Machine gaining knowledge of is a swiftly changing place. As a result, whether or not operating with system getting to know approach or analysing the impact of system studying techniques, there are a few elements to do not forget.”

Methods of Machine Learning

In machine gaining knowledge of, obligations are regularly divided into large categories. These classifications are based totally on how data is acquired and the way the system responds to it.

Two of the maximum widely used gadget getting to know methods are unsupervised mastering, which gives the algorithm with out a labelled information so as for it to find structure inside its input statistics, and supervised mastering, which trains algorithms primarily based on example input and output facts that is labelled via humans. Let's take a deeper study each of those strategies.

Supervised Learning

“The laptop is furnished sample inputs and their anticipated consequences are labelled in supervised getting to know. The cause of this strategy is for the set of rules to "research" through comparing its actual output to the "found" outputs that allows you to find flaws and replace the version because of this. As a result, patterns are utilized in supervised learning to are looking ahead to label values on unlabeled records”.

“Using supervised gaining knowledge of, an set of rules might be given records with images of sharks labelled as fish and snap shots of seas labelled as water. After being educated on this facts, the supervised getting to know machine need if you want to realise unlabeled shark photographs as fish and unlabeled ocean snap shots as water”.

“A common use of supervised studying is to utilise preceding information to predict statistically in all likelihood destiny events. It is probably used to forecast destiny stock marketplace moves or to clean out direct mail emails based on historic stock marketplace information. Untagged dog snap shots can be used as enter data in supervised studying to classify tagged dog pics”.

Unsupervised Learning

Because the information in unsupervised getting to know isn't labelled, it's as much as

the mastering set of rules to find commonalities within the information it's given. Because unlabeled records is more considerable than labelled statistics, gadget studying algorithms that favour unsupervised learning are relatively effective.

Unsupervised learning might have the primary cause of discovering hidden styles in a dataset, or it can have the purpose of characteristic getting to know, which allows a computational gadget to discover the representations needed to classify raw statistics on its own.

Unsupervised gaining knowledge of is substantially used for transactional data. You may also have a massive database of customers and their transactions, but as a human, you're not going a good way to deduce what related traits can be inferred from purchaser profiles and purchase sorts.

With this records positioned into an unmonitored getting to know device, it could have the ability to deduce that women of a given age group who use unscented soaps are probable to be pregnant, and so a advertising campaign emphasising pregnancy and toddler objects can be targeted to this institution to promote purchases.

Without being provided a "right" response, unsupervised gaining knowledge of systems may have a look at complicated cloth this is greater broad and looks unconnected with a purpose to organise it in doubtlessly applicable ways.
Unsupervised gaining knowledge of is generally used in anomaly detection, consisting of identifying fraudulent credit card purchases and recommender structures that suggest what to buy subsequent. Untagged dog pics can be used as enter information for an algorithm that makes use of unsupervised studying to locate likenesses and organization canine pix collectively.

Approaches

As a field, gadget gaining knowledge of is carefully associated with computational information, therefore having a heritage in statistics assist you to better realise and use machine mastering strategies.

“For the ones who've now not studied records before, a definition of correlation and regression, two frequently used methodologies for exploring the connection among quantitative statistics, is a high-quality area to start. Correlation is a diploma of the relationship among variables that aren't installed or independent of every different.
Regression is used to observe the connection amongst one primarily based and one impartial variable at its most fundamental diploma. Because they'll be used to forecast

the dependent variable while the independent variable is thought, regression information provide prediction competencies."

Programming Languages

When deciding on a language to specialise in the usage of gadget getting to know, recall the talents required on contemporary activity commercials in addition to libraries available in numerous languages that can be utilised for system mastering approaches.

According to data from Indeed.Com activity posts in December 2016, Python is the most sought-after programming language within the gadget studying professional commercial enterprise. Python is observed by Java, R, and subsequently C++.

"Python's rise might be attributed to the trendy improvement of deep getting to know frameworks for this language, together with TensorFlow, PyTorch, and Keras. Python, as a language with clean syntax and the capability for use as a scripting language, famous itself to be powerful and easy each for getting geared up statistics and working immediately with information. The scikit-examine system gaining knowledge of toolkit is constructed on pinnacle of a number of popular Python applications, such as NumPy, SciPy, and Matplotlib, which Python programmers are in all likelihood already familiar with".

Java is appreciably utilised in corporate programming, and it's far often hired by the front-end laptop software builders who are also running on machine studying at the corporation level. It isn't the first desire for folks who are new to programming and need to find out about system mastering, however it's miles favoured via human beings who've worked with Java before and need to apply it to device studying. For network security, together with cyber attack and fraud detection use instances, Java is extra widely used than Python in industrial device studying systems.

"MALLET (MAchine Learning for LanguagE Toolkit) allows device gaining knowledge of packages on textual content, together with herbal language processing, difficulty be counted modelling, record type, and clustering; and Weka is a hard and fast of device reading algorithms for records mining obligations. Deeplearning4j is an open-supply and allotted deep-getting to know library written in Java and Scala; MALLET (MAchine Learning for LanguagE Toolkit) permits system mastering packages on text, which includes herbal language processing, subject matter modelling, document type"

R is a statistical computing laptop language this is loose and open-supply. It has risen in reputation in latest years, and many teachers find it attractive. Although R has increased in reputation in commercial packages as interest in information technology has increased, it is not frequently hired in business production conditions. Caret (short for

Classification And REgression Training) is a famous R system studying bundle for building predictive models, randomForest for classification and regression, and e1071 for facts and possibility principle.

C++ is the language of choice for device studying and artificial intelligence in gaming and robotic packages (inclusive of robot locomotion). Embedded computer hardware builders and electronics engineers are more likely to apply C++ or C in gadget gaining knowledge of packages due to their talent and degree of control over the language.Two gadget gaining knowledge of libraries that you could utilise with C++ are the scalable mlpack and Dlib, which each provide a extensive variety of machine learning abilties.

1.5 Organization

> Chapter I, Contains the Introduction, Problem statement, scope, Objectives of the System or Project.

> Chapter 2, the literature survey discusses an abstract survey of the published papers and if any disadvantages are identified in the paper.

> Chapter 3 discusses the detailed requirement of the problem identified for the major project, system architecture and implementation details

> Chapter 4 discusses the Performance Analysis of the model

> Chapter 5 Concludes the Report

> Chapter 6 Discusses any Future Scope

Chapter 02:- LITERATURE SURVEY

2.1 Survey

"A Fraud act due to the fact the unlawful or criminal deception meant to result in economic or personal benefit. It is a deliberate act that is a crime, rule or insurance with an aim to obtain unauthorized monetary benefit. Numerous literatures concerning anomaly or fraud detection in this place have been posted already and are available for public usage. A comprehensive survey achieved thru Clifton Phua and his pals have revealed that techniques employed on this location encompass statistics mining applications, computerized fraud detection, adversarial detection. Unconventional techniques together with hybrid information mining/complicated community class set of regulations is able to understand unlawful instances in an actual card transaction facts set, primarily based totally on community reconstruction set of regulations that permits developing representations of the deviation of 1 instance from a reference institution have proved green commonly on medium sized on line transaction. The fraud detection is a complex undertaking and there can be no device that efficaciously predicts any transaction as fraudulent". The houses for an top notch fraud detection system are:

1. Should discover the frauds correctly.
2. Should come upon these frauds rapid.
3. Shouldn't classify a proper transaction as fraud.

Outlier detection is a vital mission as outliers imply strange walking conditions from which considerable overall performance degradation may additionally appear. Techniques used in fraud detection may be divided into two:

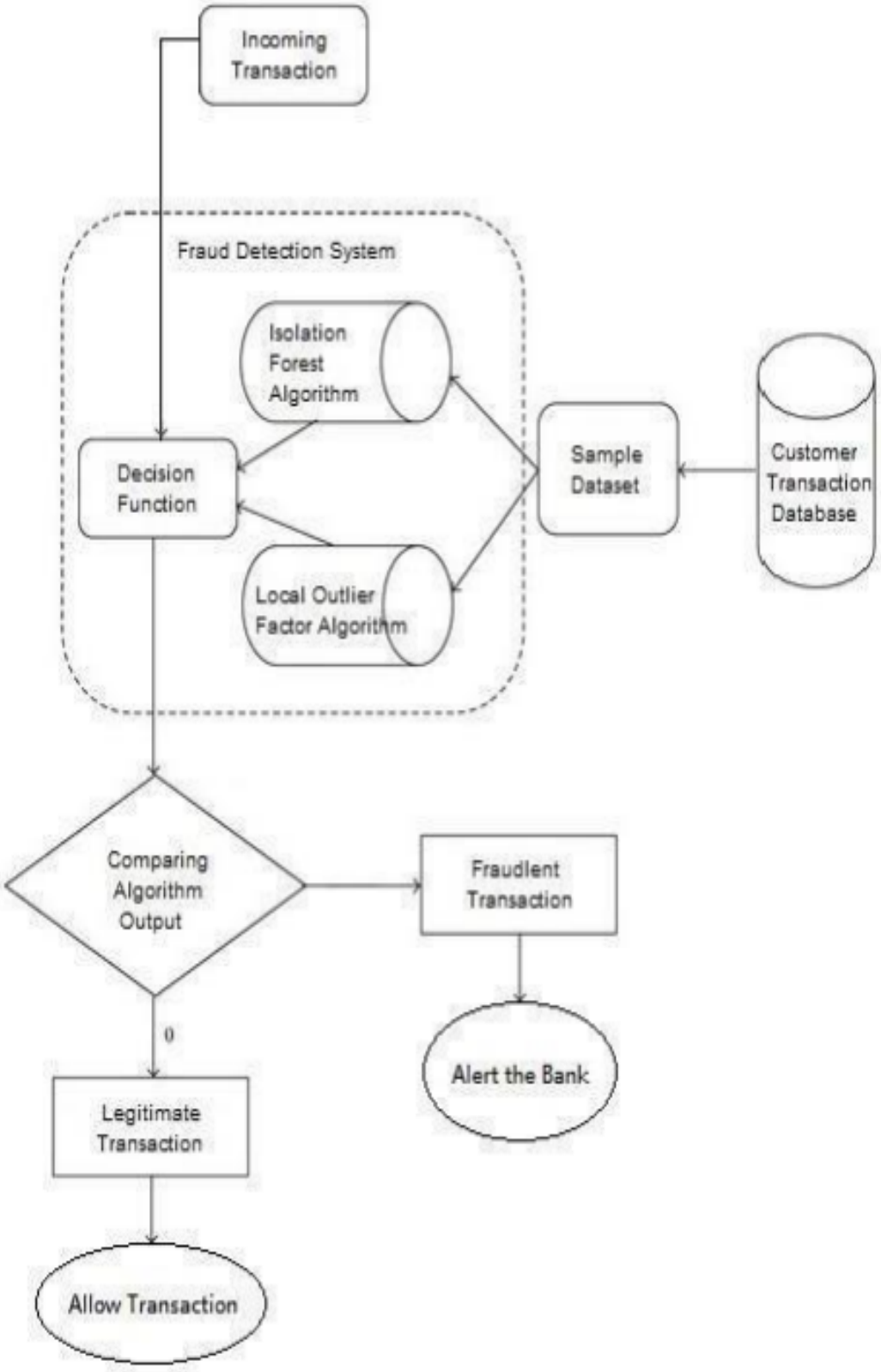
1) Supervised strategies in which beyond recognized valid/fraud instances are used to build a model for you to produce a suspicion score for the new transactions.

2) Unsupervised are those wherein there aren't any prior units wherein the country.

2.2 Proposed System

"Our Project most important purpose is to creating Credit Card Fraud Detection awaring to human beings from credit score rating card online frauds. The precept factor of credit score card fraud detection gadget is vital to safe our transactions & safety. With this machine, fraudsters do now not have the hazard to make a couple of transactions on a stolen or counterfeit card earlier than the cardholder is privy to the fraudulent hobby. This version is then used to pick out out whether a brand new transaction is fraudulent or no longer".

Our aim right here is to stumble on 100% of the fraudulent transactions at the same time as minimising the wrong fraud classifications.With the assist of following block diagram we can apprehend the functionalities of a Credit Card Fraud Detection.The following diagram suggests the entire Block Diagram :


Incoming Transaction
Fraud Detection System
Isolation Forest Algorithm
Decision Function
Sample Dataset
Customer Transaction Database
Local Outlier Factor Algorithm
Comparing Algorithm Output
Fraudlent Transaction
0
Alert the Bank
Legitimate Transaction
Allow Transaction

2.3 Description of the dataset

The data constitutes transactions performed by a cardholder over a two-day period, i.e. 2 days in Sept. 2013. There are a total of 284,807 transactions, of which 492 (or 0.172 percent) are fraudulent. Fraudulent transactions are transactions. This is a really imbalanced dataset. Due to the fact that disclosing transaction information of a user. The customer is classified as a confidential problem since the majority of the dataset's features have been changed. PCA stands for principal component analysis (PCA). PCA applied features and 'time' are represented by V1, V2, V3,..., V28, respectively. 'Amount' and 'class' are non-PCA applied characteristics, as seen in the table.

"It only accepts numerical input variables which have been converted using PCA. Due to confidentiality troubles, we are unable to provide the unique features and different historical past facts about the records. Features V1, V2,... V28 are the principal additives received with PCA; the simplest capabilities no longer changed with PCA are 'Time' and 'Amount.'The characteristic 'Time' shops the range of seconds that have handed between every transaction and the primary transaction in the dataset. The function 'Amount' represents the transaction Amount and may be used for example-based fee-sensitive gaining knowledge of. 'Class,' the reaction variable, has a value of one while there is fraud and 0 while there isn't.".

S. No.	Feature	Description
1.	Time	Time in seconds to specify the elapses between the current transaction and first transaction.
2.	Amount	Transaction amount
3.	Class	0 - not fraud 1 – fraud

2.4 Dataset acknowledgement

"Worldline and the Machine Learning Group (http://mlg.Ulb.Ac.Be) of ULB (Université Libre de Bruxelles) collaborated on the statistics collection and analysis as a part of a large facts mining and fraud detection research venture."

"More data on contemporary and former fraud detection tasks may be located at https://www.Researchgate.Net/challenge/Fraud-detection-five and the DefeatFraud task page."

Please include the following works in your reference list:

Gianluca Bontempi, Andrea Dal Pozzolo, Olivier Caelen, Reid A. Johnson Calibrating Probability for Unbalanced Classification Using Undersampling 2015 IEEE Symposium on Computational Intelligence and Data Mining (CIDM).

Andrea Dal Pozzolo; Olivier Caelen; Yann-Ael Le Borgne; Serge Waterschoot; Gianluca Bontempi Expert systems with applications, 41,10,4915-4928, 2014, Pergamon, Learned lessons in credit card fraud detection from a practitioner viewpoint Andrea Dal Pozzolo, Giacomo Boracchi, Olivier Caelen, Cesare Alippi, and Gianluca Bontempi. Credit card fraud detection: a realistic modelling and a unique learning technique, IEEE Transactions on Neural Networks and Learning Systems, vol. 29, no. 8, pp. 3784-3797, IEEE Press, 2018.

Andrea Dal Pozzolo, Andrea Dal Pozzolo, Andrea Dal Pozzolo, Andrea Dal Pozzolo.

2.5 Feasibility Study

Any key step in the software development process has been acquired. Allows developers to receive a working product that has been tested. Refers to product research that might be done in terms of product outcomes, application performance, and technical assistance needed to use it. A possible investigation should be carried out based on a variety of circumstances and conditions.

2.4.1) Economic Feasibility

Economic recovery is the difference between the advantages or results we get from a product and the overall cost we spend to enhance it.

The creation of a new product enhances system correctness and speeds up application and reporting processing in the present system.

2.4.2) Probability Feasibility

The performance of a product to make it work is referred to as availability. Some items may perform admirably during manufacture and usage, yet they may malfunction in real life. lt entails researching the needed personalities as well as their technical knowledge.

The contained data, updated information, and reports for generations are accurate and quick in the present system

2.4.3) Technical Feasibility

The term "technical performance" relates to whether or not the software currently available on the market is capable of completely supporting the present system.

lt investigates the benefits and drawbacks of utilizing specific development software, as well as its viability. lt also learns how much more time customers will need to make the app function. The user interface of the present system is usable and does not need a great deal of knowledge or training.

lt just takes a few mouse clicks to complete activities and generate reports. Because consumers want rapid access to websites with a high level of security, the software used to upgrade is best suited for current applications. This is accomplished by combining a web server and a data server in the same physical location.

Chapter 03:- SYSTEM DEVELOPMENT

3.1 Tools & Technologies Used

3.1.1) Python

Python is a robust programming language with a wide range of capabilities. Its broad features make working with targeted programs (including meta-programming and meta-objects) simple, and it fully supports object-oriented programming. Many additional paradigms, such as contract generation and logic programming, are supposed through extensions. Python takes advantage of power typing as well as the integration of reference computation and waste management waste collecting. lt also supports advanced word processing (late binding), which binds the way the words change during the process.

Patches to less essential sections of C Python that can give a minor improvement in performance at an obvious price are rejected by Python developers who try to prevent premature execution. When speed is crucial, the Python program developer can use mod-written modules in C-languages or PyPy, a timely compiler, to submit time-sensitive jobs. Cython is a Python interpreter that converts Python scripts to C and invokes the Python interpreter directly from the C-1evel API. Python developers strive to make the language enjoyable to use. Python's architecture supports Lisp culture in terms of functionality. Filters, maps, and job reduction, as well as a list comprehension, dictionaries, sets, and generator expressions, are all included.

Two modules (itertools and functools) in the standard library use realistic Haskell and Standard ML tools.

3.1.2) Why Python?

"We are using the python language due to the fact that Python may be applied at the significant majority of the ranges like Windows, Mac, Linux, Raspberry Pi, etc. Python is a stage-unfastened language. Python is comparative and as sincere as the English language. Python bolsters bunches of libraries and it has a sincere linguistic shape like English even as java and C++ have complicated codes. Python applications have fewer lines than a few other programming languages. That is the motive we use the Python language for guy-made awareness, AI, Dealing with large quantities of information. Python is an editorial organized language."

3.1.3) Machine Learning

Using Machine Learning, we will justify whether a credit card is fraudulent or not in this kernel.

- "Machine learning is the branch of science that studies how computers can learn without being explicitly programmed (Arthur Samuel, 1959)."
- "It is a type of artificial intelligence that is founded on the idea that robots may improve themselves without the need for human involvement by using data and patterns."
- "Machine learning algorithms are 'taught' to discover patterns and characteristics in massive amounts of data so that they can make judgements and predictions based on fresh information."
- "Machine learning is currently widely applied in a variety of fields. It also plays an important part in fraud detection services."
- "Various machine learning methods can be used to detect this type of fraud."

"Credit card transactions made via European cardholders in September 2013 are protected inside the information. We have 492 scams in this dataset, out of 284,807 transactions inside the previous two days. The statistics is critically skewed, with frauds accounting for simplest 0.172 percentage of all transactions."

The enter variables are translated into numerical using PCA transforms due to confidentiality issues.

3.2 Methods & Techniques Used

(1)Logistic Regression
(2)Decision Trees
(3)Random Forest
(4)KNN
(5) Isolation Forest
(6) Local Outlier Factor

A kind of supervised and semi-supervised device studying strategies are used for fraud detection, but our goal with the cardboard fraud dataset is to overcome 3 most important challenges: sturdy elegance imbalance, the inclusion of labelled and unlabelled samples, and the capability to system a big range of transactions."

"A series of operations conducted to prevent money or property from being gained under false pretences is known as fraud detection."

"The majority of detection structures use several fraud detection datasets to construct a linked picture of both prison and invalid payment statistics in an effort to make a end. When making this selection, IP deal with, geolocation, tool identification, "BIN" records, worldwide latitude/longitude, transaction history patterns, and real transaction records have to all be taken under consideration. In instruction, this means traders and issuers use a hard and fast of business guidelines or analytical algorithms to inner and external facts on the way to hit upon fraud."

“SUPERVISED ALGORITHMS”

Logistic regression is a supervised learning classification approach for predicting a target variable's probability. The goal or variable quantity has a dichotomous character, which means there are only two possible classes. In plain English, the variable quantity is doubled, with information coded as either 1 (represents success/yes) or 0 (represents disappointment/no). P(Y=1) is predicted as an element of X by a computed relapse model. It's one of the greatest ML calculations, which will be used for a variety of concerns such as spam detection, diabetes prediction, malignant growth detection, and so on.

Assumptions for Logistic Regression

Before we dive into the implementation of logistic regression, it's important to remember the following assumptions regarding the same.

- Because the target variables in binary logistic regression must always be binary, the desired outcome is represented by factor level 1.
- There should be no multicollinearity in the model, which means that the unbiased variables must be independent of one another.
- For logistic regression, we need to choose an excessive sample size.

Logistic Regression's Benefits

- On linearly separable datasets, logistic regression works well.
- While logistic regression is less prone to overfitting, it can still overfit in high-dimensional datasets. In these cases, regularisation (L1 and L2) approaches should be used to minimise over-fitting.
- Logistic regression not only provides a proportion of how important an indicator (coefficient size) is, but also the direction in which it is related (positive or

negative).

- “**Decision Trees** is an example of supervised gadget mastering algorithms.The supervised gaining knowledge of algorithms own family includes the Decision Tree set of rules. Unlike different supervised gaining knowledge of algorithms, the selection tree approach may be used to tackle regression and class troubles. The intention of the usage of a Choice Tree is to assemble a training version that can be used to predict the magnificence or cost of the target variable by way of gaining knowledge of simple choice policies learned from earlier data (training statistics). When making use of Decision Trees to forecast a document's magnificence label, we start on the pinnacle of the tree. The report's attribute and the basis attribute's values are as compared. Based on the contrast, we follow the department that corresponds to that price and visit the subsequent node.”

“Decision trees are used to illustrate or inform choice-making with the aid of visually representing choices. When handling gadget gaining knowledge of and facts mining, choice trees are used as a prediction model. These models translate data observations into conclusions about the facts's intention value. The intention of selection tree mastering is to create a version that may predict the cost of a target primarily based at the enter facts. The leaves of the predictive model contain judgments approximately the facts's purpose price, while the branches represent facts attributes received via statement.”

- “With an emphasis on meta-classifiers and meta-getting to know approaches for managing extremely unbalanced credit card fraud statistics, the overall performance of Logistic Regression is significantly skewed credit card fraud information. To assign observations to a discrete set of instructions, the type algorithm logistic regression is utilised. Classification challenges include e mail unsolicited mail or not unsolicited mail, on-line transaction fraud or not fraud, and tumour malignant or benign.”

- “Random forest’s a supervised system studying technique for solving type and regression troubles. It uses the majority vote for class and the common for regression to generate selection timber from diverse samples. Due to the incorporated characteristic choice in the version constructing technique, Random Forests have a excessive prediction accuracy and may accommodate a large number of functions.”

Working of Random Forest Algorithm

With the help of the stages below, we can learn how the Random Forest algorithm works.

- Begin by determining irregular cases from a given dataset.
- Following that, for each case, this calculation will generate a choice tree. Then it will get the desired outcome from each decision tree at that point.
- For each expected outcome, a polling form will be projected in this movement. Finally, as the last forecast outcome, choose the most casted ballot expectation result.

Working Example of Algorithms

We started with the Train Test split, but made the switch to the K Fold method because the Train Test split frequently contributes to the formation of models with high bias and the possibility of selecting test data with similar values, i.e. non-random values, resulting in an inaccurate assessment of model performance. We designed a function to execute numerous algorithms with different K values of KFold, such as Logistic Regression, Random Classifier, KNN, and Multinomial Naive Bayes.

- The K Nearest Neighbour set of rules is a kind of supervised system learning method for fixing classification and regression prediction problems. In industry, however, it's basically used to tackle class and prediction problems. The following characteristics could be a terrific manner to explain KNN:
    (1) “Because it would not have a wonderful education section and as a substitute uses all of the records for education and classification, KNN is a lazy studying set of rules.”
    (2) “Because it would not have a wonderful education section and as a substitute uses all of the records for education and classification, KNN is a lazy studying set of rules.”

> The k-nearest neighbour technique is a sample popularity approach that can be used to classify and predict statistics. The okay in ok-nearest neighbour is a fantastic integer that is often small, consequently the time period is from time to time abbreviated as ok-NN. In either category or regression, the enter can be the k closest schooling cases interior a area. The okay-NN class algorithm might be the point of interest of our interest.Class club is the final results of this process. This adds a brand new object to the magnificence that has the maximum members among its k closest neighbours. When k = 1, the item is assigned to the class of the unmarried nearest neighbour.”

When KNN is used to group events, the result is determined by the class with the absolute best recurrence from the K-most similar events. In essence, each instance votes for his or her class, and the class with the highest votes is chosen as the winner of the prediction. If you're using K and have a large number of classes (e.g. 2), choosing a K value with an odd number to avoid a tie is a good idea. As a result, when you have an odd number of classes,

select a decent number for K. Ties can be consistently broken by increasing K by one and looking at the class of the most comparable occurrence in the preparation dataset. Here are a few things to remember:

As we decrease the worth of K to 1, our predictions abate stable.

Conversely, as we increment the value of K, our expectations become more steady on account of larger part casting a ballot/averaging, and accordingly, bound to make more precise forecasts (up to a specific degree) In the end, we begin to observe an expanding number of mistakes. It's now we all know we've pushed the worth of K too far.

In circumstances where we are taking a larger part vote (in favor of model: picking the mode during a characterization issue) among marks, we normally make K an odd number to have a tiebreaker.

Advantages:

- The algorithm is simple and straightforward to implement.
- There's no need to create a model, tune a few parameters, or make more assumptions.
- The algorithm can be changed as needed. It has applications in categorization, regression, and search.

Disadvantage

- The calculation gets essentially slower because the number of models as well as indicators/free factors increases.

“UNSUPERVISED ALGORITHMS”

1) Isolation Forests : “One of the latest techniques to come across anomalies is referred to as Isolation Forests. The set of rules is based totally at the reality that anomalies are data factors that are few and extraordinary. As a end result of those properties, anomalies are prone to a mechanism known as isolation”

“This method is noticeably useful and is essentially one of a kind from all present methods. It introduces the usage of isolation as a more powerful and green method to come across anomalies than the usually used basic distance and density measures. Moreover, this approach is an set of rules with a low linear time complexity and a small memory requirement. It builds a great performing version with a small range of timber the usage of small sub-samples of constant length, irrespective of the size of a statistics set. Typical machine getting to know strategies tend to work better when the styles they are trying to learn are balanced, which means the same amount of good and awful behaviors are gift in the dataset”

“How Isolation Forests Work The Isolation Forest set of rules isolates observations with the aid of randomly selecting a feature and then randomly selecting a break up value between the maximum and minimum values of the selected function. The common sense argument goes: isolating anomaly observations is easier due to the fact just a few conditions are had to separate those instances from the regular observations. On the alternative hand, keeping apart ordinary observations require greater situations. Therefore, an anomaly rating may be calculated as the quantity of conditions required to separate a given commentary”

“The way that the algorithm constructs the separation is with the aid of first developing isolation trees, or random choice trees. Then, the score is calculated as the path period to isolate the observation”.

2) "The LOF algorithm is an unmanaged outlier detection approach which computes the nearby density deviation of a given records factor with admire to its acquaintances. It considers as outlier samples which have a appreciably decrease density than their neighbors".

"The range of friends taken into consideration, (parameter n_neighbors) is normally selected 1) more than the minimal variety of items a cluster has to contain, so that different gadgets can be local outliers relative to this cluster, and 2) smaller than the most number of nearby items that can doubtlessly be local outliers. In exercise, such informations are commonly now not available, and taking n_neighbors=20 appears to paintings properly in preferred".

3.3 Requirements for Hardware and Software

> Pentium 4. Intel Core i3, 5, i7. and 2 GHz processor RAM must be at 512MB.

> Hard disc with a capacity of at least 10 GB

> Input Keyboard and Mouse are the devices that are used.

> Monitor or PC as an output device

> Versions of Windows 7, 10, and above are supported.

> Jupyter Notebook as a platform

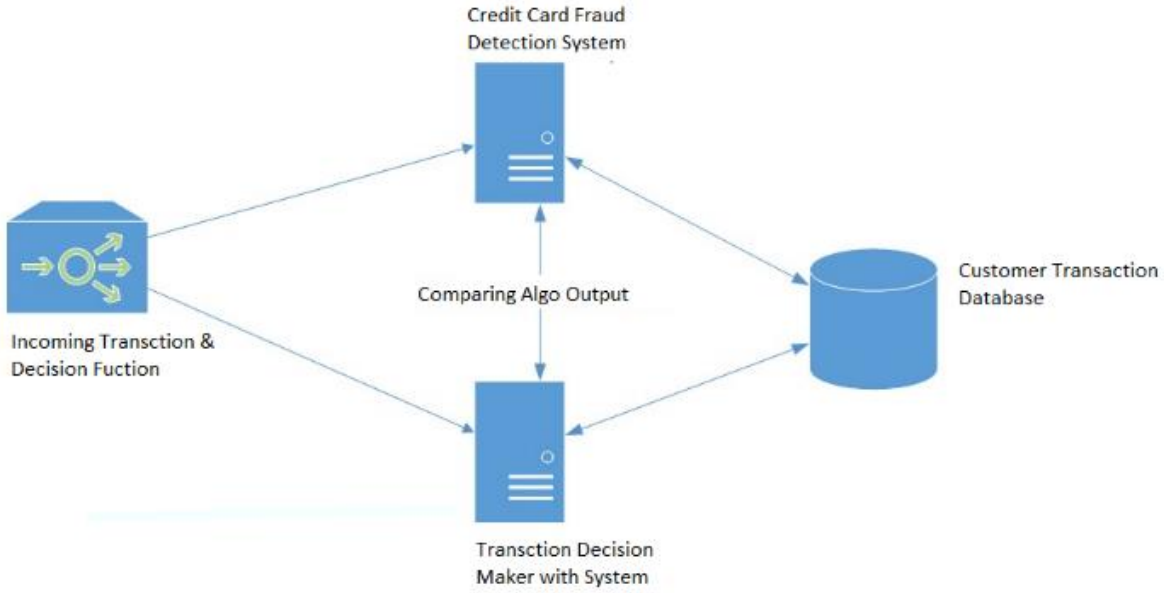
> Python Django (previous)

> Python, PostgreSQL. and Files as a Background

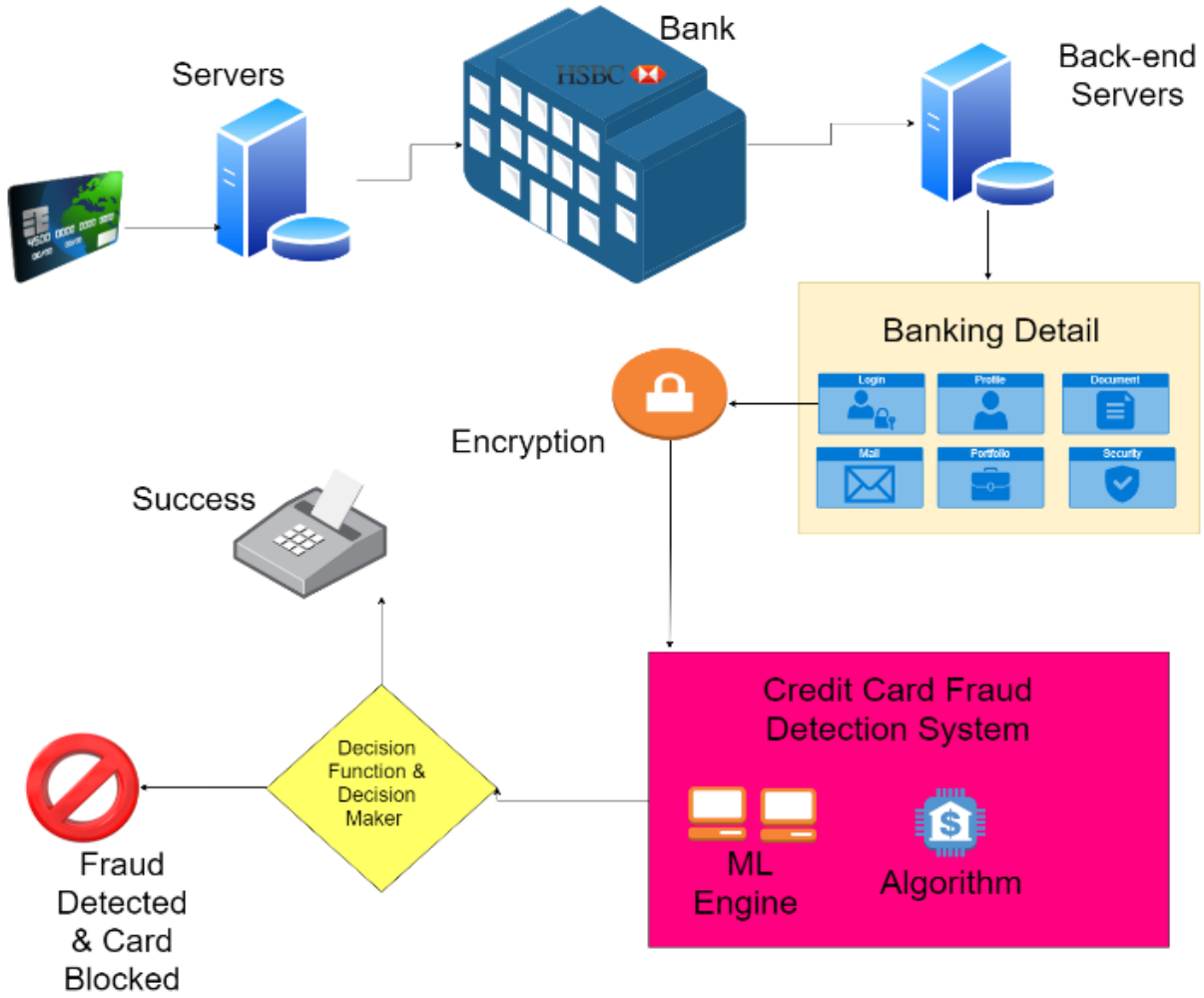
> Python is the programming language.

3.4 Methodology

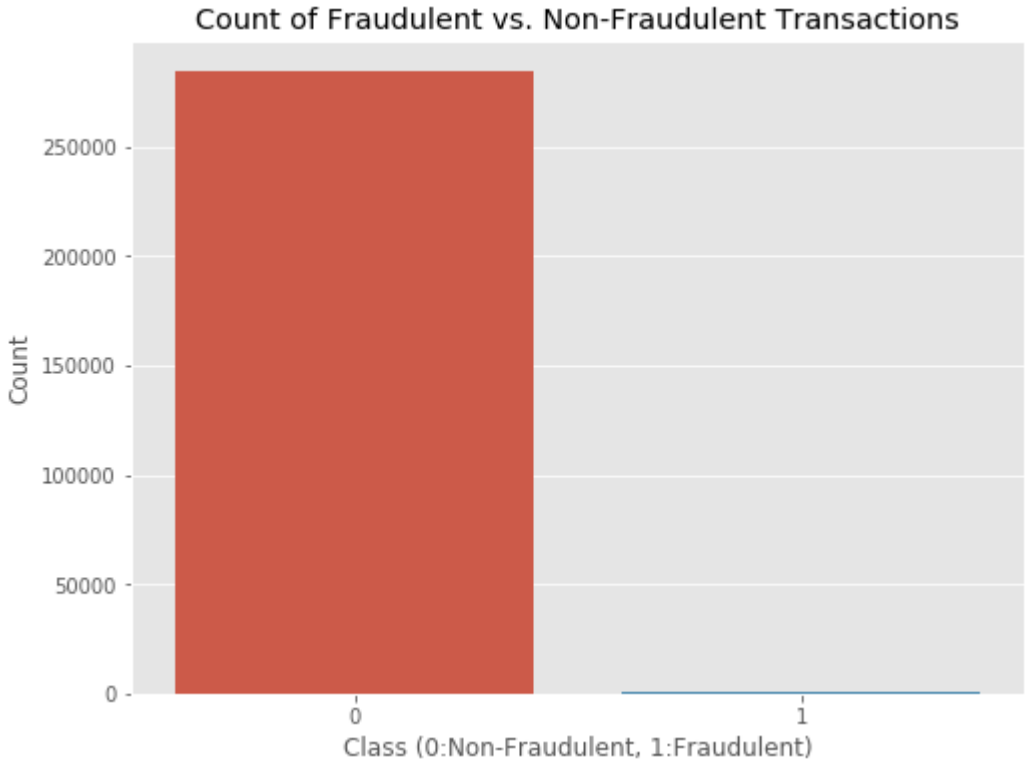
The technique that this paper proposes, makes use of the present day device getting to know algorithms to locate anomalous sports, known as outliers. The primary tough structure diagram may be represented with the following discern:



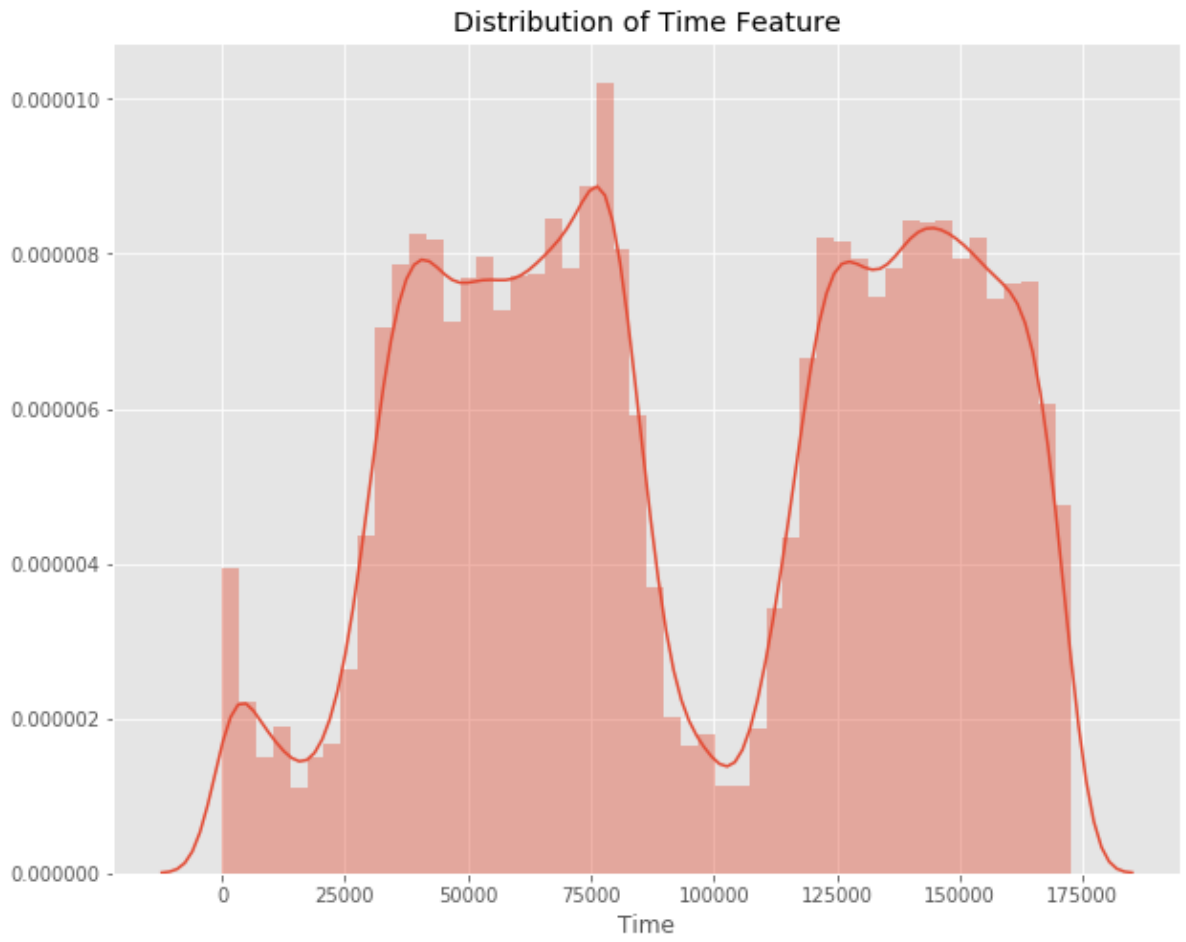
When checked out in detail on a larger scale at the side of real existence factors, the full architecture diagram may be represented as follows:



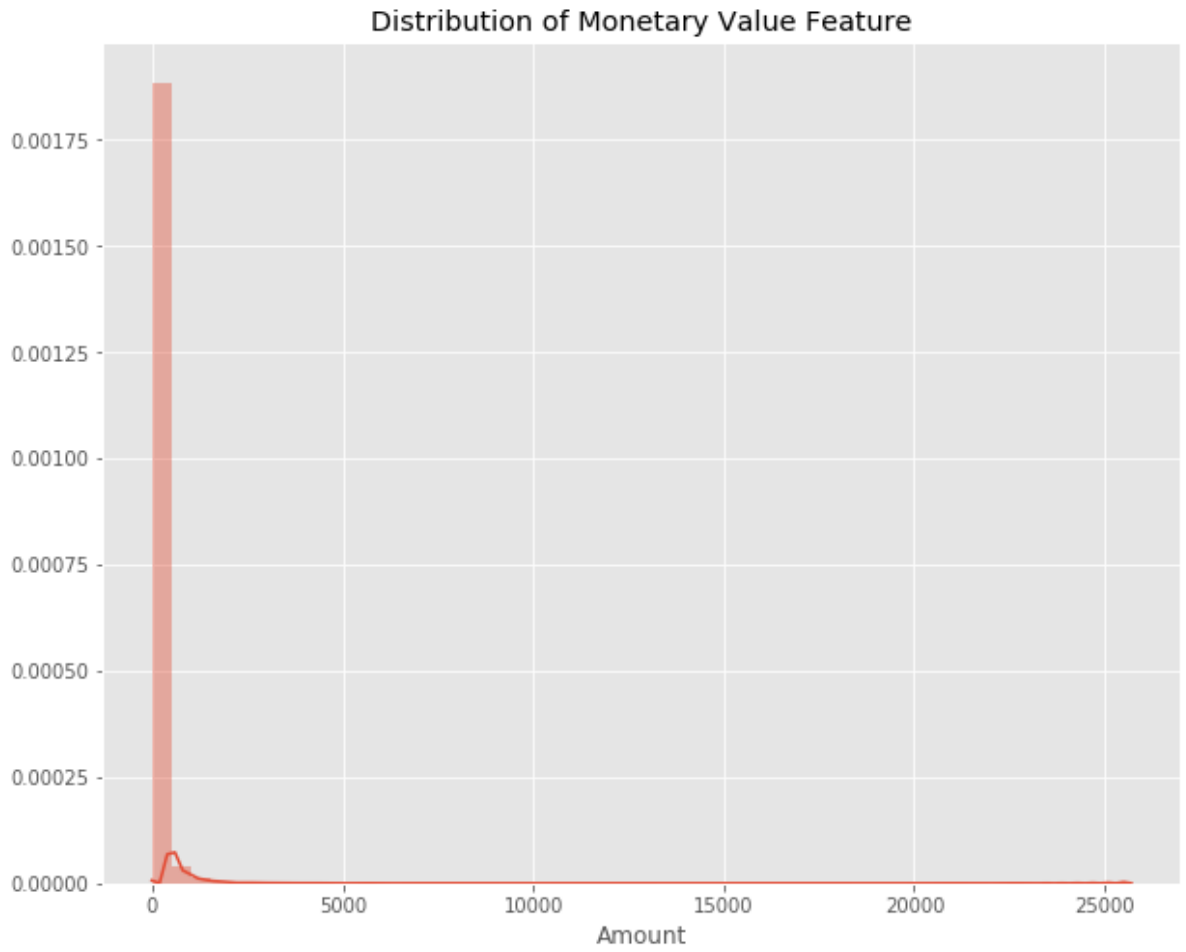
First of all, we acquired our dataset from Kaggle, a information analysis internet site which gives datasets.Inside this dataset, there are 31 columns out of which 28 are named as v1-v28 to shield sensitive facts.The other columns constitute Time, Amount and Class. Time suggests the time hole between the primary transaction and the following one. Amount is the amount of cash transacted. Class zero represents a valid transaction and 1 represents a fraudulent one.We plot specific graphs to check for inconsistencies inside the dataset and to visually realize it:



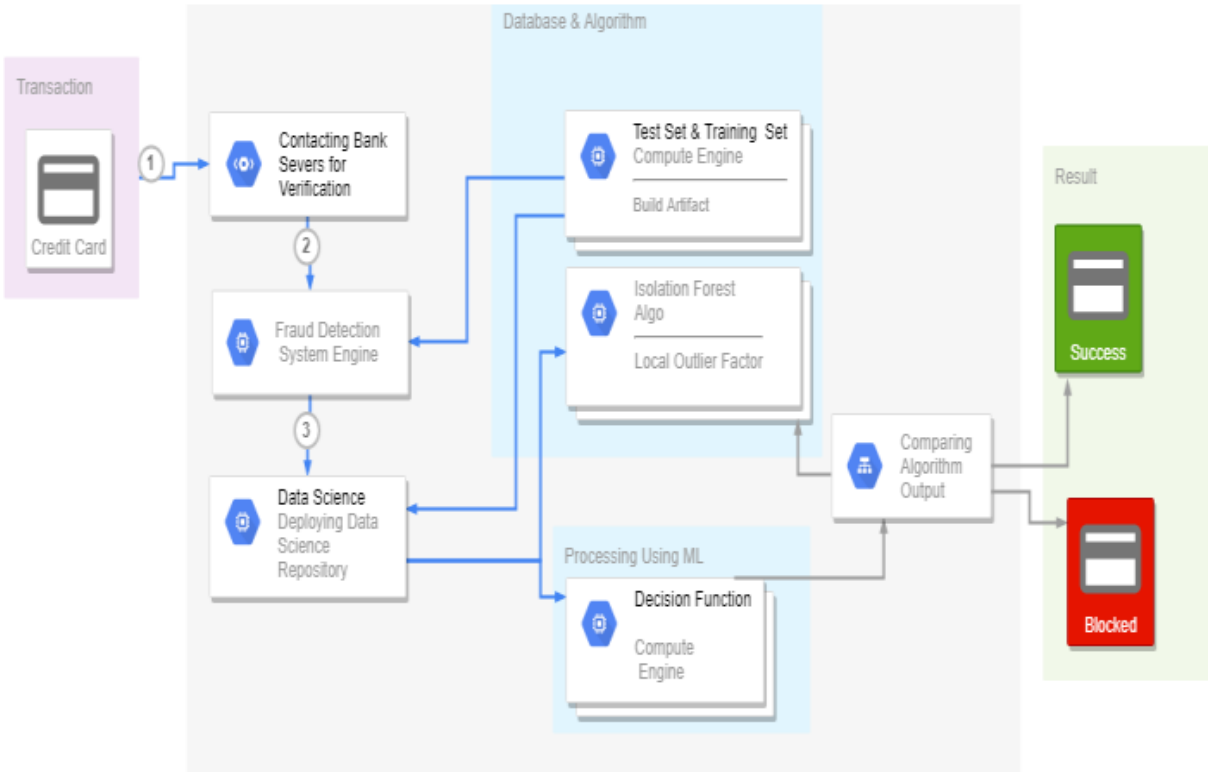
This graph shows that the number of fraudulent transactions is much lower than the legitimate ones.



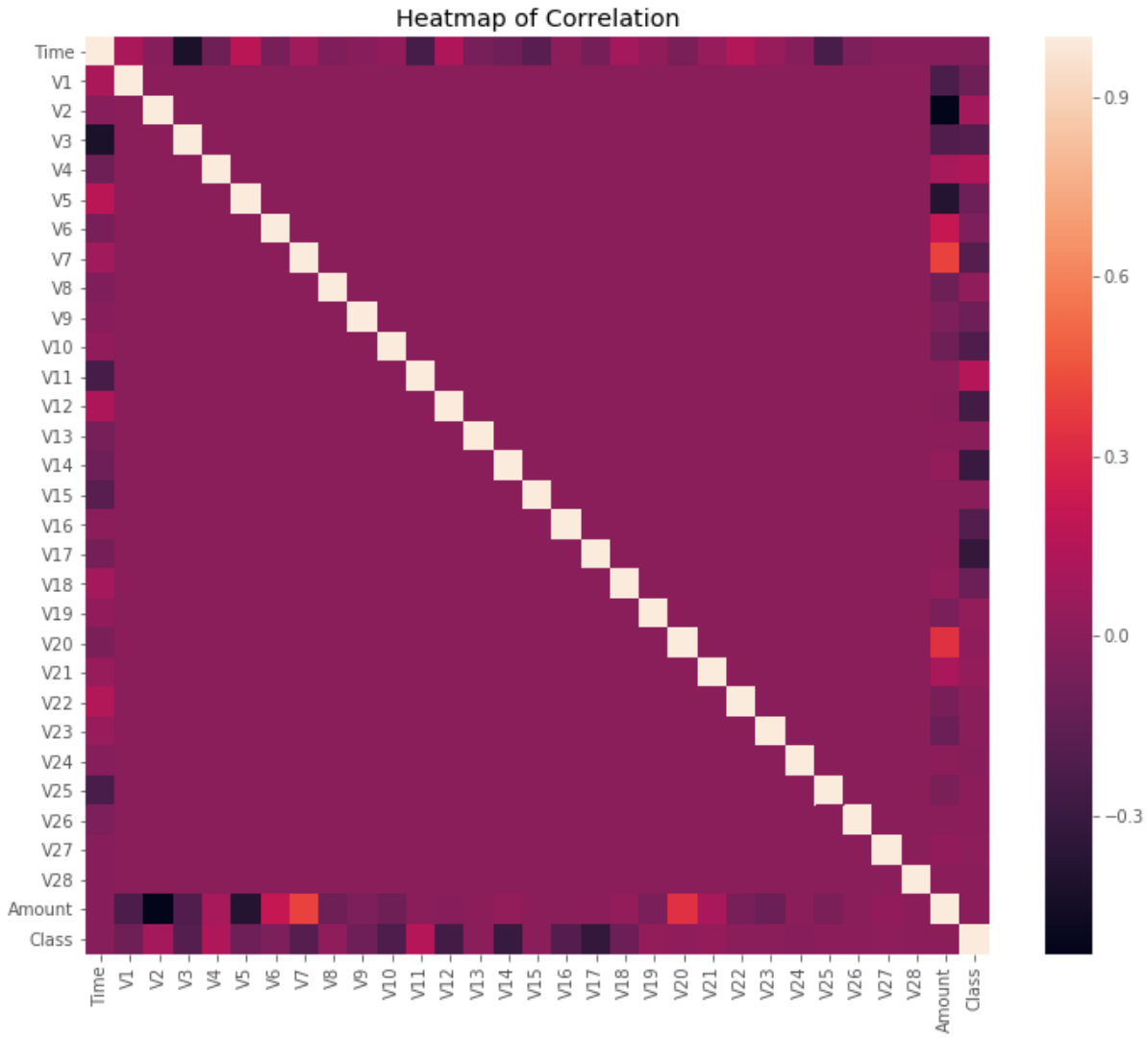
“This graph suggests the instances at which transactions have been finished inside days. It can be visible that the least number of transactions had been made during night time time and maximum all through the days.”



This graph represents the quantity that became transacted. A majority of transactions are especially small and only a handful of them come close to the most transacted amount.After checking this dataset, we plot a histogram for every column. This is done to get a graphical illustration of the dataset which may be used to confirm that there aren't any lacking any values in the dataset. This is done to make sure that we don’t require any lacking cost imputation and the gadget studying algorithms can system the dataset smoothly.



After this evaluation, we plot a heatmap to get a coloured representation of the records and to look at the correlation among out predicting variables and the class variable. This heatmap is shown underneath:



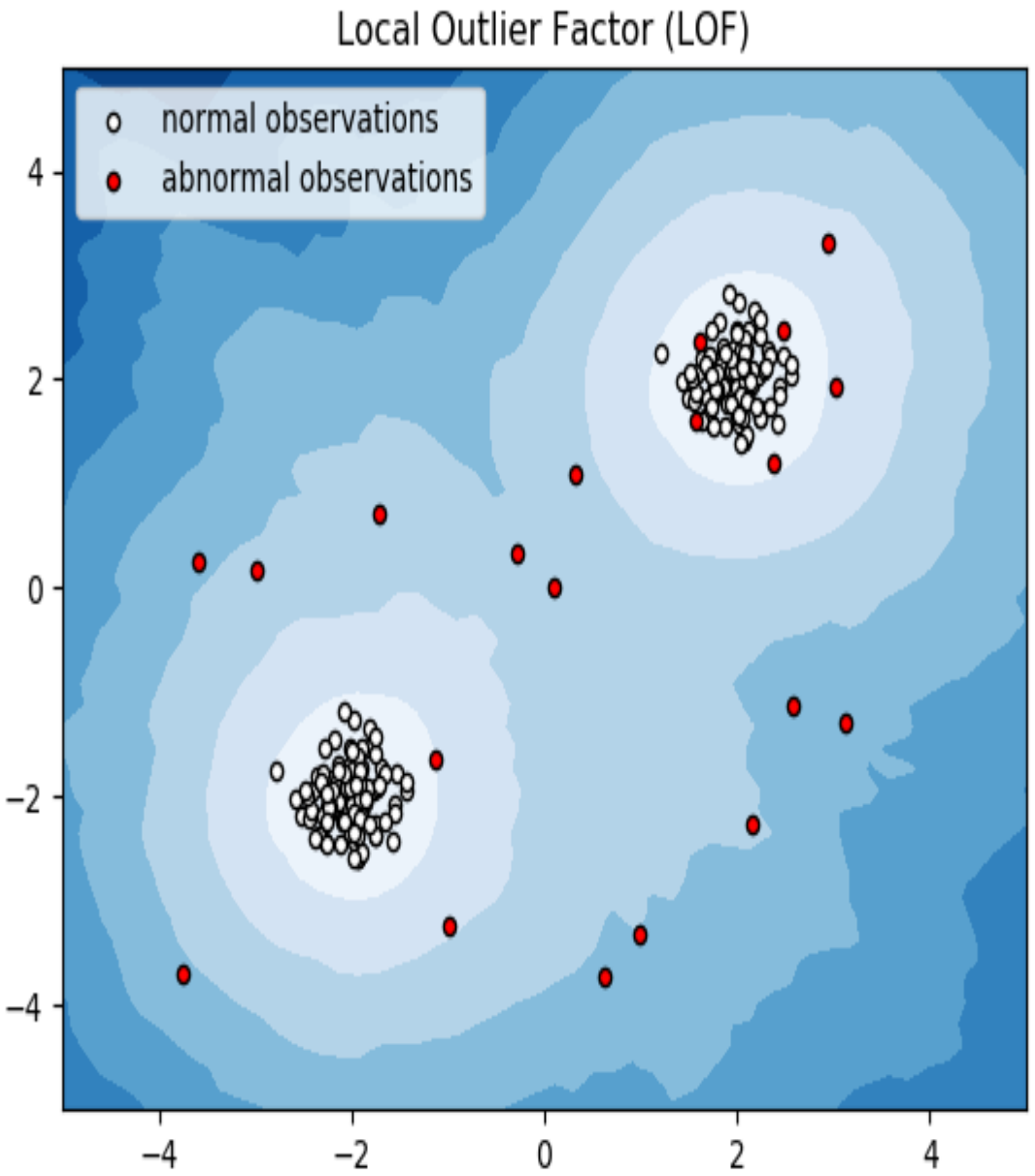
The dataset is now formatted and processed. The time and quantity column are standardized and the Class column is eliminated to make certain fairness of evaluation. The information is processed through way of a hard and fast of algorithms from modules. The following module diagram explains how these algorithms paintings together:

“This records is match right into a model and the subsequent outlier detection modules are accomplished on it:
• Local Outlier Factor
• Isolation Forest Algorithm
These algorithms are a part of sklearn. The ensemble module in the sklearn bundle consists of ensemble-primarily based definitely strategies and talents for the magnificence, regression and outlier detection.
This free and open-source Python library is constructed the usage of NumPy, SciPy and matplotlib modules which offers a ramification of easy and green tools which may be used for facts evaluation and tool learning. It capabilities numerous category, clustering and regression algorithms and is designed to interoperate with the numerical and scientific libraries.We’ve used Jupyter Notebook platform to make a software in Python to demonstrate the technique that this paper indicates. This application also can be completed on the cloud the use of Google Collab platform which supports all python pocket book documents.”
Detailed factors about the modules with pseudocodes for their algorithms and output graphs are given as follows:

A. Local Outlier Factor: “It is an Unsupervised Outlier Detection set of policies. 'Local Outlier Factor' refers to the ambiguity score of every pattern. It measures the neighborhood deviation of the sample statistics with understand to its neighbours. More precisely, locality is given with the resource of ok-nearest neighbours, whose distance is used to estimate the nearby statistics.”The pseudocode for this set of policies is written as:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import IsolationForest

rng = np.random.RandomState(42)

# Generate train data
X = 0.3 * rng.randn(100, 2)
X_train = np.r_[X + 2, X - 2]
# Generate some regular novel observations
X = 0.3 * rng.randn(20, 2)
X_test = np.r_[X + 2, X - 2]
# Generate some abnormal novel observations
X_outliers = rng.uniform(low=-4, high=4, size=(20, 2))

# fit the model
clf = IsolationForest(behaviour='new', max_samples=100,
                      random_state=rng, contamination='auto')
clf.fit(X_train)
y_pred_train = clf.predict(X_train)
y_pred_test = clf.predict(X_test)
y_pred_outliers = clf.predict(X_outliers)

# plot the line, the samples, and the nearest vectors to the plane
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf.decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

On plotting the results of Local Outlier Factor algorithm, we get the following figure:

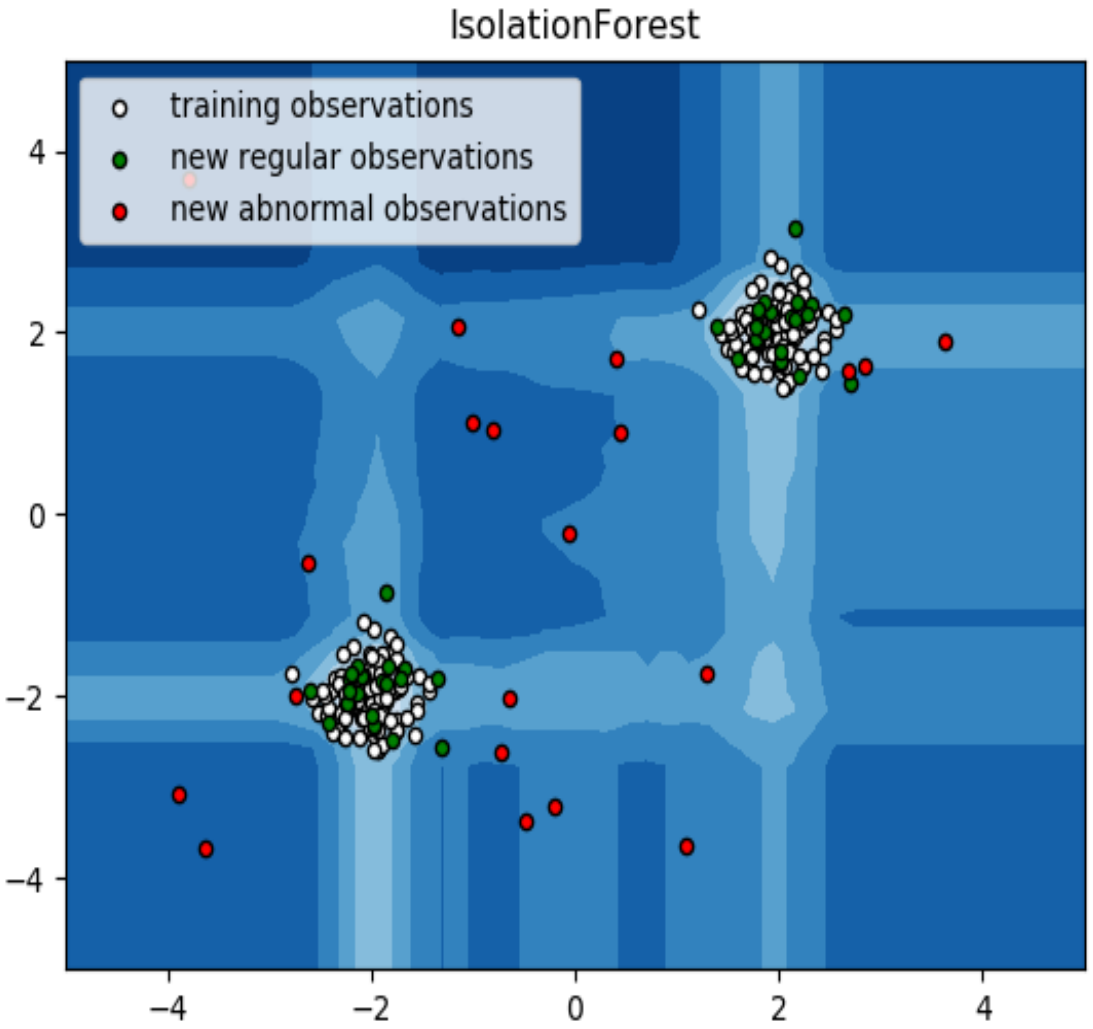


“By evaluating the community values of a sample to that of its neighbours, you could nonetheless identify samples which can be considerably decrease than their neighbours. These values are quite amanous and they are considered as outliers.As the dataset might be very large, we used handiest a fragment of it in out checks to lessen processing times.The very last stop result with the complete dataset processed is likewise decided and is given in the outcomes section of this paper.”

B. Isolation Forest Algorithm: “The Isolation Forest ‘isolates’ observations by means of the use of arbitrarily deciding on a characteristic after which randomly selecting a break up rate some of the maximum and minimum values of the positive function. Recursive partitioning can be represented by means of a tree, the range of splits required to isolate a pattern is equal to the course length root node to terminating node. The not unusual of this course length gives a degree of normality and the selection characteristic which we use.”The pseudocode for this algorithm can be written as:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn.neighbors import LocalOutlierFactor

np.random.seed(42)

# Generate train data
X = 0.3 * np.random.randn(100, 2)
# Generate some abnormal novel observations
X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
X = np.r_[X + 2, X - 2, X_outliers]

# fit the model
clf = LocalOutlierFactor(n_neighbors=20)
y_pred = clf.fit_predict(X)
y_pred_outliers = y_pred[200:]

# plot the level sets of the decision function
xx, yy = np.meshgrid(np.linspace(-5, 5, 50), np.linspace(-5, 5, 50))
Z = clf._decision_function(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)
```

On plotting the results of Isolation Forest algorithm, we get the following figure:



Partitioning them randomly produces shorter paths for anomalies. When a forest of random trees at the same time produces shorter route lengths for particular samples, they're extremely probable to be anomalies. Once the anomalies are detected, the gadget may be used to file

them to the concerned government. For checking out functions, we are comparing the outputs of those algorithms to decide their accuracy and precision.

3.5 Implementation

This concept is tough to put into effect in actual lifestyles because it calls for the cooperation from banks, which aren't inclined to percentage statistics due to their market opposition, and also due to legal reasons and protection of information of their users. Therefore, we seemed up some reference papers which observed similar approachesand accrued consequences. As said in such a reference papers:"This approach become carried out to a full software records set supplied by using a German financial institution in 2006. For banking confidentiality motives, simplest a summary of the consequences acquired is offered underneath. After applying this technique, the level 1 listing encompasses a few instances however with a high chance of being fraudsters. All individuals referred to on this listing had their playing cards closed to avoid any danger because of their high-risk profile. The situation is extra complex for the other listing. The degree 2 listing remains limited appropriately to be checked on a case by case foundation. Credit and series officers considered that 1/2 of the instances in this listing can be taken into consideration as suspicious fraudulent behaviour. For the closing list and the biggest, the work is equitably heavy. Less than a 3rd of them are suspicious. In order to maximize the time performance and the overhead prices, a opportunity is to include a brand new element in the question; this element may be the 5 first digits of the cellphone numbers, the e-mail deal with, and the password, as an instance, the ones new queries may be implemented to the level 2 listing and stage three list.".

Chapter 04:- PERFORMANCE ANALYSIS

4.1 Dataset

The data constitutes transactions performed by a cardholder over a two-day period, i.e. 2 days in Sept. 2013. There are a total of 284,807 transactions, of which 492 (or 0.172 percent) are fraudulent. Fraudulent transactions are transactions. This is a really imbalanced dataset. Due to the fact that disclosing transaction information of a user. The customer is classified as a confidential problem since the majority of the dataset's features have been changed. PCA stands for principal component analysis (PCA). PCA applied features and 'time' are represented by V1, V2, V3,..., V28, respectively. 'Amount' and 'class' are non-PCA applied characteristics, as seen in the table.

4.2 Performance Evaluation

The code prints out the wide variety of false positives it detected and compares it with the actual values. This is used to calculate the accuracy score and precision of the algorithms. The fraction of data we used for faster checking out is 10% of the complete dataset. The entire dataset is likewise used at the end and each the results are printed.These consequences together with the classification report for each algorithm is given within the output as follows, where class 0 manner the transaction turned into determined to be legitimate and 1 means it become decided as a fraud transaction.This end result matched against the magnificence values to test for fake positives.

Results whilst 10% of the dataset is used:

```
Isolation Forest
Number of Errors: 71
Accuracy Score: 0.99750711000316

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.28      0.29      0.28        49

    accuracy                           1.00     28481
   macro avg       0.64      0.64      0.64     28481
weighted avg       1.00      1.00      1.00     28481


Local Outlier Factor
Number of Errors: 97
Accuracy Score: 0.9965942207085425

              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

    accuracy                           1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```

Results with the complete dataset is used:

```
Isolation Forest
Number of Errors: 659
Accuracy Score: 0.9976861523768727

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.33      0.33      0.33       492

    accuracy                           1.00    284807
   macro avg       0.66      0.67      0.66    284807
weighted avg       1.00      1.00      1.00    284807


Local Outlier Factor
Number of Errors: 935
Accuracy Score: 0.9967170750718908

              precision    recall  f1-score   support

           0       1.00      1.00      1.00    284315
           1       0.05      0.05      0.05       492

    accuracy                           1.00    284807
   macro avg       0.52      0.52      0.52    284807
weighted avg       1.00      1.00      1.00    284807
```

Chapter 05:- CONCLUSION

5.1 Conclusion

Credit card fraud is truely an act of criminal dishonesty. This article has indexed out the most common techniques of fraud in conjunction with their detection strategies and reviewed recent findings on this area. This paper has additionally defined in element, how machine gaining knowledge of may be carried out to get better outcomes in fraud detection along side the set of rules, pseudocode, explanation its implementation and experimentation results.

While the algorithm does attain over ninety nine.6% accuracy, its precision remains best at 28% whilst a 10th of the data set is considered. However, whilst the complete dataset is fed into the set of rules, the precision rises to 33%. This excessive percentage of accuracy is to be anticipated due to the large imbalance among the wide variety of legitimate and wide variety of real transactions.

Since the entire dataset consists of most effective days' transaction data, its only a fragment of records that can be made to be had if this project had been to be used on a commercial scale. Being based totally on device mastering algorithms, this system will best increase its efficiency over the years as extra data is placed into it.

5.2 Future Scope

While we couldn't reach out purpose of one hundred% accuracy in fraud detection, we did turn out to be growing a gadget that may, with enough time and records, get very near that intention. As with this sort of task, there's some room for development right here.The very nature of this undertaking lets in for more than one algorithms to be included collectively as modules and their consequences can be combined to boom the accuracy of the very last result. This version can similarly be progressed with the addition of extra algorithms into it.

However, the output of these algorithms needs to be within the equal format because the others. Once that circumstance is glad, the modules are clean to add as performed within the code. This affords a awesome degree of modularity and flexibility to the mission.More room for development can be determined in the dataset. As validated earlier than, the precision of the algorithms increases whilst the size of dataset is elevated. Hence, extra

statistics will clearly make the model more correct in detecting frauds and decrease the quantity of fake positives. However, this calls for authentic support from the banks themselves.

REFERENCES

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

[2] CLIFTON PHUA1, VINCENT LEE1, KATE SMITH1 & ROSS GAYLER2 " A Comprehensive Survey of Data Mining-based Fraud Detection Research" published by School of Business Systems, Faculty of Information Technology, Monash University, Wellington Road, Clayton, Victoria 3800, Australia

[3] "Survey Paper on Credit Card Fraud Detection by Suman" , Research Scholar, GJUS&T Hisar HCE, Sonepat published by International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 3 Issue 3, March 2014

[4] "Research on Credit Card Fraud Detection Model Based on Distance Sum – by Wen-Fang YU and Na Wang" published by 2009 International Joint Conference on Artificial Intelligence

[5] "Credit Card Fraud Detection through Parenclitic Network Analysis By Massimiliano Zanin, Miguel Romance, Regino Criado, and SantiagoMoral" published by Hindawi Complexity Volume 2018, Article ID 5764370, 9 pages

[6] "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy" published by IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 8, AUGUST 2018

[7] "Credit Card Fraud Detection-by Ishu Trivedi, Monika, Mrigya, Mridushi" published by International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 1, January 2016

[8] David J.Wetson,David J.Hand,M Adams,Whitrow and Piotr Jusczak "Plastic Card Fraud Detection using Peer Group Analysis" Springer, Issue 2008.

APPENDICES

[2]:
```
import numpy as np
import pandas as pd
import sklearn
import scipy
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import classification_report,accuracy_score
from sklearn.ensemble import IsolationForest
from sklearn.neighbors import LocalOutlierFactor
from sklearn.svm import OneClassSVM
from pylab import rcParams
rcParams['figure.figsize'] = 14, 8
RANDOM_SEED = 42
LABELS = ["Normal", "Fraud"]
```

[3]:
```
data = pd.read_csv('creditcard.csv',sep=',')
data.head()
```

t[3]:

	Time	V1	V2	V3	V4	V5	V6	V7	V8	V9	...	V21	V22	V23	V24	V25	V26	
0	0.0	-1.359807	-0.072781	2.536347	1.378155	-0.338321	0.462388	0.239599	0.098698	0.363787	...	-0.018307	0.277838	-0.110474	0.066928	0.128539	-0.189115	0.133
1	0.0	1.191857	0.266151	0.166480	0.448154	0.060018	-0.082361	-0.078803	0.085102	-0.255425	...	-0.225775	-0.638672	0.101288	-0.339846	0.167170	0.125895	-0.008
2	1.0	-1.358354	-1.340163	1.773209	0.379780	-0.503198	1.800499	0.791461	0.247676	-1.514654	...	0.247998	0.771679	0.909412	-0.689281	-0.327642	-0.139097	-0.055
3	1.0	-0.966272	-0.185226	1.792993	-0.863291	-0.010309	1.247203	0.237609	0.377436	-1.387024	...	-0.108300	0.005274	-0.190321	-1.175575	0.647376	-0.221929	0.062
4	2.0	-1.158233	0.877737	1.548718	0.403034	-0.407193	0.095921	0.592941	-0.270533	0.817739	...	-0.009431	0.798278	-0.137458	0.141267	-0.206010	0.502292	0.219

5 rows × 31 columns

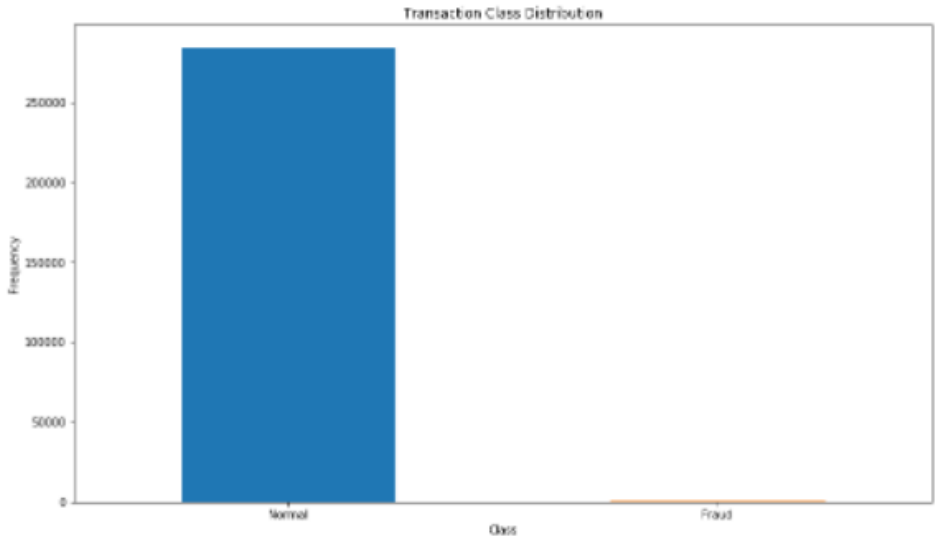
Exploratory Data Analysis

In [5]:
```
data.isnull().values.any()
```

Out[5]:
```
False
```

In [7]:
```
count_classes = pd.value_counts(data['Class'], sort = True)

count_classes.plot(kind = 'bar', rot=0)

plt.title("Transaction Class Distribution")

plt.xticks(range(2), LABELS)

plt.xlabel("Class")

plt.ylabel("Frequency")
```

Out[7]:
```
Text(0, 0.5, 'Frequency')
```



In [9]:
```
## Get the Fraud and the normal dataset

fraud = data[data['Class']==1]

normal = data[data['Class']==0]
```

In [10]:
```
print(fraud.shape,normal.shape)
```

```
(492, 31) (284315, 31)
```

In [12]:
```
## We need to analyze more amount of information from the transaction data
#How different are the amount of money used in different transaction classes?
fraud.Amount.describe()
```

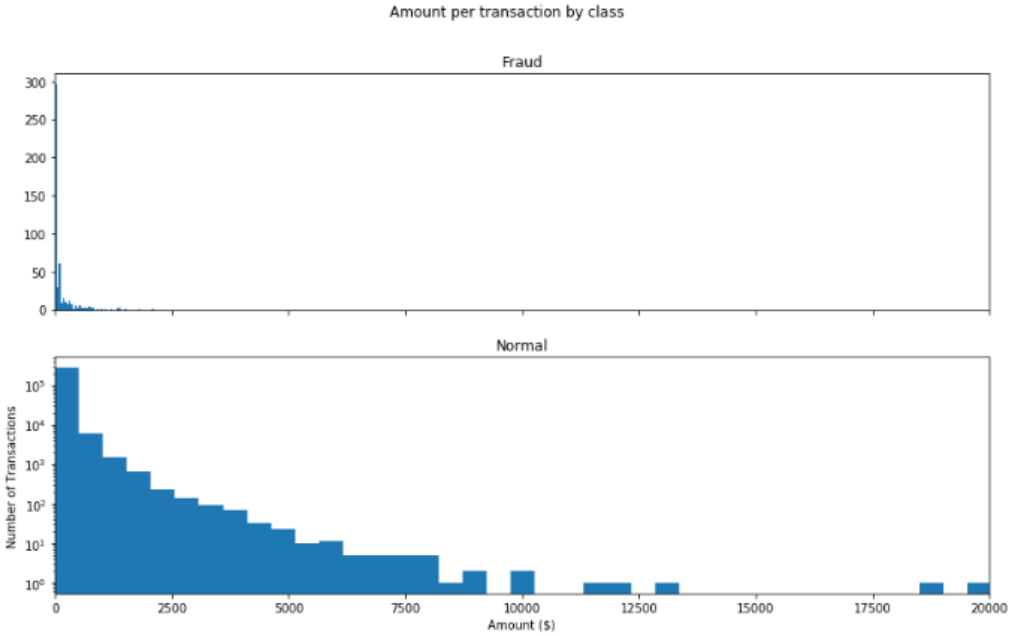
Out[12]:
```
count     492.000000
mean      122.211321
std       256.683288
min         0.000000
25%         1.000000
50%         9.250000
75%       105.890000
max      2125.870000
Name: Amount, dtype: float64
```

In [13]:
```
normal.Amount.describe()
```

Out[13]:
```
count    284315.000000
mean         88.291022
std         250.105092
min           0.000000
25%           5.650000
50%          22.000000
75%          77.050000
max       25691.160000
Name: Amount, dtype: float64
```

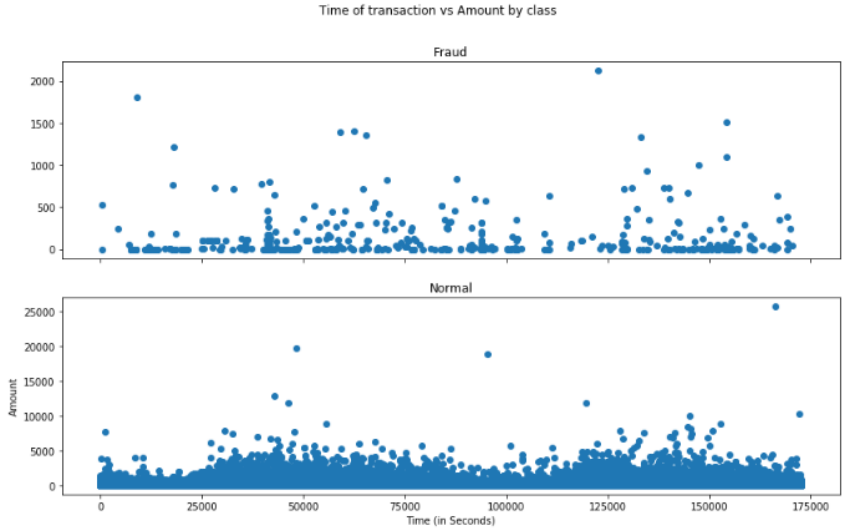
In [15]:

```
f, (ax1, ax2) = plt.subplots(2, 1, sharex=True)
f.suptitle('Amount per transaction by class')
bins = 50
ax1.hist(fraud.Amount, bins = bins)
ax1.set_title('Fraud')
ax2.hist(normal.Amount, bins = bins)
ax2.set_title('Normal')
plt.xlabel('Amount ($)')
plt.ylabel('Number of Transactions')
plt.xlim((0, 20000))
plt.yscale('log')
plt.show();
```



In [16]:

```
# We Will check Do fraudulent transactions occur more often during certain time frame ? Let us find out with a visual representation.

f, (ax1, ax2) = plt.subplots(2, 1, sharex=True)
f.suptitle('Time of transaction vs Amount by class')
ax1.scatter(Fraud.Time, Fraud.Amount)
ax1.set_title('Fraud')
ax2.scatter(Normal.Time, Normal.Amount)
ax2.set_title('Normal')
plt.xlabel('Time (in Seconds)')
plt.ylabel('Amount')
plt.show()
```



In [17]:
```
## Take some sample of the data

data1= data.sample(frac = 0.1,random_state=1)

data1.shape
```

Out[17]:
```
(28481, 31)
```

In [18]:
```
data.shape
```

Out[18]:
```
(284807, 31)
```

In [19]:
```
#Determine the number of fraud and valid transactions in the dataset

Fraud = data1[data1['Class']==1]

Valid = data1[data1['Class']==0]

outlier_fraction = len(Fraud)/float(len(Valid))
```

In [20]:
```
print(outlier_fraction)

print("Fraud Cases : {}".format(len(Fraud)))

print("Valid Cases : {}".format(len(Valid)))
```

```
0.0017234102419808666
Fraud Cases : 49
Valid Cases : 28432
```

In [23]:
```
#Create independent and Dependent Features
columns = data1.columns.tolist()
# Filter the columns to remove data we do not want
columns = [c for c in columns if c not in ["Class"]]
# Store the variable we are predicting
target = "Class"
# Define a random state
state = np.random.RandomState(42)
X = data1[columns]
Y = data1[target]
X_outliers = state.uniform(low=0, high=1, size=(X.shape[0], X.shape[1]))
# Print the shapes of X & Y
print(X.shape)
print(Y.shape)
```

```
(28481, 30)
(28481,)
```

In [24]:

```
##Define the outlier detection methods

classifiers = {
    "Isolation Forest":IsolationForest(n_estimators=100, max_samples=len(X),
                                       contamination=outlier_fraction,random_state=state, verbose=0),
    "Local Outlier Factor":LocalOutlierFactor(n_neighbors=20, algorithm='auto',
                                              leaf_size=30, metric='minkowski',
                                              p=2, metric_params=None, contamination=outlier_fraction),
    "Support Vector Machine":OneClassSVM(kernel='rbf', degree=3, gamma=0.1,nu=0.05,
                                         max_iter=-1, random_state=state)

}
```

In [26]:

```
type(classifiers)
```

Out[26]:

```
dict
```

In [27]:

```
n_outliers = len(Fraud)
for i, (clf_name,clf) in enumerate(classifiers.items()):
    #Fit the data and tag outliers
    if clf_name == "Local Outlier Factor":
        y_pred = clf.fit_predict(X)
        scores_prediction = clf.negative_outlier_factor_
    elif clf_name == "Support Vector Machine":
        clf.fit(X)
        y_pred = clf.predict(X)
    else:
        clf.fit(X)
        scores_prediction = clf.decision_function(X)
        y_pred = clf.predict(X)
    #Reshape the prediction values to 0 for Valid transactions , 1 for Fraud transactions
    y_pred[y_pred == 1] = 0
    y_pred[y_pred == -1] = 1
    n_errors = (y_pred != Y).sum()
    # Run Classification Metrics
    print("{}: {}".format(clf_name,n_errors))
    print("Accuracy Score :")
    print(accuracy_score(Y,y_pred))
    print("Classification Report :")
    print(classification_report(Y,y_pred))
```

```
Isolation Forest: 73
Accuracy Score :
0.9974368877497279
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.26      0.27      0.26        49

   micro avg       1.00      1.00      1.00     28481
   macro avg       0.63      0.63      0.63     28481
weighted avg       1.00      1.00      1.00     28481

Local Outlier Factor: 97
Accuracy Score :
0.9965942207085425
Classification Report :
              precision    recall  f1-score   support

           0       1.00      1.00      1.00     28432
           1       0.02      0.02      0.02        49

   micro avg       1.00      1.00      1.00     28481
   macro avg       0.51      0.51      0.51     28481
weighted avg       1.00      1.00      1.00     28481
```