****

**Athlone Institute of Technology**

**Project Thesis – Academic Year 2019/20**

**Credit Card Fraud Detection**

**in Machine Learning**

**Project Supervisor: Martina Curran**

**Chhaya Sharma**

**A00268860**

**Bachelor of Engineering (honors) Software Engineering**

# **Abstract**

**Credit Card Fraud Detection using Machine Learning**

**By**

**Chhaya Sharma**

In this technical era where whole world is in express train of digitalization, number of online transactions are also increasing rapidly as e-commerce and other online sites are using Online Payment mode which has grown chances of fraud. Fraud is an act of criminal deception. As of now Credit Card transaction has become popular mode of transactions in both online and offline payments which is escalating the chances of fraud. Recent study shows loss of credit card, account takeover, counterfeit card, fraudulent credit applications have been major accounting reasons of credit card fraud.

As enormous data generated from trace logs or from user history can be analyzed to detect patterns and these patterns can help us in detecting future theft, fraud, and user behaviors etc. Anomalies are also called as outliers, novelties, and deviations. Machine Learning offers wide range of supervised and unsupervised algorithm which can be used for recognizing these anomalies. Anomalies detected by this project would help in detecting frauds in credit card statement or trace files.

# **Declaration**

I hereby declare that I am the sole author of this thesis and that all the work presented in it, unless otherwise referenced, is my own. I also declare that this work has not been submitted, in whole or in part, to any other university or college for any degree or qualification.

**Chhaya Sharma**  **Martina Curran**

# **Acknowledgement**

First and foremost, I would like to express my earnest to my project supervisor Martina Curran for her persistent help and guidelines. Her remarkable feedback, motivation and immense knowledge from her research background helped me a lot and kept me motivated from the very beginning of the project. She has always helped me out all the time from research phase till the end of documentation with continuous guidelines. I could not have imagined better supervisor and mentor for my final year project.

I world also like to thank my friends who helped me to overcome from situation where I could not think properly and clearly. They always encouraged me to try new technologies and share their experience of working with similar technologies.

# **Project Summary**

This report will highlight what are different types of Frauds, how they are impacting today’s world. This will also outline what were different rule-based approaches used to address these kinds of frauds and why these approaches failed to respond a desired result. Further we will discuss how Data Science approach can meet the challenges. Also, it will outline what are challenges for Fraud Detection Model and how these challenges are tackled.

In addition, this report will show libraries like **NumPy**, **SciPy**, **Sklearn**, **Matplotlib** etc. This report will also explore what all different libraries were used in building the project. It will show how we used matplotlib function to show correlation metrics. Also, it will how **Isolation** **Forest** and **Local** **Outlier Factor,** and **Support Vector Machine** Algorithm can be used to detect anomalies in a dataset to isolate anomalies and it will show comparison between performance of both.

# **Index**

Table of Contents

[**Abstract** 2](#_Toc38916184)

[**Declaration** 3](#_Toc38916185)

[**Acknowledgement** 4](#_Toc38916186)

[**Project Summary** 5](#_Toc38916187)

[**Index** 6](#_Toc38916188)

[**Introduction** 7](#_Toc38916189)

[**Aims and Objective** 9](#_Toc38916190)

[**Scope** 10](#_Toc38916191)

[**Tools & Technologies** 11](#_Toc38916192)

[**Jupyter Notebook:** 11](#_Toc38916193)

[**Python:** 11](#_Toc38916194)

[**Research Introduction** 12](#_Toc38916195)

[**Types of Fraud** 14](#_Toc38916196)

[**Rule Based Approach** 15](#_Toc38916197)

[**Machine Learning Approach** 16](#_Toc38916198)

[**Two Machine Learning Techniques for Anomaly Detection** 19](#_Toc38916199)

[**Flow chart of the system** 24](#_Toc38916200)

[**Implementation** 25](#_Toc38916201)

[**Environment Setup:** 25](#_Toc38916202)

[**Code Snippets** 26](#_Toc38916203)

[**Conclusion:** 38](#_Toc38916204)

[**References** 39](#_Toc38916205)

# **Introduction**

The purpose of this is to show the implementation of Credit Card Fraud Detection using Machine Learning. This report will describe what is Fraud and how rapid rise of Credit Card Fraud is affecting companies, organizations, and citizens. This report will also show how Credit Card Fraud can happen. What are different types of Fraud. What was rate of Credit card frauds in past decade.

The report will show the limitation of traditional rule-based approach that were used to predict credit card frauds and how these algorithms failed to predict any new kind of anomaly.

Further, report will show why did I chose Machine Learning based project and how is Machine Learning is a big thing in the market, and why did I peculiarly go for Credit Card Fraud Detection project. It will also set out what is classification, Regression, Clustering, and anomaly detection in Machine Learning. Different types of Anomalies in a dataset and how we have used them in our project.

Report will show what different types of problem can be faced with credit card dataset as sometimes data can be very unbalanced. And how Random under-sampling, over-sampling and both can be used to deal with unbalanced data.

This report will also include Tools, environment and algorithms used in project i.e. **Jupyter Notebook** online, Libraries like **Pandas, NumPy, SciPy, Matplotlib** and **Sklearn**. Use of **Anaconda** and different unsupervised algorithm like **Isolation Forest** and **Local Outlier Factor,** and **Support Vector Machine** to get anomalies from dataset. Also, it will describe why all these algorithms have shown this level of precision.

Next, what are Sequential Supervised Learning techniques like Kernel Based Technique (**kNN**, Clustering), **Window Based Technique, Markovian Technique and Hidden Markov Model Based Technique.**

As the dataset we have used in the project is **PCA** dataset, so report will also highlight what is Principal Component Analysis and advantage of using PCA on a dataset. Also, it will show different classes of dataset and relation among the values. Also, it will outline different features of dataset may vary for the legitimate and Fraud transactions.

This report shows the implementation of Credit Card Fraud Detection System based on Anomaly Detection in Machine Learning. This report will touch need of this kind of system. It will show usage and compare different Anomaly Detection Algorithm. This system can help any Credit Card Company to predict customer behavior, suspected threat, and frauds etc.

The layout of this interim report falls into following sections:

* **Section 1: Literature Review**
  + Study of statistic of Credit Card Frauds happening and types of fraud
  + Study and comparison of existing anomaly detection algorithms
  + Selection of a dataset for analysis
* **Section 2: Specification and Implementation**
  + Selecting algorithms to predict abnormal patterns
  + Implementation of the specified algorithms
* **Section 3: Evaluation and Visualization of Results**
  + Understanding various evaluation metrics that can be used for my use-case
  + Using Python libraries such as **matplotlib** to depict and compare results
* **Section 4: Conclusion**
  + This section includes the result of this project
  + Also shows the learning and knowledge I acquired from this project
  + Thesis preparation
* **Section 5: References**
  + This section contains all the links and references that I have conferred during the development of this undertaking and making of this report.

# **Aims and Objective**

This report will trace a Credit Card Fraud detection system based on Machine Learning. The motive of developing this project is to detect fraudulent transactions in a credit card dataset by a credit card company so that customers are not charged for something they did not buy.

In 2016, US approximate 15.4 million people experienced fraud. Fraudsters stole $6 billion from banks last year. According to **Javelin Study & Research,** it takes 40+ days to detect fraud for brick-and-mortar financial institutions.

To overcome these challenges, we need some real-time claim system which can address all these problems with improved accuracy of fraud detection. Machine Learning provides a wide range of algorithms to predict these frauds with its libraries and algorithms. My aim is to develop a system which can respond to these frauds in real-time and can respond to future frauds with more accuracy.

This system aims to predict anomalies in a credit card dataset. Many different approaches have been identified to detect patterns in data which can be supervised or unsupervised. Clustering in machine learning is also a technique to group similar items and recognize similarities in types of data and their coupling with other similar data group.

Unsupervised Machine Learning algorithms can detect anomalies by clustering unknown or new data sample in a different cluster. Anomalies detected by this system can be used by a Credit Card Company for detecting frauds inn credit card statements or trace files.

Objective:

* To decide which Machine Learning tools to use.
* To learn different Python libraries and their use-case
  + Pandas
  + NumPy
  + Scikit-learn
  + Matplotlib
  + SciPy
* To decide which Anomaly Detection algorithm to use.
* To reduce the dimensions from the dataset.
* To determine correlation in variables using heatmaps from matplotlib
* To implement chosen algorithms and compare the results.
* To visualize the results of all the algorithms

# **Scope**

The scope of this project is very large as it highlights the benefits of using Machine Learning in real-life application. This system aims to detecting fraud from Streaming Transaction data, with a goal of analyzing past behavior of the customers and extract the transaction behavioral patterns. Where cardholders with similar transactions details are clustered together in a group and are distinguished from those with different transaction amount. Then 3 classifiers are trained over the groups separately and then classifiers with most accuracy can be used for future prediction.

Not only this system provides a more trusted way to determine anomalies in a credit card dataset, but it also gives real-time results which was not possible with traditional rule-based approaches. Thus, it can be used by any Credit Card Company and Financial Institution, Bank to predict their customer’s transactional behavior and can be utilized to further to aware them of any possible threat. Amount of accuracy in predicting these patterns is crucial as it is one of the biggest challenges faced by Credit Card Companies and false results sometimes can make them loose their customers.

To sum it up this project has immense scope and its advantages cannot be overlooked by both banks and financial institutions. As a reliable system attracts more customer. There is no bank or company which would not want to secure its transactions details. Considering frauds happened in past decade this System can serve as boon for any company.

# **Tools & Technologies**

Since Credit Card Fraud Detection is entirely Machine Learning based project so the challenge was to choose an environment which can provide everything in one place if possible. I chose Jupyter Notebook which is an open-source web-based application that allows us to create and share the code and documents.

### **Jupyter Notebook:**

1. User can write code, run it, see the result, and visualize the outcome and can see the result at the same terminal. All these amenities make it a popular tool for performing end-to-end data-science workflow I.e. Data Cleaning, Statistical modelling, building, and training Machine Learning Models, Visualizing the data etc.
2. User can write code in independent cells and run them individually and independent which allows user to see the execution of specific block.
3. Jupyter Notebook is incredibly flexible, interactive tool for Data Science and Machine Learning.
4. It allows to run Python, R, and SQL etc.

### **Python:**

Next, which language to choose. For this project **Python** suited the most as Python offers a wide range of library with which can curtail code into few lines only and hence makes the development easy and less time consuming. Here are few of libraries offered by Python:

1. **Pandas:**

This library provides function to deal with different kind of data resources. For example, reading CSV files.

1. **NumPy**

NumPy is used to deal with numbers and arrays.

1. **SciPy**

SciPy is fundamental library for computing.

1. **Scikit-learn**

Scikit supports various Machine Learning algorithms. In this project 3 Anomaly Detection Algorithms are used

1. **Matplotlib & Seaborn**

These libraries provide a wide range of graphs and maps to visualize the dataset which plays a vital role in understanding dataset and correlation among values.

# **Research Introduction**

Fraud is an act of deception used to illegally deprive another person, his money, property, or legal rights. Here are few major Credit Card Frauds evidence in recent years which made highlights

* More than 3.2 million fraud cases were reported to Federal Trade Commission (FCI) in 2019, Identity theft is 20.33% of all cases and was most common type of fraud.
* Referring to results of over 270,000 reports, credit card fraud accounted for the most common type of identity theft last year and more than double from 2017 to 2019.
* Also, in 2018, $24.26 Billion was lost due to payment card fraud worldwide.
* A close up of a map

  Description automatically generatedThe Capital One Cyber incident was biggest data breach of 2019 as it disclosed the personal data of approximately 100 million customers in US.

#### **How Credit Card Fraud Can Happen?**

* The account number can also be taken, along with pin and security code.
* Scammers can use **skimmers on point-of-sale system**, to get user information and use it to get a duplicate card. This duplicate card can be used anywhere and can be dangerous if card has lots of balance.
* If your credit card is misplaced or taken, it can be used for unauthorised in person or online transactions.
* Hackers can obtain your info using unsecure or imposter website and data breaches, then they can use that info to make purchase without needing the card present.

***Credit card not present fraud is now 81% more likely than point-of-sale fraud according to Javelin Strategy***

A close up of a logo

Description automatically generated

## **Types of Fraud**

#### **Payment Card Fraud**

* This type of fraud happens when someone uses stolen or counterfeit card to make some purchases or cash withdrawals.
* This type of fraud can also happen if someone uses the card details over phone or internet.

#### **Invoice Redirection Fraud**

* This type of fraud includes criminals contacting sellers or business owners by email or by phone calls.
* Generally, criminal pretends as trusted goods supplier or service provider and change the account details of legit supplier due to which on next transaction payment will be transferred to criminal’s account.
* These kinds of frauds are dangerous for business as now they also have an outstanding bill to pay along with loss.

#### **CEO Fraud:**

* In such frauds junior employee of companies receive fake emails from criminals pretending to be CEO of the company.
* Email generally stats that an urgent payment is pending and needs to be immediately transferred. By the time fraud is detected money is already gone into scammers account.

#### **Email Fraud: Phishing**

* This type of fraud evident by criminals making contact and can take several forms.
* It pretends to be coming from reputable company however with on click on attachment may download malicious software into computer
* These software keeps a track of all the activities happening onto the PC. Hence, they also get the account details of the user.
* Both individuals and companies can be victims of this type of crime.

#### **Phone Fraud: Vishing or Smishing**

* This type of fraud happens by criminals contacting by either phone (Vishing) or text (Smishing) pretending to be from user’s bank, credit card issuer company or any government services.
* During this conversation they try to trick the customer to get their account credentials.
* They use this information against customer on other platforms as well.

#### **Advance Fee Fraud**

* This type of frauds involves scammers asking victims for advance payment of some goods or services for which victim usually does transactions which makes them believe the scammers as legit suppliers.
* Rental Fraud, Romance Fraud are some of major example of Advance Fee Fraud

## **Rule Based Approach**

These approaches of dealing with frauds are the oldest and time-consuming. These approaches include certain rules and actions defined or written by Data Analysts. For example, if a transaction is 10-12 times than average transaction amount of that customer than a notification is given or if wrong pin is entered more than 3 times block the account.

All these rules are flexible and making any variations into these rules requires lot of time and efforts. These rules might not respond to any new anomaly. In short Rule based approach are:

* Algorithms written by Fraud Analysts
* Based on strict rules
* Changes for detecting new frauds were made manually.

#### **Drawbacks of Rule Based Approach**

* **Cannot recognize hidden patterns**
* **More likely to have False-Positive**

By rules-based approach chances of False-positive were high which sometimes may result in dishonor of the customer and eventually resulted in loosing their interest from respective credit card company or bank.

* **Cannot respond to new situations, not trained, or explicitly** **programmed**

Rule based approach works on predefined set of rules and these rules often fail in predicting new changes. These rigid approaches are not adaptable for an environment which is continuously evolving, and new kind of data set is resulting. So, these approaches are not right choice for such data.

* **Expansive for bigger data set**

As customer base increases it becomes very expansive to maintain that data and are proved to be less scalable.

* **Time-proven approaches**

Rule based approaches require lots of time in a complex and highly regulated industries such as finance services. Inability to access the real-time insights or identifying emerging fraud patterns has restricted financial service providers which made them less effective and efficient.

## **Machine Learning Approach**

To leverage the vast amount of data collected from online transactions and model it in a wat that allows us to flag or predict fraud in future transactions. To meet the challenges faced by Rule-based approach Machine Learning and Deep Learning approaches are obvious solution.

Machine Learning comprises a large collection of algorithms and techniques used in classification, regression, clustering, or anomaly detection.

* **Classifications**

Classification in machine learning is a supervised machine learning approach in which the computer program learns the data given to it and make result set or observations.

* **Regression**

Regression is a ML algorithm which can be trained to predict real time outputs such as temperature, stock price etc. In training phase, the hidden parameters are optimized w.r.t the inputs presented in training the model.

* **Clustering**

Clustering is technique of dividing the data with similar items into same groups. Clustering is unsupervised Machine Learning approach.

* **Anomaly Detection**

Anomalies detection is identifying of rare items, events, or observations from the rest of dataset. It can be used in various domains, such as intrusion detection, fraud detection, fault detection, detecting events in sensors. It is often used in preprocessing to remove anomalous data from dataset.

Now, Anomalies fall into 3 main categories as:

1. Global Outliers
2. Contextual Outliers
3. Collective Outliers
4. **Global Outliers:**

These kinds of outliers exist far from outside of dataset. As it is because of one point being different from other which makes it appear like a point is anomaly to rest of the dataset, it is also called as Point Anomaly.

A close up of a logo

Description automatically generated

1. **Contextual Outliers**

These anomalies significantly deviate from other data points that are there in the same context. In this case one thing that is anomaly to one dataset might not be anomaly if referred in a different context.

These types of anomalies are common in time series data because those datasets are records of specific quantities in given period. The value exists in global expectations but may appear anomalous within certain seasonal data patterns.

A close up of a mans face

Description automatically generated

1. **Collective Outliers**

When a subset of data points collectively acts as anomaly to the entire dataset, then all those values together are called collective outliers. In this type of anomaly individual value is not anomaly globally or contextually.

A picture containing map

Description automatically generatedThese types of anomalies are observed when examining distinct time series together. As individual behavior may not deviate from normal range in a specific time series dataset. As soon as combined with another time series dataset, results become clearer.

## **Two Machine Learning Techniques for Anomaly Detection**

* **Supervised Machine Learning for Anomaly Detection**

This type of method requires a labeled training data which both normal and anomalous data values for creating the predictive model.

I.e. Supervised Algorithms learns from the data which is already trained this helps in predicting unforeseen data which can be used in various real-time applications such as Weather forecasting, Time of the day, fraud detection etc. Successfully building, scaling, and deploying accurate supervised machine learning data science model which requires a team of highly skilled data scientists. Moreover, these models need to be rebuilt to assure that all insights given are correct.

A screenshot of a cell phone

Description automatically generated

* + Theoretically, supervised approach is believed to provide better results/ prediction than unsupervised machine learning approaches.
  + Some popular supervised algorithms are:
    - Supervised Neural Networks
    - Parameterization of training model
    - Support Vector Machine
    - K-nearest Neighbor (k-NN)

K Nearest Neighbor is standard nonparametric techniques which is used for supervised learning or anomaly detection.

* **Unsupervised Machine Learning for Anomaly Detection**

Unsupervised Machine Learning algorithms does not require to train the dataset. These algorithms require 2 basic assumptions:

* + They assume most of the network connection are normal traffic.
  + Malicious traffic is statistically different from normal traffic.

Unsupervised learning allows the model to work on its own to discover information. Unsupervised learning algorithm allows to perform more complex processing tasks compared to Supervised learning,

Because of these two assumptions data groups of similar instances that appear frequently are treated as normal traffic and those date instances that appear rarely are malicious.

Most used Unsupervised ML algorithms are:

* + Self-organizing maps (SOM)
  + K-means
  + C-means
  + Isolation Forest Algorithm
  + Local Outlier Factor
  + Expectations-maximization meta-algorithm (EM)
  + Adaptive Response Theory (ART)
  + One-class support vector Machine

Main objective of the SOM is to reduce the dimension of data visualization.

Benefits of Unsupervised Learning:

* With this technique, all kind of unknown patterns can be detected.
* It helps in finding features which can be used for categorization.
* It is easier to get unlabeled data than labelled data which needs manual intervention.

In this project, following Python libraries will be used:

* **Pandas:**

Pandas is a Python packages which is most powerful and agile open source data analysis or manipulation tool available. Pandas can be used for different kinds of data as:

* + Tabular data with miscellaneous data like in SQL or excel spreadsheet.
  + Arbitrary matrix data with rows and columns
  + Ordered and unordered time-series data.
  + Any other form of observational or statistical data sets.

Two fundamental data structures of Pandas:

1. Series (1 Dimensional)
2. Data Frame (2 Dimensional)

These data structure handle vast majority of the data and typical use case in finance, statistics, and Social science.

Few features of Pandas:

* Easy handling of missing data (represented as NaN)
* Columns can be inserted and deleted from Data-frame and higher dimensional objects.
* Time-series specific functionality I.e. Data range generation and frequency conversion, moving window statistics, and data shifting etc.
* Hierarchical labelling of data axes (Multiple labels per ticks)
* Provides robust tools for loading flat files like (CSV and delimited), Excel files, database, and saving/loading data from the ultrafast HDF5 format
* Intelligent label-based fancy indexing, slicing, and subsetting.
* All these features Pandas makes it a powerful package to deal with data-resources. Pandas is helpful for data-scientists in working at various data stages as cleaning data, analyzing/modelling it then organizing the results of the analysis.
* **NumPy**

NumPy is an open-source elementary package for specific computing with Python. It provides following

1. N-dimensional array object
2. Tools for integrating C/C++ and Fortran Code
3. Sophisticated functions

Additional to scientific usage, NumPy can be efficient multi-dimensional container of generic data. Datatypes can be defined arbitrary which allows NumPy to integrate seamlessly and speedily integrate with variety of databases.

Using NumPy increases performance because of its data structure perform better in:

* + **Size:** NumPy data structure takes less space.
  + **Performance:** These are faster than **lists** in Python and have a need of speed.
  + **Functionality:** Both NumPy and SciPy have built in linear algebra operations for optimized functions.
  + **Memory:** Using less memory consumption and better runtime behavior.
  + NumPy is designed to have all the capabilities of Numeric and Numarray and some more than both.
* **Matplotlib**

Matplotlib offers a wide range of maps and graphs to help in visualizing data. Data visualization is representing data in pictorial or graphical format to visualize the data patterns which helps in recognizing new trends and patterns.

The main benefits of data visualization:

* It simplifies the complex quantitative information
* It helps in understanding and identifying new areas and patterns that can draw some conclusions.
* It has high-quality graphics and plots to visualize and print for a wide range of graphs such as histogram, bar chart, pie chart, and heat maps.
* Its integration with Jupyter gives the developer more flexibility rather than struggling with compatibility.
* It has large community support and cross-platform as it is open source tool.

**Scikit-learn**

* It is simple and powerful tools for predictive data analysis
* Built on NumPy, SciPy and Matplotlib
* It is also open source and commercially usable tool
* It provides a wide range of algorithm that can be used for both supervised machine learning and unsupervised machine learning.
* **Classification:** Identifying which category an object belongs to
  + **Applications**: Spam Detection, Image Recognition
  + **Algorithms**: SVM, nearest neighbors, and random forest etc.
* **Regression:** Predicting a continuous values attribute associated with an object
  + **Applications**: Drug Response, Stock Prices
  + **Algorithms**: SVR, nearest neighbors, random forest etc.
* **Clustering:** Automatic grouping of similar objects into sets
  + **Applications**: Customer Segmentation, Grouping Experiment Outcomes
  + **Algorithms**: k-means, special clustering, mean-shift etc.
* **Dimensionality Reduction:** Reducing random number variable
  + **Applications**: Visualization, Increases efficiency
  + **Algorithms**: k-means, feature selection, non-negative matrix etc.
* **Processing:** Feature extraction and normalization
  + **Applications**: Transformation of input data for use with machine learning, like text.
  + **Algorithms**: preprocessing, feature extraction
* **Model selection:** Comparing, validating and choosing parameters and models
  + **Applications:** Improved accuracy via parameter tuning
  + **Algorithms:** grid search, cross validation, metrics

# **Flow chart of the system**

A close up of a sign

Description automatically generated

Above flowchart explains the flow of Credit Card detection system in a broad picture. Implementation of this system included following major steps:

1. Importing Dataset
2. Exploring Dataset
3. Visualization of dataset
4. Marking Fraud and valid transactions
5. Creating dependent and independent features to apply model
6. Applying algorithms like Isolation forest, Local Outlier Factor and Support Vector Machine
7. Comparing the results from different algorithms.

# **Implementation**

## **Environment Setup:**

Environment setup for this system development requires only few basic tools:

* Anaconda
* Python
* Jupyter Notebook

Anaconda is a powerful distribution of Python mainly used for Scientific purpose. It makes it easy to install a complete environment for scientific computation by providing all types of required modules, packages, an IDE and a package manager for maintaining that working environment which is its own not **pip.**

Few advantages of Anaconda over other distributions which makes it a popular choice of data scientists (*Canopy, Scientific Python Package & Analysis Environment, Python-xy.Github.io by Python-xy*)

* conda is a package manager provided by Anaconda which acts ad both package manager and a virtual environment.
* It offers a stripped-down installation called Miniconda, consisting of Python, conda, and other required packages and one can install these packages as per need either into default or virtual environment.
* Conda manages to install package that require compilation of C modules

I have used Jupyter Notebook online because:

1. By using this, server was already provided no need of separately having it.
2. With Jupyter user can write, run the code in separate cells which are independent of execution of each other.
3. Also, User can visualize datasets and outcome at same place.
4. All these features made it best suited for this project.

## **Code Snippets**

A screenshot of a cell phone

Description automatically generated**Importing required libraries**

**Importing data set or CSV file**

Now, this dataset was PCA transformed which means hiding sensitive information to ensure confidentiality as it is crucial data of 2 days transactions of a bank in Europe as a result of which maximum attributes as account number, name, age etc. are kept hidden (V-features). Only time, amount and class are visible.

This dataset contains 31 columns. But in this project, we will be having only time, amount and class to work on. Let us understand these 3 columns first:

**Time:** Time of transaction which will be required to see how frequently fraudulent transaction were happening

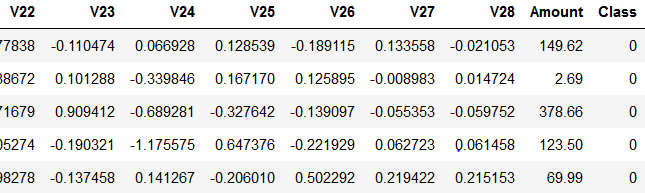
**Amount:** This column shows for how amount transaction was done.

**Class:** This column is most important as it shows if a particular transaction is fraudulent or legitimate transaction, it has only 2 values possible I.e. 1 or 0

**1 indicates FRAUD** transaction

**0 indicates LEGITIMATE** transaction

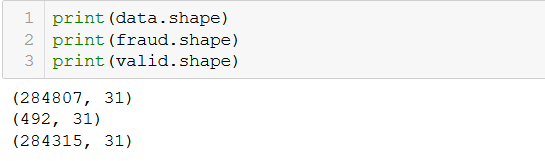
**Head of the dataset** (last few columns to show amount and class)



Labelling Class with value 1 as Fraud transaction and those with value 0 as valid transaction

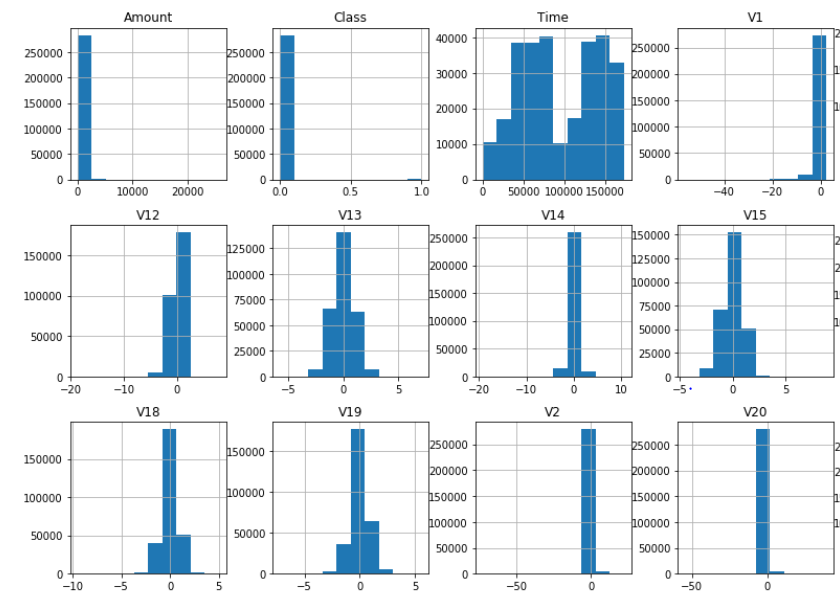


Printing the shape of data, fraud and legit transactions (Total no. of transactions)

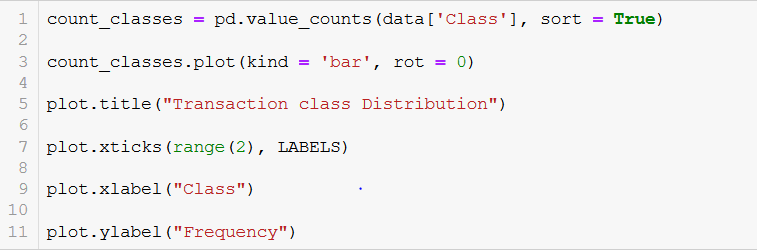


Plotting histogram to visualize the dataset

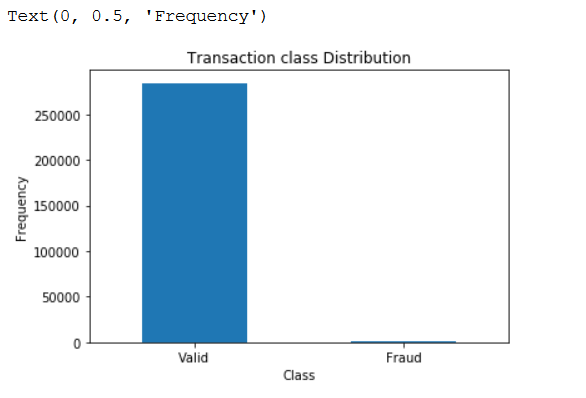




As to predict fraudulent and valid transaction we need to visualize Class column, so for that I have plotted a separate bar chart using matplotlib showing class column vs Frequency



Here is the resultant bar chart



Observation from chart is **Data is highly imbalance** which means there are very less fraud transactions as compare to those of legitimate transactions as there are total of 284315 legit transaction while there were only 492 fraudulent transaction giving a big difference in dataset. This is most common problem faced in Machine Learning. To avoid this imbalance dataset problem few approaches are used as mentioned below:

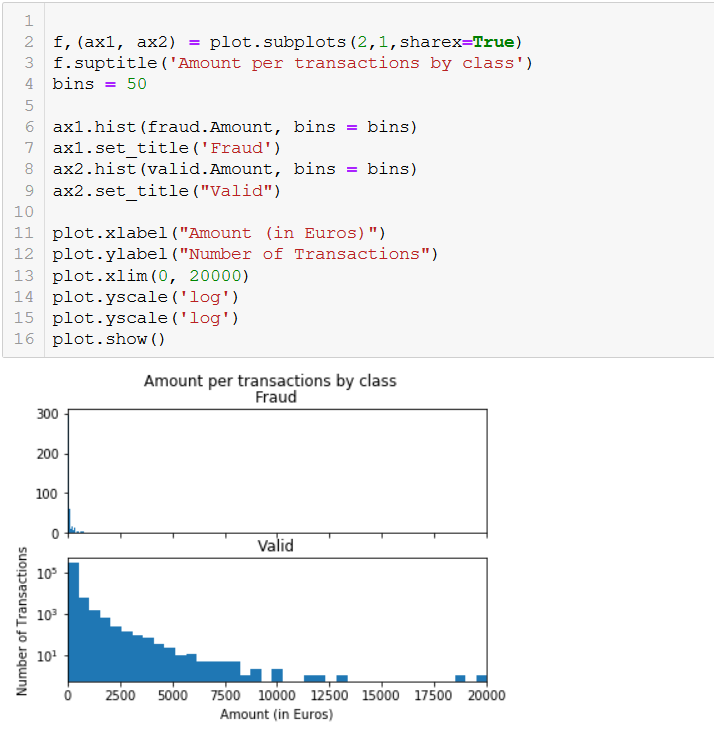
**Random Over-Sampling:**  In this sampling method, the class which has less result values are over-sampled so that an almost equal distribution is achieved.

**Random Under-sampling:** In this method, the majority class is under sampled I.e. some of the transaction are removed till a desired distribution is achieved.

**Both:** This includes a balance use of both Random over-sampling and random under-sampling

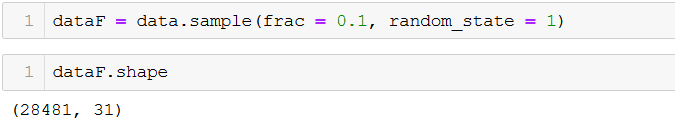
Although this credit card system data has highly unbalanced data still Isolation Forest and Local Outlier Factor has shown amazing results and there was no stiff requirement of handling the unbalanced dataset separately which shows a big advantage of both the algorithms.

To see for how much amount frauds were occurring, by using matplotlib again, I showed relevance of Number of transactions to amount for both fraud and valid transaction using histogram

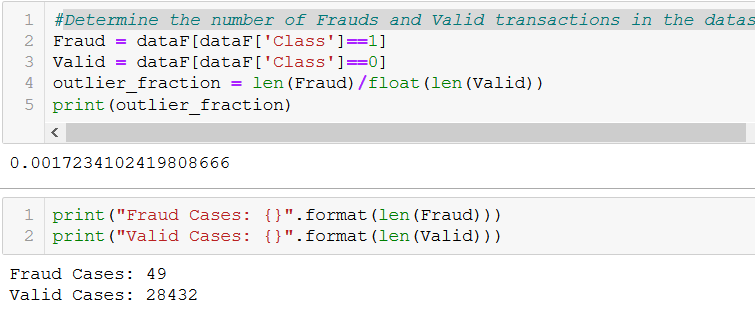


From this histogram we can conclude that fraudulent transactions were for small amount while legitimate transactions can be seen for big amount as well.

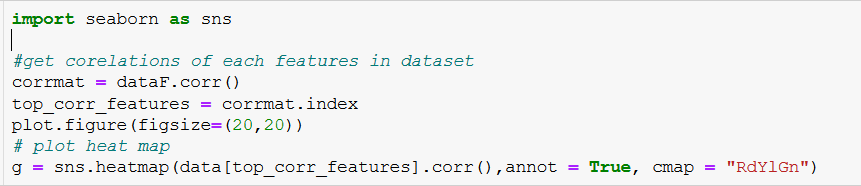
Since this dataset has almost twenty-eight thousand transactions which would require lots of preprocessing, so to avoid I proceed with a fraction of data as 10% of original dataset.

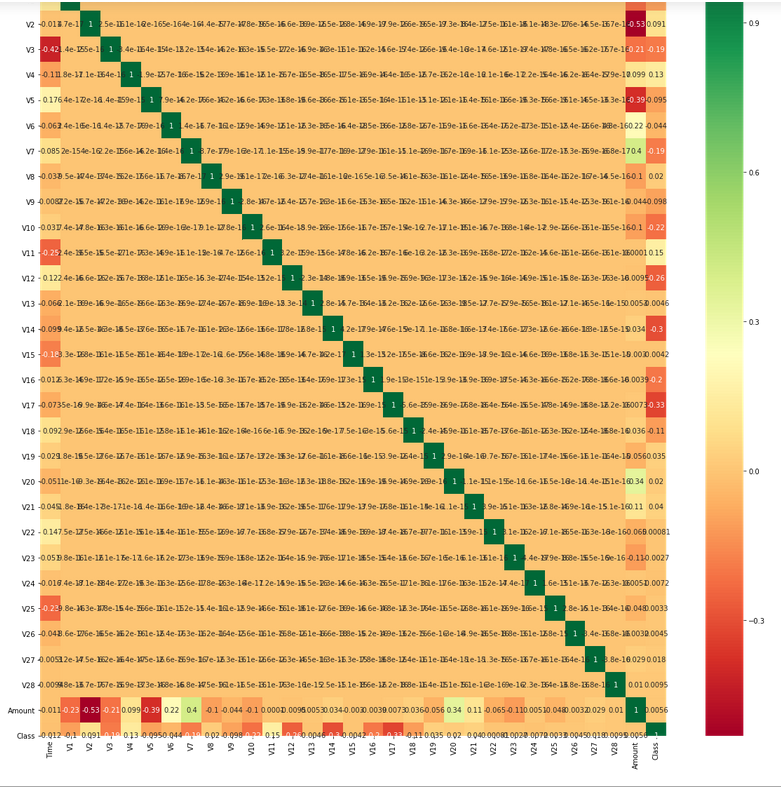


To determine the number of Frauds and Valid transactions in the dataset & printing No of Fraud and legitimate transactions in resulted data sample.



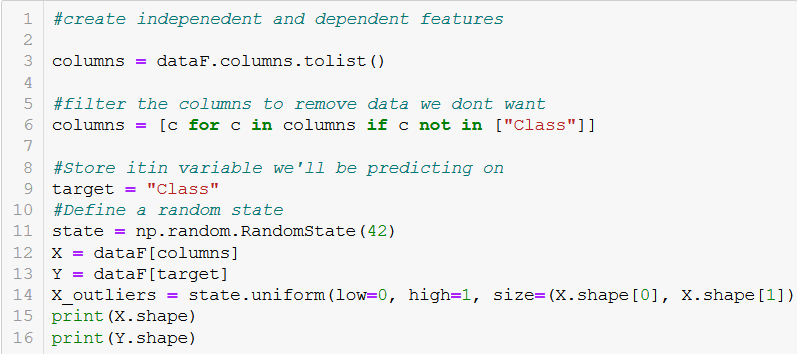
To better visualize the whole dataset at once and see the correlation among the values, I have also used heatmap from seaborn



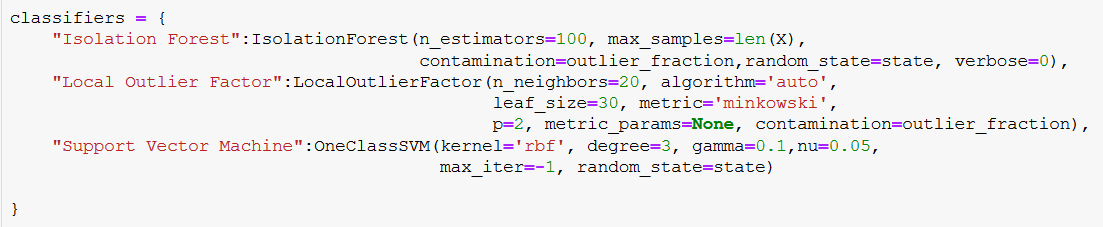


* From above heatmap we can understand that in Class there are lots of values have are close to zero which signifies that these are legitimate transactions
* There is not strong correlation into values I.e. they are quite different from each other
* There is not any strong correlation for both time and amount so from those columns one cannot predict whether these are fraudulent transactions or not

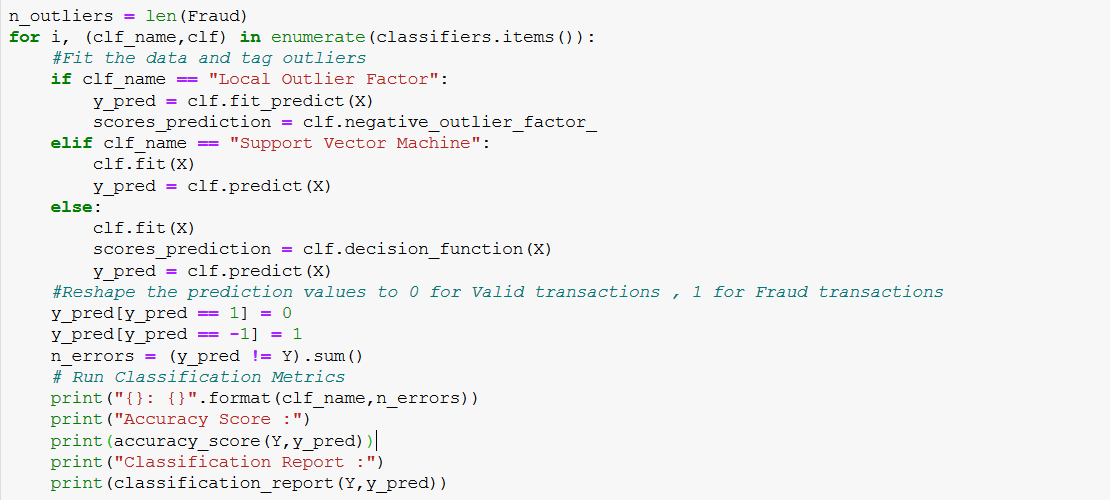
Separating dependent and independent features:



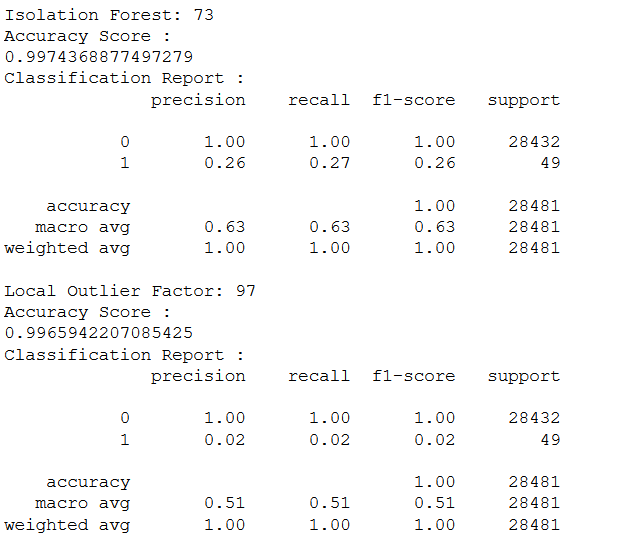
Following three classifiers were used



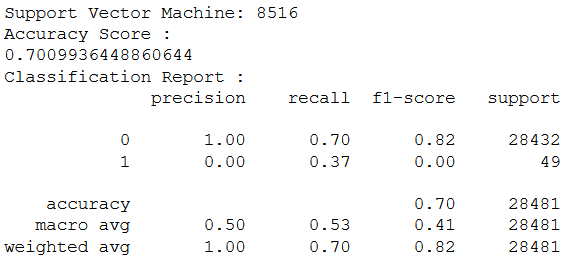
Now, the last step was to fit the data and tag outliers followed by running the classification metrics as shown in snippet



**Results shown by Isolation Forest and Local Outlier Factor algorithms are**



**By Support Vector Machine**



Now from above results we can see that Isolation Forest algorithm has worked best here giving **accuracy score of 0.99743 with total of 73 anomalies** or fraud transaction which is close to real fraudulent data I.e. 49. Hence, it can be used for future predictions in similar applications.

Similarly, Local Outlier Factor has also worked well but it has given lesser accuracy score 0.99659 or 97 Fraud Transaction.

But, SVM did not work good here as it gave 8516 anomalies I.e. accuracy score of 0.7009 which shows its not a good alternative to go for when data is highly imbalance.

**Why did Isolation Forest Algorithm work well?**

As Isolation forest creates a decision tree which means outliers are getting separated in the beginning itself due to large difference in values hence it required less condition to separate them from rest of the data.

# **Conclusion:**

There are so many things I got to learn from this undertaking which primarily includes learning new technologies, exploring of Identity thefts happening in today’s world and their impact on economy, people, companies, and their customer base.

Before starting with this project, I had no knowledge and experience of Machine Learning, but before starting with machine learning it also required some knowledge of Python libraries as Pandas, NumPy, Matplotlib and Scikit etc. so learning that became a new milestone for me which was quite a bit challenging in the beginning but after working with Pandas and NumPy things become interesting. While working at this project, I also got to explore some more areas of machine learning which shows Machine Learning is a next big thing and if used it can make life more convenient.

Now Credit Card Frauds is the biggest identity since 2018 and it is affecting dangerously the customers and companies equally. Also, there are many techniques used to achieve so and if in any case everything goes as per planned by the criminals then it can do so much harm to the company by the time they will get to know about that as Yahoo Data Breach 2014, the biggest information breach in history, in which more than 2 Billion its user accounts were compromised which in result became the threat for American citizens and government employees data was at risk. So suitable measures should be taken in advance to ensure that there is no hack, theft and intrusion etc. and if it happens suitable security steps should be taken.

Also, these Credit Card Fraud are proving as big lose because. In 2018, $24.26 Billion was lost due to payment card fraud worldwide. United state has most credit card fraud in last year. Also, what makes it worse is that 69% of frauds starts with customer getting call or email which appears from a bank or their company.

So, identifying these frauds and their pattern is very important so they can be used to detect and avoid any future mishappening.

**References**

<https://shiftprocessing.com/credit-card-fraud-statistics/>

<https://www.garda.ie/en/Crime/Fraud/What-are-the-6-most-common-types-of-fraud-in-Ireland-today-and-how-to-avoid-becoming-a-victim-.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.svm.OneClassSVM.html>

<https://pandas.pydata.org/>

<https://numpy.org/>

<https://www.geeksforgeeks.org/machine-learning/>