Online Credit Card Fraud Detection System

Team Members

K. Satya Ishyanth-700735513

B. Anil Kumar-700731647

G. Pranav-700741488

Goals and Objectives:

Online transactions are fast rising, and credit cards will be used mostly. Loss of physical credit cards or loss of credit card information will result in a substantially higher payment. The hackers are there to commit fraud against people. As a result, there was a requirement to detect fraudulent transactions and safeguard online credit card transactions. To analyse this problem and combat credit card theft, we created a new system called "Online credit card fraud detection and prevention system" that employs machine learning. We tested the system's accuracy by experimenting with several algorithms. This system use the random forest algorithm to determine if a transaction is legitimate or fraudulent.

Motivation and Significance

Many businesses are expanding rapidly over the world at the moment. Companies strive to deliver the greatest services to their clients. Companies process massive amounts of data on a daily basis for this reason. This data also includes the clients' personal and financial information. As a result, firms must store data in order to handle it, and data security is critical. If this data is not secured, it may be exploited by other companies or, in the worst-case situation, stolen. In rare circumstances, financial information is taken and utilized for fraudulent transactions, causing harm to the parties involved.

Today, online buying has become a widespread daily purchase habit. The perpetrators are engaging in malicious operations such as Trojan and spoofing. When criminals steal cardholder information, the increase in fraud incidents has become a severe issue. Credit card fraud detection is an important subject that has piqued the interest of the machine learning and computational intelligence fields, where a plethora of automation solutions have been presented. Data is available all around the world, and businesses of all sizes are loading information with great volume, variety, speed, and value. This data is derived from a variety of sources, including social media followers, likes, and comments, as well as user purchasing habits. All of this data is utilized to analyze and visualize hidden data patterns.

Objective

The goal is to distinguish and precisely recognize erroneous extortion recognition. There are several methods for differentiating credit cards. misrepresentation, such as neural networks, genetic algorithms, and k-means clustering. The cost of consistent deception in the economy is more than $4 trillion globally. The solution consists in relying on cutting-edge investigation and performing extensive data stockpiling skills that aid the utilization of computerized reasoning (AI) and AI (ML) approaches to deal with staying one step ahead of lawbreakers. ML capabilities, misrepresentation, and consistency teams can focus their efforts on more sophisticated extortion concerns.

Our Objective is to Implementing below Algorithm for fraud detection

- **Logistic Regression**
- **SVM (Support Vector Machine)**
- **DT(Decision-tree)**

Related Work (Background)

There are significant financial losses as a result of credit card fraud incidents. Criminals utilize Trojan and phishing technologies to steal credit card information from others. card. As a result, the fraud detection approach is critical.

Because of fraud detection methods, we can detect fraud when a criminal uses a false card to defraud a consumer. Two types of random forest algorithms are employed in this paper to train the behavior feature of normal and fraudulent transactions.

The purpose of data analytics is to discover hidden patterns and use them to make informed decisions in a range of scenarios. The publicly available datasets on credit card fraud are highly mismatched. We analyze the most essential variables that may contribute to improved credit card fraudulent transaction detection accuracy.

Theft of sensitive credit card information or physical theft of a credit card is considered credit card fraud. For credit card identification, there are several machine learning algorithms available.

Multiple algorithms are utilized in this study to identify whether a transaction is fraudulent or not. This study made use of a credit card fraud detection dataset.

Oversampling was accomplished using the "Synthetic Minority Over Sampling Technique (SMOTE)." The dataset was split into two parts: training and testing data. The methods used in the study included Logistic Regression, Random Forest, Nave Bayes, and Multilayer Perceptron. The results indicate that each algorithm can detect credit card theft with high accuracy.

Survey

Techniques for Detecting Credit Fraud Currently in Use

This subsection focuses on the analysis of some reliable data mining methods applied specifically to the data-rich areas of insurance, credit cards, and telecommunications credit fraud detection in order to detect credit fraud in the data-rich areas of insurance, credit cards, and telecommunications.

To include a few of them Each technique and its uses are briefly described. The research contrasts Bayesian Belief Networks (BBN) with Artificial Neural Networks (ANN) (ANN). The STAGE algorithm for BBNs and the BP algorithm for ANNs are utilized in credit fraud detection. The results show that while BBNs are more accurate and faster to train, they are slower to deploy. The same as in equation [1,] but applied to new instances. Data from actual credit cards

Dataset

Dataset Link: https://www.kaggle.com/datasets/kartik2112/fraud-detection

Motive for a Dataset!

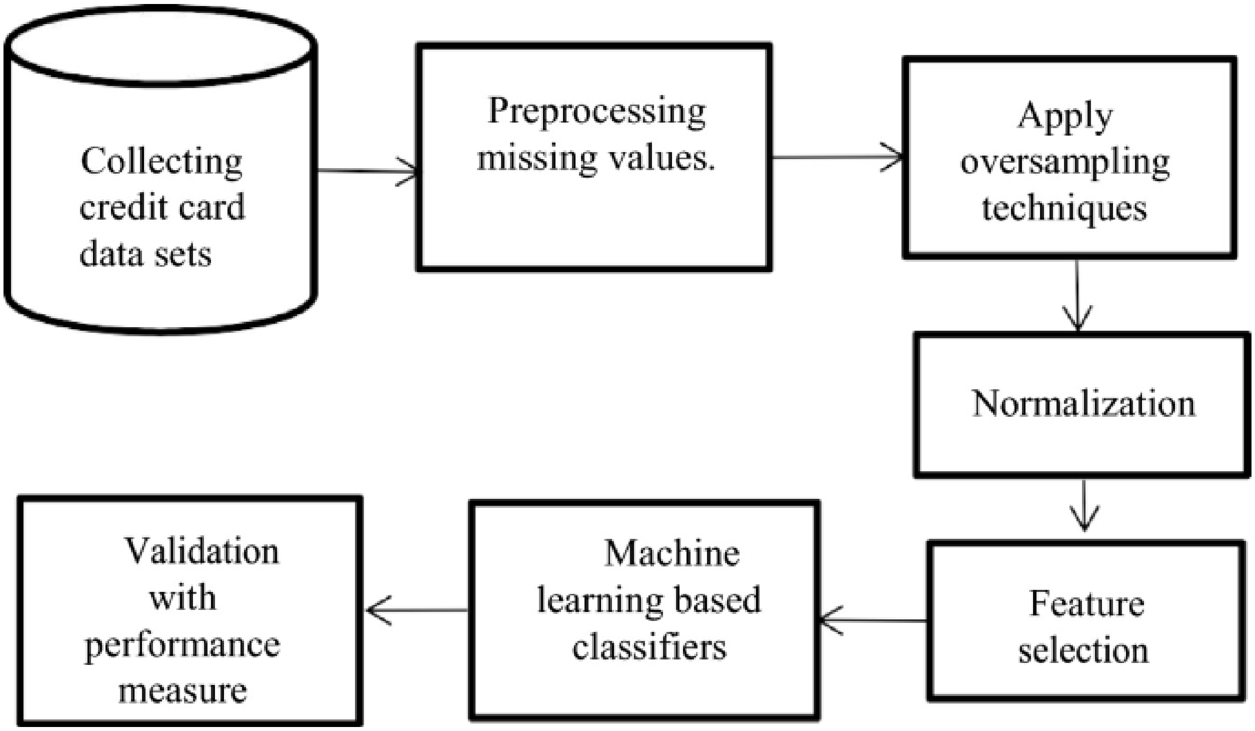
This data collection has been published in order to gain insights into Credit Card Defaulters based on the relevant attributes! Inside? In the Application Data Set, we have attributes such as Income Total, AMTAPPLICATION, AMT CREDIT, and around 122 Columns. The fascinating part is that if you want to see patterns and variances, we can also leverage the PREVIOUS APPLICATION data set to gain more insights.

Inspiration

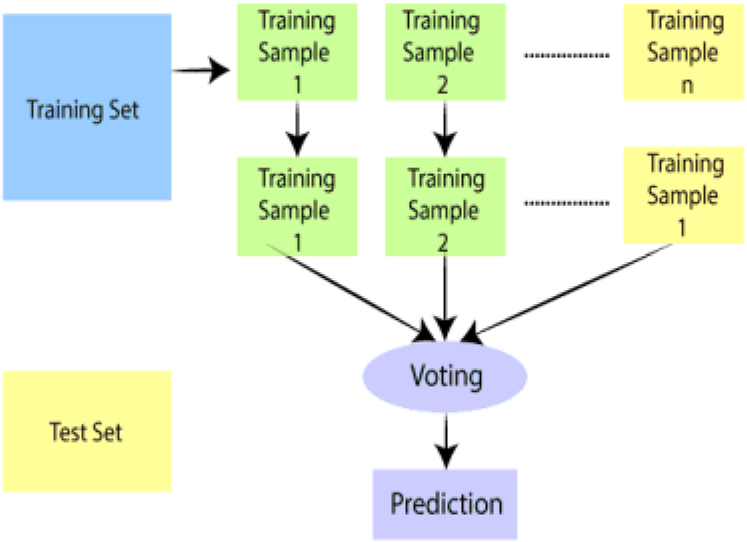
We accepted this data set as our homework and tried our hardest to complete the EDA to the best of our abilities!

Detail design of Features

Architecture


Collecting credit card data sets
Preprocessing missing values.
Apply oversampling techniques
Normalization
Feature selection
Machine learning based classifiers
Validation with performance measure

Workflow for multimodel



Analysis

Fraud detection is a set of operations used to prevent money or property from being obtained fraudulently. false representation Fraud can be committed in a variety of ways and in a wide range of industries. To make a decision, the majority of detection systems use a range of fraud detection datasets to generate a connected picture of both legitimate and invalid payment data. This decision must take into account IP address, geolocation, device identity, "BIN" data, global latitude/longitude, historical transaction trends, and transaction information. In practice, this implies that merchants and issuers use analytically based solutions to detect fraud, which leverage internal and external data to apply a set of business rules or analytical algorithms.

Credit Card Fraud Detection Using Machine Learning is a data research process. by a team of data scientists and the creation of a model that will yield the greatest outcomes in terms of detecting and preventing fraudulent transactions This is accomplished by aggregating all relevant information of card users' transactions, such as Date, User Zone, Product Category, Amount, Provider, Client's Behavioral Patterns, and so on. The data is then fed into a subtly trained model that looks for patterns and rules to determine if a transaction is fraudulent or lawful. All major banks, including Chase, employ fraud monitoring and detection systems.

Credit Card Fraud Techniques and Prevention

Associations and banks that use them make excellent proposals. arrangements for security to address these concerns, but at the same time Following the development of the same fraudsters' simple methods a period of time as a result, it is critical for future development. Techniques for recognizing and counteracting Location of Fraud because the primary goal of avoidance is to distinguish to distinguish between legitimate and bogus exchanges and to prevent phony movement. In the event that the framework fails to detect and prevent bogus exercises, extortion detection dominates. Frameworks for managed extortion discovery in light of this, new exchanges are labelled as fraudulent or certified. characteristics of both deceptive and genuine exercises Anomalies exchanges are identified as potentially fake. trades in individual extortion location frameworks

System configuration

This project may be run on standard hardware. We ran the entire project on an Intel I5 processor with 8 GB RAM and a 2 GB Nvidia Graphic Processor. It also has two cores that run at 1.7 GHz and 2.1 GHz. The first half of the process is the training phase, which takes about 10-15 minutes, and the second part is the testing phase, which just takes a few seconds to generate predictions and calculate accuracy.

Hardware Requirements:

- RAM: 4 GB

- Storage: 500 GB

- CPU: 2 GHz or faster

- Architecture: 32-bit or 64-bit

Software requirements

• Python 3.5 in Google Colab is used for data pre-processing, model training and prediction.

• Operating System: windows 7 and above or Linux based OS or MAC OS.

Implementation

Data visualization is the discipline of attempting to understand data by displaying it in a visual context in order to highlight patterns, trends, and connections that might otherwise go undetected.

Python has a number of excellent graphing packages that are jam-packed with useful functionality. Python provides a great library for creating dynamic or highly customizable charts.

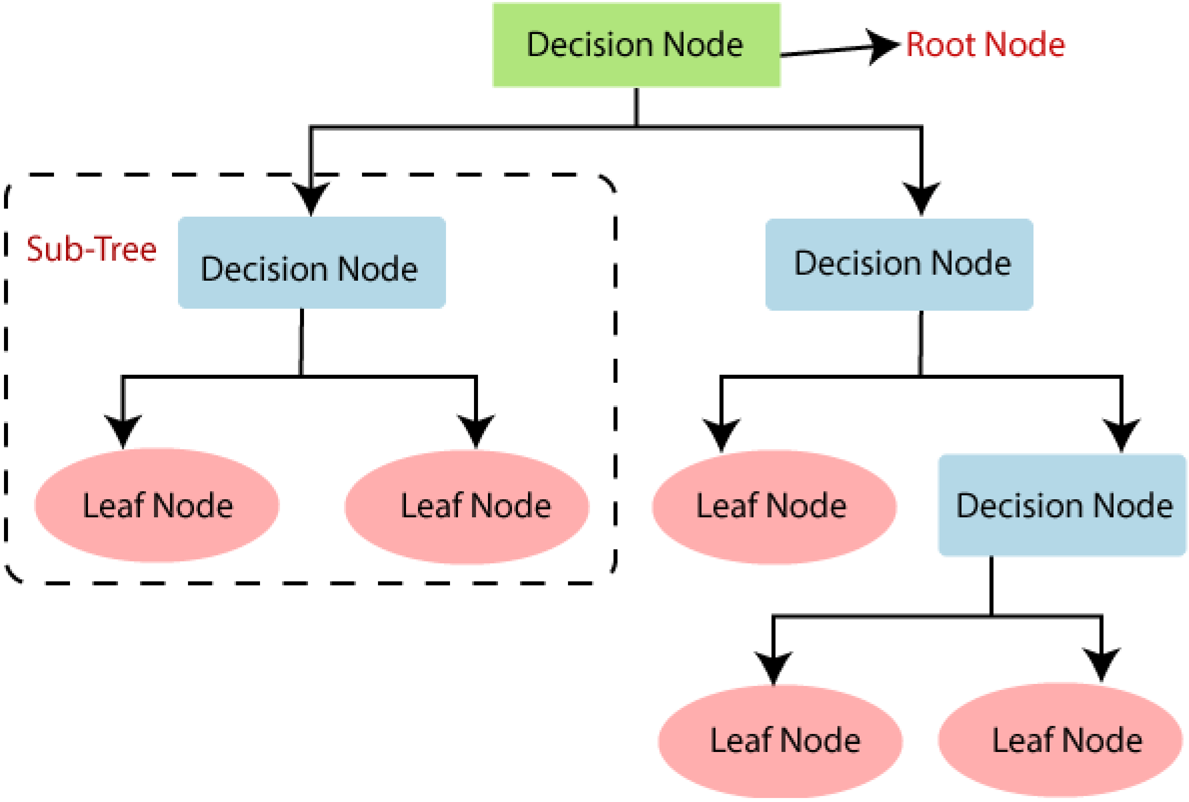
To get a little overview, here are a few popular plotting libraries:

- **Matplotlib:** low level, provides lots of freedom
- **Pandas Visualization:** easy to use interface, built on Matplotlib
- **Seaborn:** high-level interface, great default styles
- **plotnine:** based on R's ggplot2, uses Grammar of Graphics
- **Plotly:** can create interactive plots

Algorithm of Decision Tree Classification

- Decision Tree is a Supervised learning technique that may be applied for both classification and regression issues, however it is most commonly employed for classification. It is a tree-structured classifier in which internal nodes contain dataset attributes, branches represent decision rules, and each leaf node represents the result.
- A Decision tree has two nodes: the Decision Node and the Leaf Node. Decision nodes are used to make decisions and have numerous branches, whereas Leaf nodes represent the results of those decisions and do not have any additional branches.
- The decisions or tests are based on the characteristics of the presented dataset.
- It is a graphical representation of all possible solutions to a problem/decision given certain conditions.

- It is named a decision tree because, like a tree, it begins with the root node and then branches out to form a tree-like structure.
- The CART algorithm, which stands for Classification and Regression Tree algorithm, is used to form a tree.
- A decision tree simply asks a question and divides the tree into subtrees based on the answer (Yes/No).



Flow of the Process

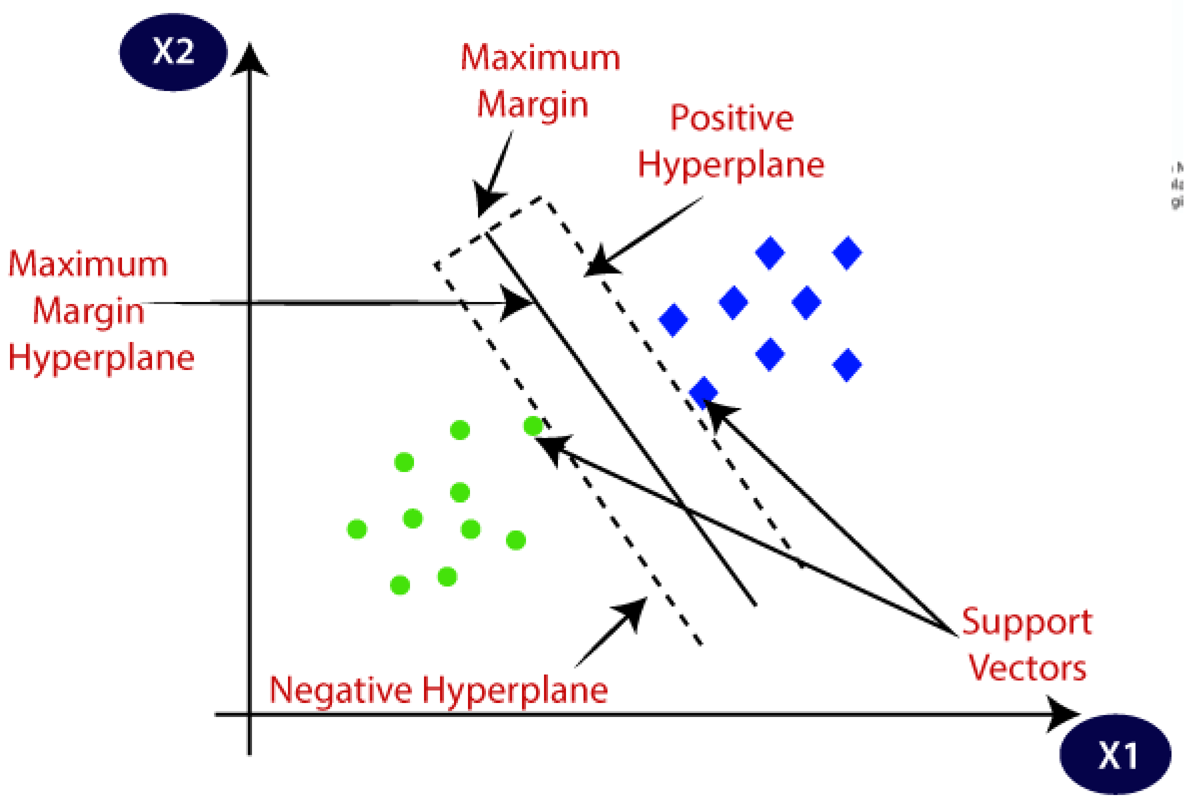
- Step 1: Begin the tree with the root node, which contains the entire dataset, says S.
- Step 2: Using the Attribute Selection Measure, find the best attribute in the dataset (ASM).
- Step 3: Subdivide the S into subsets containing potential values for the best qualities.
- Step 4: Create the decision tree node with the best attribute.
- Step 5: Create new decision trees recursively using the subsets of the dataset obtained in step 3. Continue this process until you reach a point where you can no longer categorize the nodes and refer to the final node as a leaf node.

Support Vector Machine

Support Vector Machine, or SVM, is a prominent Supervised Learning technique that is used for both classification and regression issues. However, it is mostly utilized in Machine Learning for Classification difficulties.

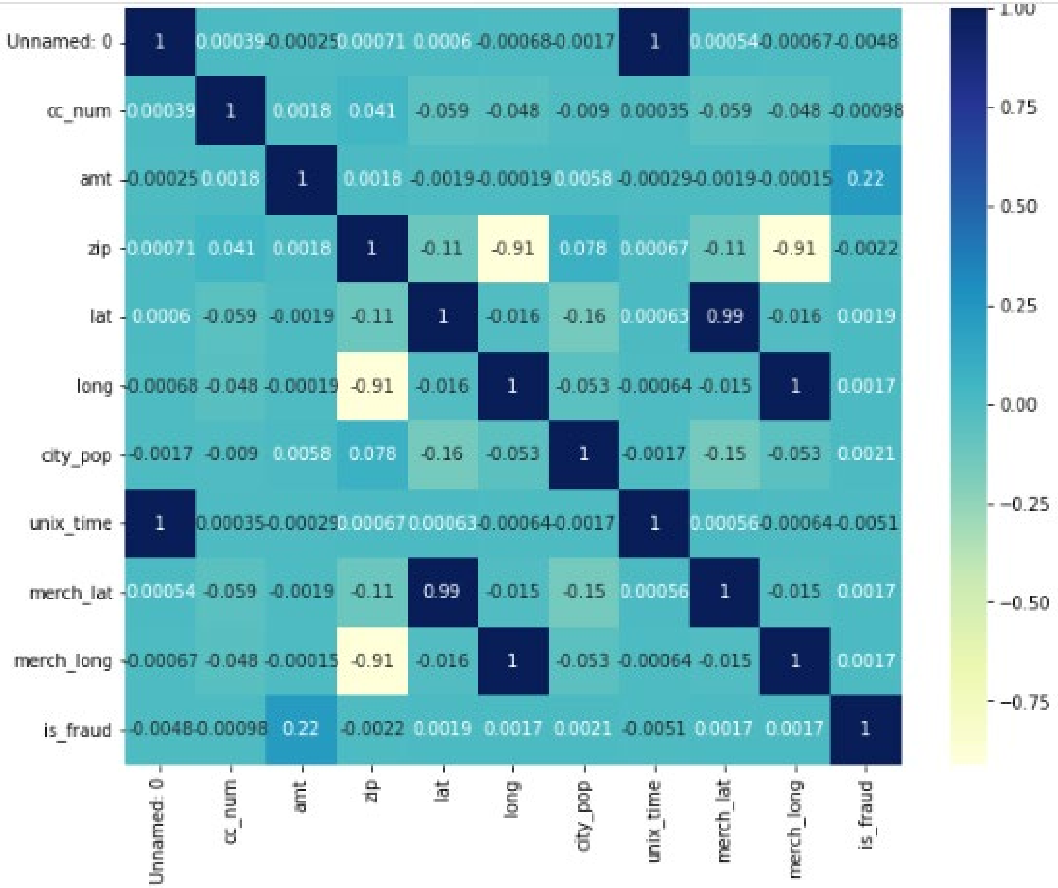
The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space so that we may easily place fresh data points in the correct category in the future. A hyperplane is the optimal choice boundary.

SVM selects the extreme points/vectors that aid in the creation of the hyperplane. These extreme examples are referred to as support vectors, and the method is known as the Support Vector Machine. Consider the picture below, which shows two distinct categories that are classified.



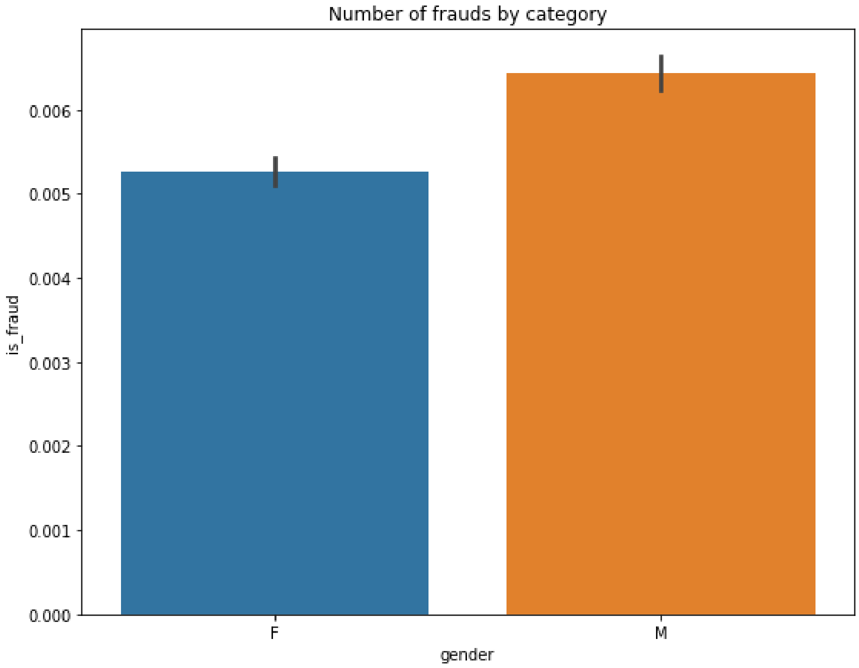
Preliminary Results

Correlation matrix

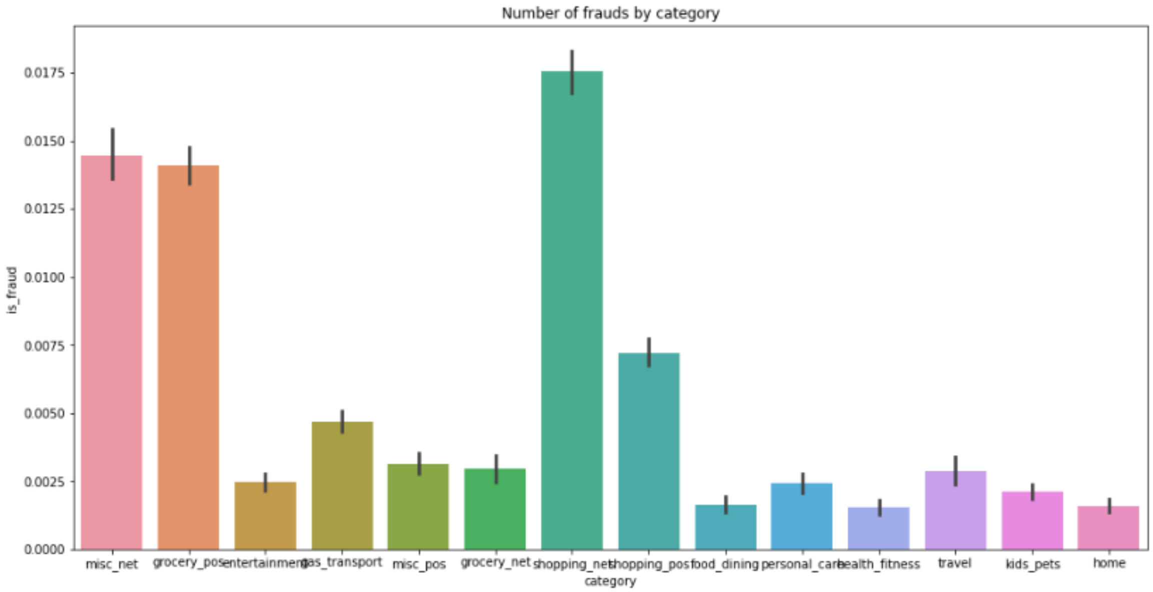


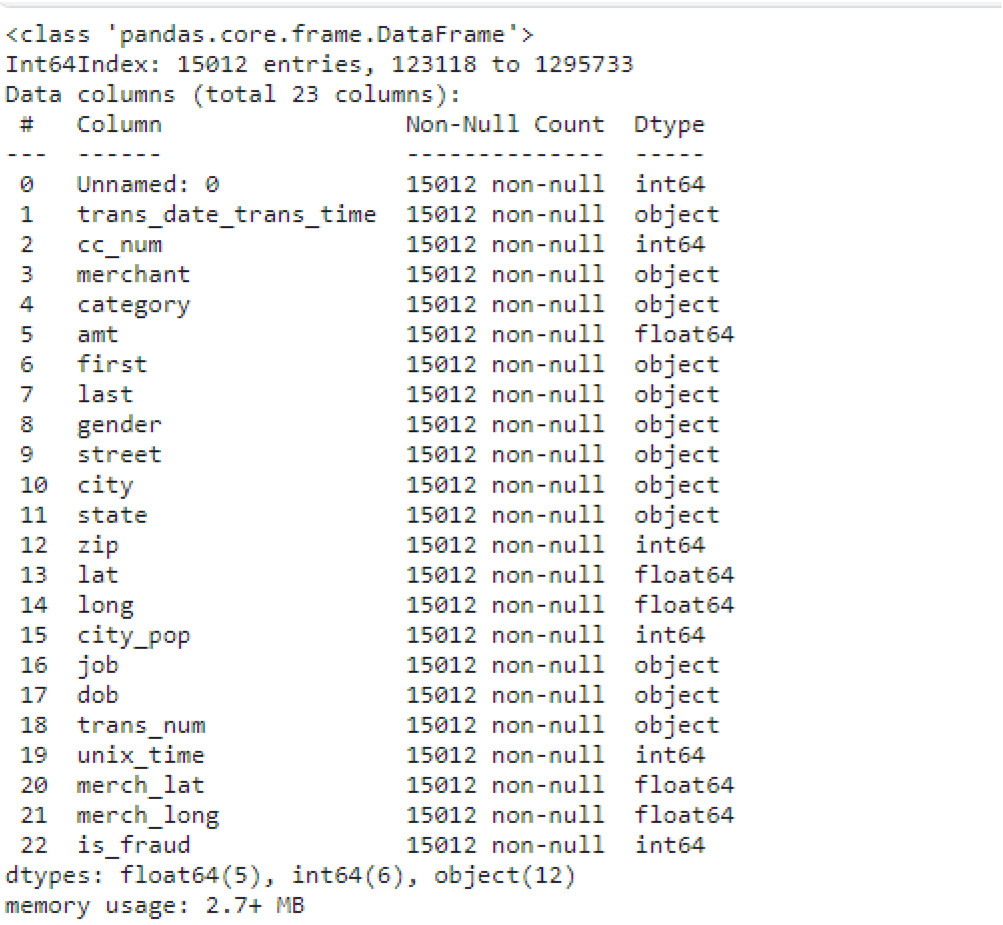
	Unnamed: 0	cc_num	amt	zip	lat	long	city_pop	unix_time	merch_lat	merch_long	is_fraud
Unnamed: 0	1	0.00039	-0.00025	0.00071	0.0006	-0.00068	-0.0017	1	0.00054	-0.00067	-0.0048
cc_num	0.00039	1	0.0018	0.041	-0.059	-0.048	-0.009	0.00035	-0.059	-0.048	-0.00098
amt	-0.00025	0.0018	1	0.0018	-0.0019	-0.00019	0.0058	-0.00029	-0.0019	-0.00015	0.22
zip	0.00071	0.041	0.0018	1	-0.11	-0.91	0.078	0.00067	-0.11	-0.91	-0.0022
lat	0.0006	-0.059	-0.0019	-0.11	1	-0.016	-0.16	0.00063	0.99	-0.016	0.0019
long	-0.00068	-0.048	-0.00019	-0.91	-0.016	1	-0.053	-0.00064	-0.015	1	0.0017
city_pop	-0.0017	-0.009	0.0058	0.078	-0.16	-0.053	1	-0.0017	-0.15	-0.053	0.0021
unix_time	1	0.00035	-0.00029	0.00067	0.00063	-0.00064	-0.0017	1	0.00056	-0.00064	-0.0051
merch_lat	0.00054	-0.059	-0.0019	-0.11	0.99	-0.015	-0.15	0.00056	1	-0.015	0.0017
merch_long	-0.00067	-0.048	-0.00015	-0.91	-0.016	1	-0.053	-0.00064	-0.015	1	0.0017
is_fraud	-0.0048	-0.00098	0.22	-0.0022	0.0019	0.0017	0.0021	-0.0051	0.0017	0.0017	1

```
[50]: <AxesSubplot:title={'center':'Number of frauds by category'}, xlabel='gender', ylabel='is_fraud'>
```


Number of frauds by category
is_fraud
0.006
0.005
0.004
0.003
0.002
0.001
0.000
F
M
gender

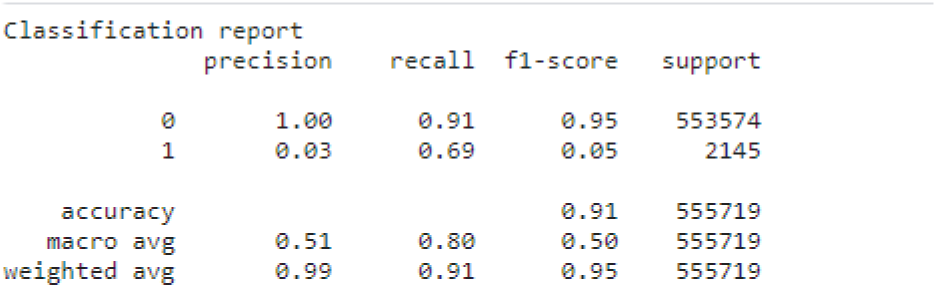
```
[51]: <AxesSubplot:title={'center':'Number of frauds by category'}, xlabel='category', ylabel='is_fraud'>
```


Number of frauds by category
is_fraud
0.0175
0.0150
0.0125
0.0100
0.0075
0.0050
0.0025
0.0000
misc_net
grocery_pos
entertainment
gas_transport
misc_pos
grocery_net
shopping_net
shopping_pos
food_dining
personal_care
health_fitness
travel
kids_pets
home
category



```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 15012 entries, 123118 to 1295733
Data columns (total 23 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   Unnamed: 0             15012 non-null  int64  
 1   trans_date_trans_time  15012 non-null  object 
 2   cc_num                 15012 non-null  int64  
 3   merchant               15012 non-null  object 
 4   category               15012 non-null  object 
 5   amt                    15012 non-null  float64
 6   first                  15012 non-null  object 
 7   last                   15012 non-null  object 
 8   gender                 15012 non-null  object 
 9   street                 15012 non-null  object 
 10  city                   15012 non-null  object 
 11  state                  15012 non-null  object 
 12  zip                    15012 non-null  int64  
 13  lat                    15012 non-null  float64
 14  long                   15012 non-null  float64
 15  city_pop               15012 non-null  int64  
 16  job                    15012 non-null  object 
 17  dob                    15012 non-null  object 
 18  trans_num              15012 non-null  object 
 19  unix_time              15012 non-null  int64  
 20  merch_lat              15012 non-null  float64
 21  merch_long             15012 non-null  float64
 22  is_fraud               15012 non-null  int64  
dtypes: float64(5), int64(6), object(12)
memory usage: 2.7+ MB
```

SVM Classification Report

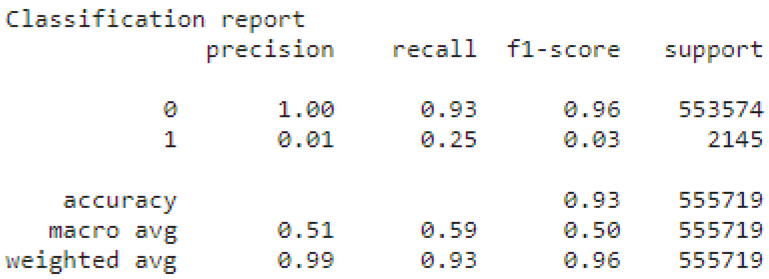


```
Classification report
              precision    recall  f1-score   support

           0       1.00      0.91      0.95    553574
           1       0.03      0.69      0.05      2145

    accuracy                           0.91    555719
   macro avg       0.51      0.80      0.50    555719
weighted avg       0.99      0.91      0.95    555719
```

DT Classification Report

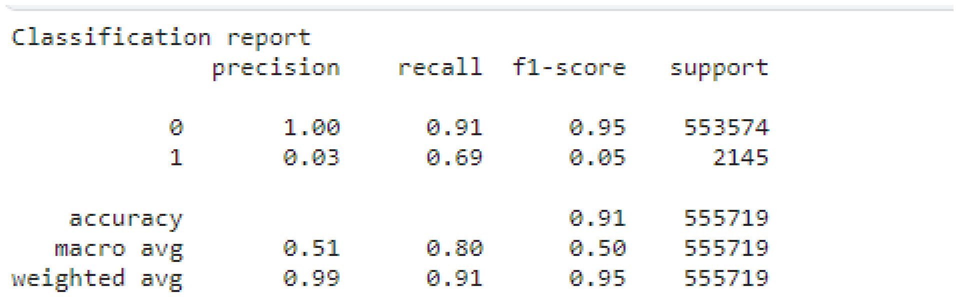


```
Classification report
              precision    recall  f1-score   support

           0       1.00      0.93      0.96    553574
           1       0.01      0.25      0.03      2145

    accuracy                           0.93    555719
   macro avg       0.51      0.59      0.50    555719
weighted avg       0.99      0.93      0.96    555719
```

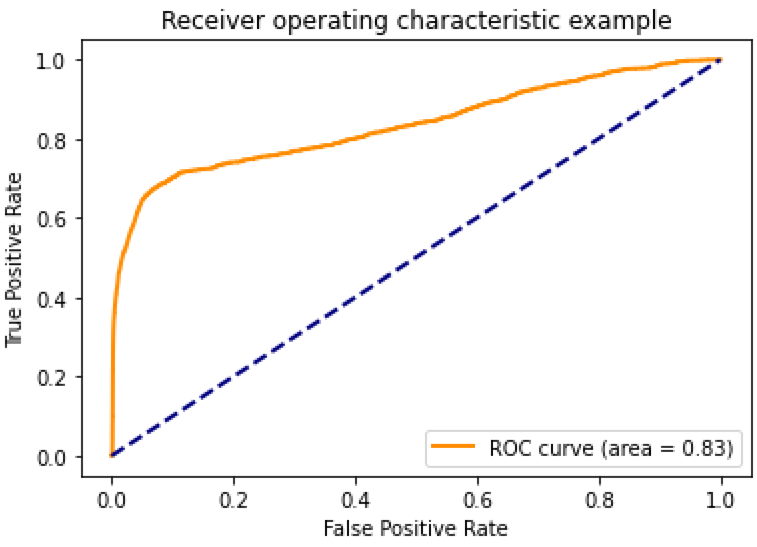
Logistic Regression



```
Classification report
              precision    recall  f1-score   support

           0       1.00      0.91      0.95    553574
           1       0.03      0.69      0.05      2145

    accuracy                           0.91    555719
   macro avg       0.51      0.80      0.50    555719
weighted avg       0.99      0.91      0.95    555719
```



Project Management

Implementation status report

Work completed:

Description	**Responsibility - Task**	**Responsibility - Person**	**Contributions - percentage**
Data Read and Preprocessing	Finding dataset, tuning and scaling features in dataset.	Pranav	100
Visualization and Data Transformation	Visualizing the data by using various visualizations techniques.	Anil Kumar	100
Model And Result Analysis	Analyzing of different fraud detection models.	Satya Ishyanth	100

Work to be completed:

Description	**Responsibility - Task**	**Responsibility - Person**	**Issues/ Concerns**
XGBoost			
Result Analysis	Improve the accuracy of the model, trying different algorithms.	Satya Ishyanth	Yield better results

References/Bibliography

- S. Bachmayer, "Artificial Immune Systems," pp. 119-131 in Artificial Immune Systems, vol. 5132, 2008.M. Krivko, "A Hybrid Model for Plastic Card Fraud, "Expert Systems with Applications, vol. 37, no. 8, pp. 6070-6076, August 2010.
- S. Bhattacharyya, S. Jha, K. Tharakunnel, and J. C. Westland, "Data mining for credit card fraud: A comparative study," Decision Support Systems, vol. 50, no. 3, Feb. 2011, pp. 602-613.
- "Plastic card fraud detection via peer group analysis," Advances in Data Analysis and Classification, vol. 2, no. 1, pp. 45-62, Mar. 2008.
- A. Mishra, C. Ghorpade, “Credit Card Fraud Detection on the Skewed Data Using Various Classification and Ensemble Techniques” 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS) pp. 1-5. IEEE.
- M. Krivko, “A hybrid model for plastic card fraud detection systems,” Expert Systems with Applications, vol. 37, no. 8, pp. 6070–6076, Aug. 2010.