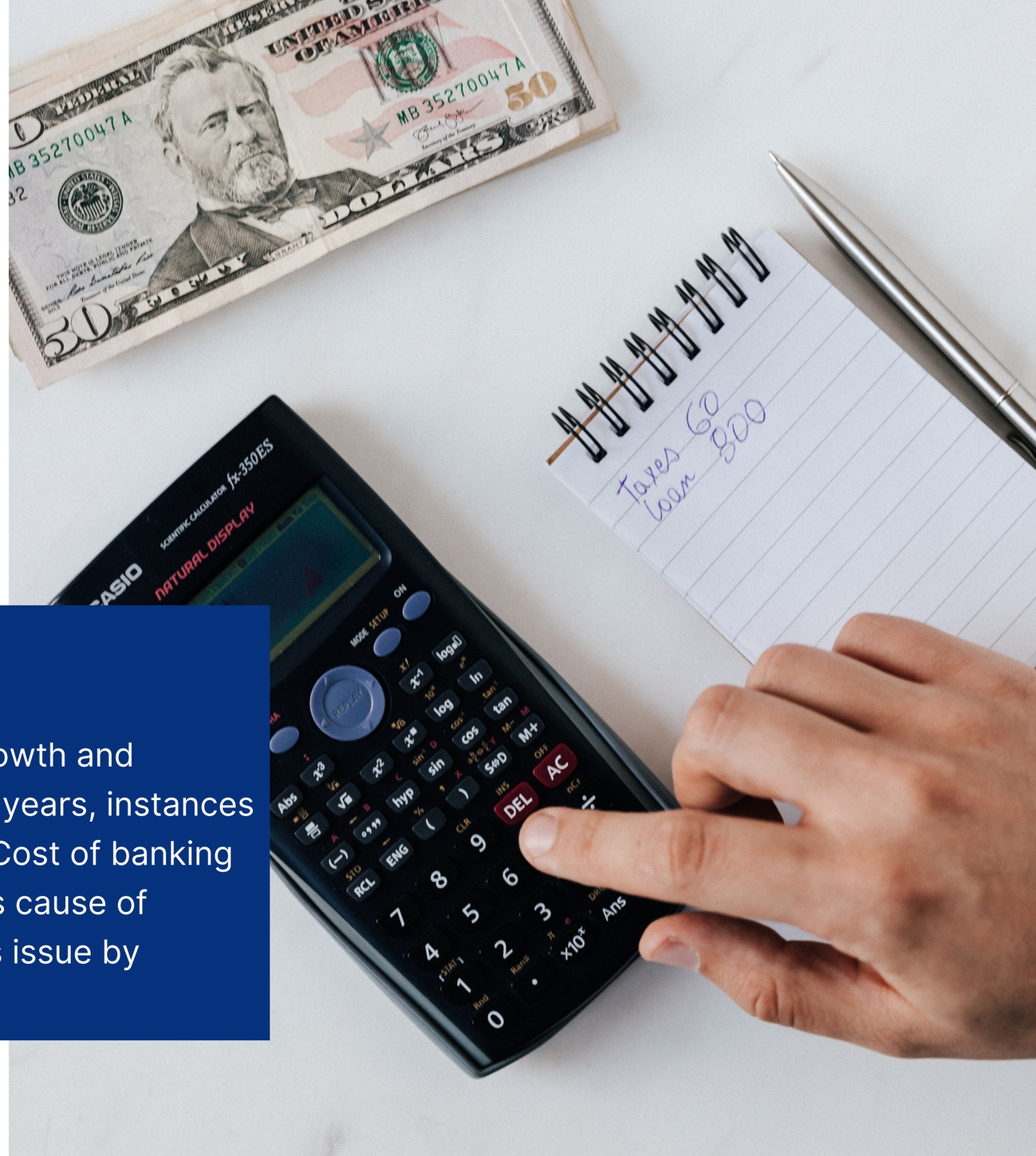
Data Science

Fraud Detection in Banking Systems

Introduction



The Indian banking sector has experienced significant growth and changes since liberalization of economy in 1991.In recent years, instances of financial frauds have regularly been reported in India. Cost of banking frauds have increased manifold resulting in a very serious cause of concern. Hence an attempt need to be made to solve this issue by providing a proper mechanism of detecting frauds.

Objectives

- To detect frauds in banking transactions.
- Analysing the variations in frauds based on amount spent.
- Finding out the nature of frauds based on categories of spending.

	Date	Open	High	Low	Close	Adj Close	Volume
0	May 03, 2022	8.06	8.24	7.93	8.11	8.11	523500
1	May 02, 2022	7.62	8.07	7.58	8.05	8.05	347500
2	Apr 29, 2022	7.66	7.87	7.59	7.65	7.65	259800
3	Apr 28, 2022	7.52	7.71	7.37	7.68	7.68	227400
4	Apr 27, 2022	7.49	7.62	7.4	7.47	7.47	270300
5	Apr 26, 2022	7.79	7.84	7.5	7.53	7.53	303500
6	Apr 25, 2022	7.6	7.87	7.52	7.86	7.86	301900
7	Apr 22, 2022	7.89	7.89	7.64	7.69	7.69	301400
8	Apr 21, 2022	8.14	8.14	7.83	7.9	7.9	285300
9	Apr 20, 2022	8.29	8.3	7.97	7.99	7.99	163000
10	Apr 19, 2022	8.08	8.31	7.97	8.16	8.16	314100
11	Apr 18, 2022	8.3	8.32	7.92	8.01	8.01	432800
12	Apr 14, 2022	8.23	8.57	8.13	8.28	8.28	373900
13	Apr 13, 2022	7.74	8.18	7.71	8.17	8.17	559300
14	Apr 12, 2022	7.95	8.14	7.67	7.71	7.71	316500
15	Apr 11, 2022	7.79	8.04	7.69	7.86	7.86	384900
16	Apr 08, 2022	7.82	8.06	7.77	7.81	7.81	269100
17	Apr 07, 2022	8.02	8.12	7.6	7.79	7.79	503800
18	Apr 06, 2022	8.05	8.11	7.9	8.07	8.07	399000
19	Apr 05, 2022	8.38	8.45	8.12	8.21	8.21	285700
20	Apr 04, 2022	8.12	8.39	8.1	8.31	8.31	408200
21	Apr 01, 2022	8.19	8.33	8.09	8.12	8.12	430700
22	Mar 31, 2022	8.21	8.23	7.92	8.17	8.17	643400
23	Mar 30, 2022	8.62	8.64	8.15	8.18	8.18	618700
24	Mar 29, 2022	8.5	8.93	8.5	8.67	8.67	722300
25	Mar 28, 2022	8.12	8.44	8.03	8.42	8.42	664000
26	Mar 25, 2022	8.16	8.26	8.05	8.09	8.09	387500
27	Mar 24, 2022	8.21	8.24	8.08	8.16	8.16	291200
28	Mar 23, 2022	8.25	8.42	8.11	8.17	8.17	319900
29	Mar 22, 2022	8.28	8.39	8.18	8.34	8.34	582900
30	Mar 21, 2022	8.55	8.55	8.15	8.23	8.23	346400
31	Mar 18, 2022	8.47	8.72	8.4	8.59	8.59	607100
32	Mar 17, 2022	8.28	8.56	8.24	8.55	8.55	458800
33	Mar 16, 2022	8.06	8.38	8	8.3	8.3	572600
34	Mar 15, 2022	7.86	8.05	7.82	7.99	7.99	381900
35	Mar 14, 2022	8.04	8.04	7.73	7.83	7.83	426900
36	Mar 11, 2022	8.43	8.48	7.92	7.93	7.93	483200
37	Mar 10, 2022	8.48	8.65	8.35	8.41	8.41	652000

Scrapped Dataset

Scrapping

Dataset was obtained by scrapping
https://finance.yahoo.com/
Scrapping tools used – Beautiful Soup, Pandas, Requests and csv

Steps involved

1. Generating the required url.
2. Generating the raw data.
3. Building the dataframe.
4. Conversion to csv.

Dataset used

https://www.kaggle.com/code/turkayavci/fraud-detection-on-bank-payments/data

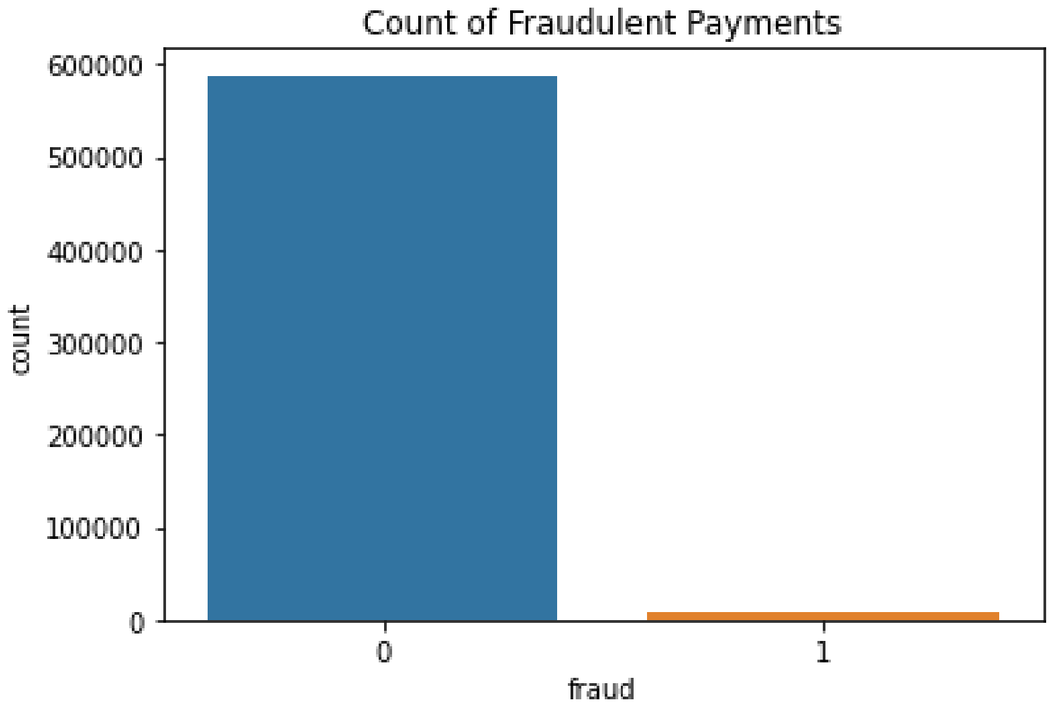
The dataset used contains the following attributes:

- step
- customer
- age
- gender
- zipcodeOri
- merchant
- zipMerchant
- category
- amount
- fraud

Data Preprocessing

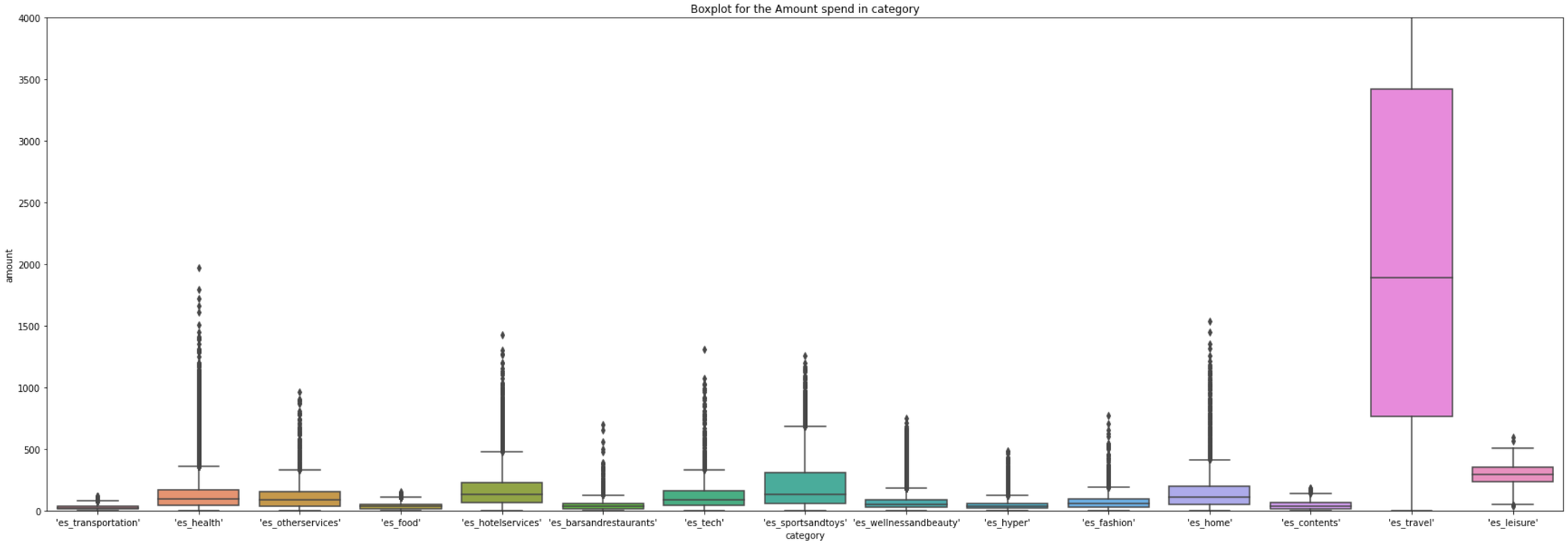
- Dropped irrelevant attributes from the dataset like zipCodeOri and zipMerchant.
- Dropped the redundant records.
- Transforming objects to categorical datatypes.
- Converting categorical datatypes into numerical.

Data Visualization



Count of fraudulent payments in dataset.

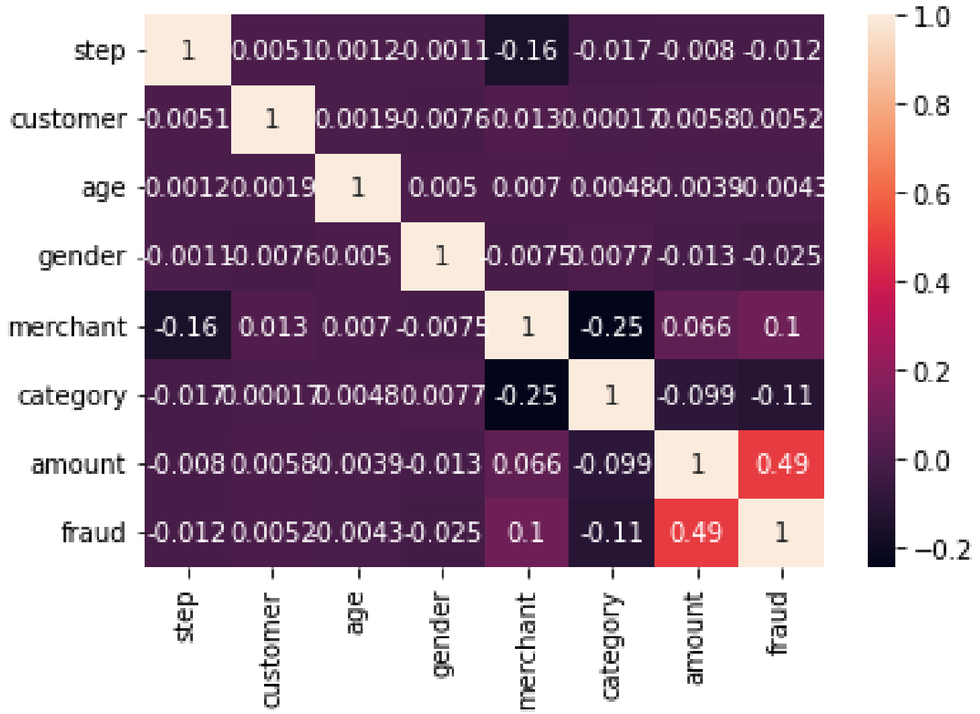
Data Visualization



Boxplot

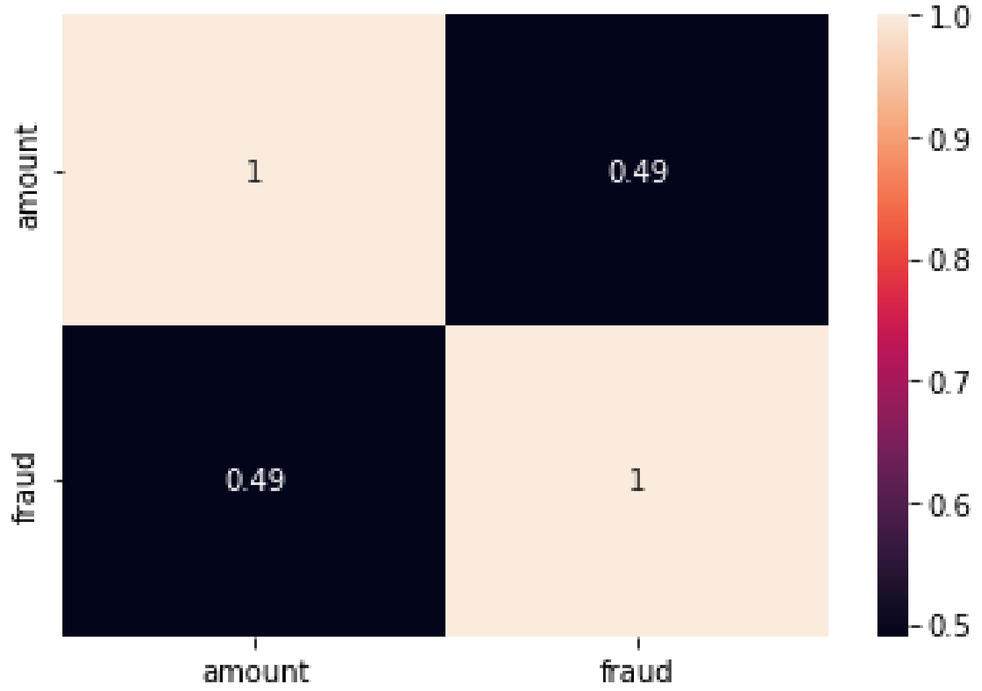
Categorical distribution of amounts spent.

Data Visualization



Heatmap showing correlation of all attributes.

Data Visualization



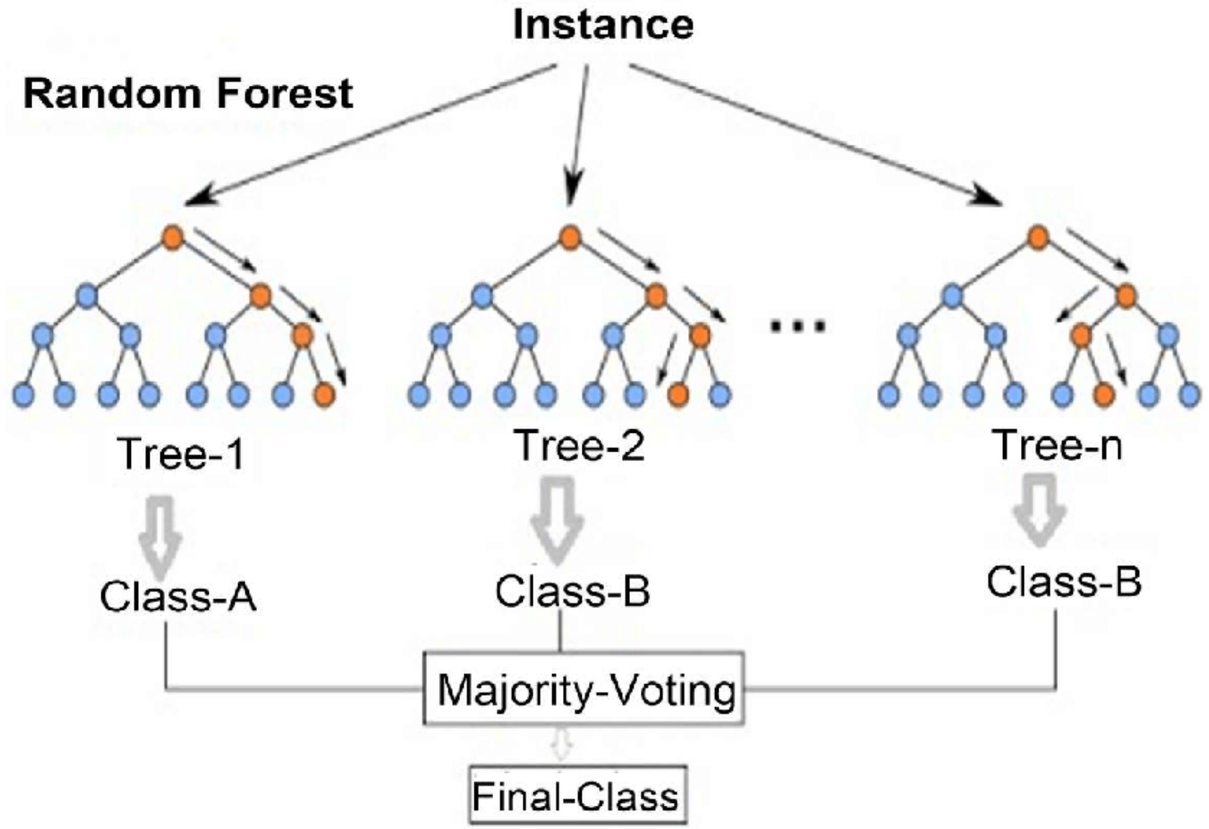
Heatmap of Amount vs Fraud

Model used

Random Forest Classifier

There are two classes present in the data that are fraudulent and legitimate transactions.

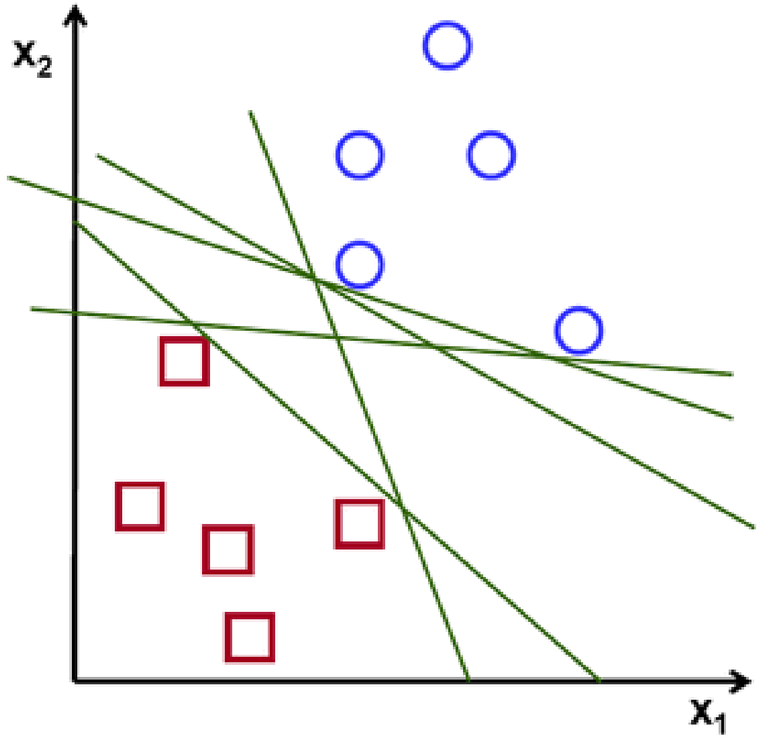
We used the Random Forest Classifier to classify the data set and detect frauds.

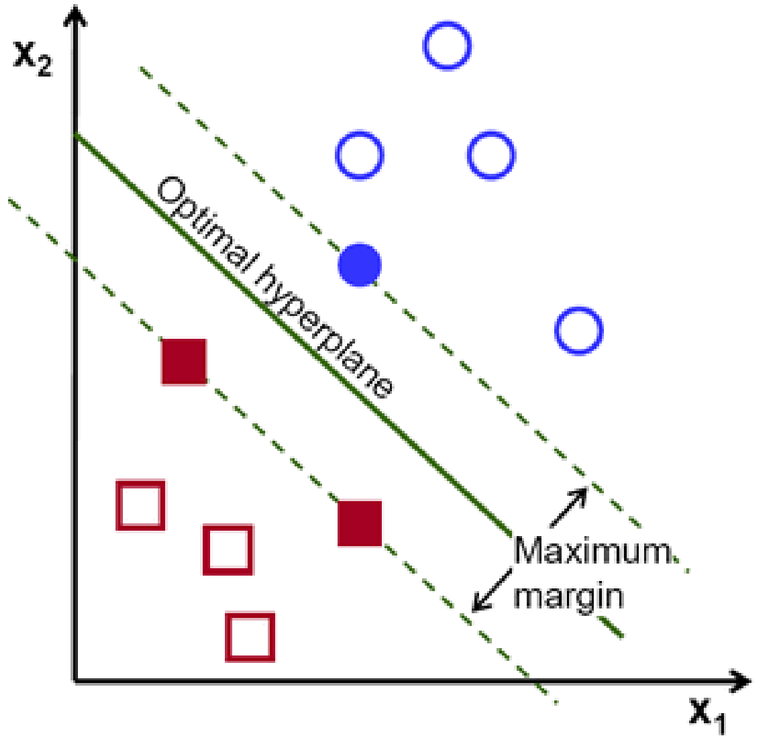


Model used

Support Vector Classification

The SVM Algorithm finds a hyperplane in an N-dimensional space where N is the number of features that distinctly classifies the data points.
This hyperplane must have maximum distance between data points of both classes.





Performance Evaluation

- Accuracy $$\frac{TP + TN}{TP + FP + TN + FN}$$
- Precision $$\frac{TP}{TP + FP}$$
- Recall $$\frac{TP}{TP + FN}$$
- F1-Score $$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$