GALGOTIAS
UNIVERSITY

(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

FRAUD DETECTION USING MACHINE LEARNING

A Report for the Evaluation 3 of Project 2

Submitted by
Gaurav Sharma
(1613112021/16SCSE112033)

in partial fulfillment for the award of the degree
of

Bachelor of Technology
IN
Computer Science and Engineering With Data Analytics

SCHOOL OF COMPUTING SCIENCE AND ENGINEERING

Under the Supervision of
Mr Sachin Minocha
Professor

APRIL / MAY- 2020

TABLE OF CONTENTS

- ✓ Abstract
- ✓ Introduction
    - i. Overall Description
    - ii. Purpose
    - iii. Scope
- ✓ Literature Survey
- ✓ Purposed Model
- ✓ Software Requirement Specification

ABSTRACT :

Recent research has shown that machine learning techniques have been applied very effectively to the problem of payments related fraud detection. Such ML based techniques have the potential to evolve and detect previously unseen patterns of fraud. In this paper, we apply multiple ML techniques based on Logistic regression and Support Vector Machine to the problem of payments fraud detection using a labeled dataset containing payment transactions. We show that our proposed approaches are able to detect fraud transactions with high accuracy and reasonably low number of false positives.

ML based approaches involving ANN (Artificial Neural Networks), SVM (Support Vector machines) ,HMM (Hidden Markov Models), clustering etc

Introduction

➢ Overall Description

We are living in a world which is rapidly adopting digital payments systems. Credit card and payments companies are experiencing a very rapid growth in their transaction volume. In third quarter of 2018, PayPal Inc (a San Jose based payments company) processed 143 billion USD in total
payment volume . Along with this transformation, there is also a rapid increase in financial fraud that happens in these payment systems.

➢ Purpose

An effective fraud detection system should be able to detect fraudulent transactions with high accuracy and efficiency. While it is necessary to prevent bad actors from executing fraudulent transactions, it is also very critical to ensure genuine users are not prevented from accessing the payments system. A large number of false positives may translate into bad customer experience and may lead customers to take their business elsewhere.

- **Scope**

A major challenge in applying ML to fraud detection is presence of highly imbalanced data sets. In many available datasets, majority of transactions are genuine with an extremely small percentage of fraudulent ones. Designing an accurate and efficient fraud detection system that is low on false positives but detects fraudulent activity effectively is a significant challenge for researchers.

In our paper, we apply multiple binary classification approaches - Logistic regression , Linear SVM and SVM with RBF kernel on a labeled dataset that consists of payment transactions.Our goal is to build binary classifiers which are able to separate fraud transactions from non-fraud transactions. We compare the effectiveness of these approaches in detecting fraud transactions.

Literature survey:

- **Logistic Regression :**

Logistic Regression is a supervised learning technique that is used when the decision is categorical. It means that the result will be either ‘fraud’ or ‘non-fraud’ if a transaction occurs.

Use Case: Let us consider a scenario where a transaction occurs and we need to check whether it is a ‘fraudulent’ or ‘non-fraudulent’ transaction. There will be given set of parameters that are checked and, on the basis of the probability calculated, we will get the output as ‘fraud’ or ‘non-fraud.’

Logistic regression is a technique used to find a linear decision boundary for a binary classifier. For a given input feature vector x, a logistic regression model with parameter θ classifies the input x using the following hypothesis $h\theta(x) = g(\theta T\ x) = 1\backslash 1+e-\theta T\ x$ where g is known as Sigmoid function.

For a binary classification problem, the output $h\theta(x)$ can be interpreted as a probability of x as belonging to class 1. The parameters θ can be given as

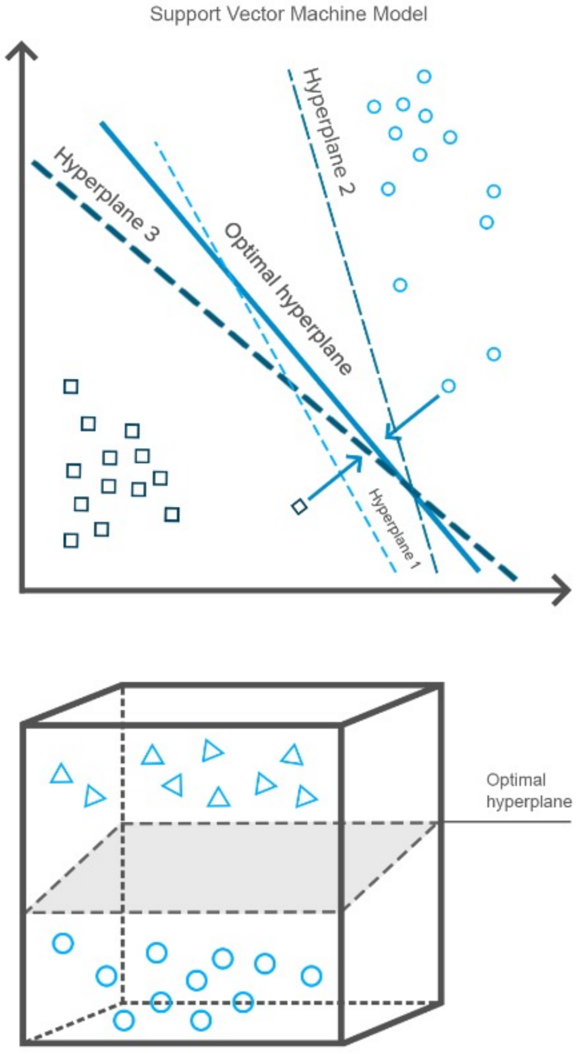
$J(\theta) = 1\backslash m\ \sum m$ and $i=1\ \log 1 + \exp(-y\ (i)\ \theta T\ x\ (i))$

✓ Support Vector Machine :

Support vector machine creates a classification hyperplane in the space defined by input feature vectors. The training process aims to determine a hyper-plane that maximizes geometric margin with respect to labeled input data. SVMs
optimization problem can be characterized b

$$\text{Min } \gamma,\psi,\beta \ 1\backslash 2||\psi||2+C\sum \varepsilon i$$

$\text{S.t.}d(i)(w(T)x(i)+b)>=1-\varepsilon i \quad i=1.........m$



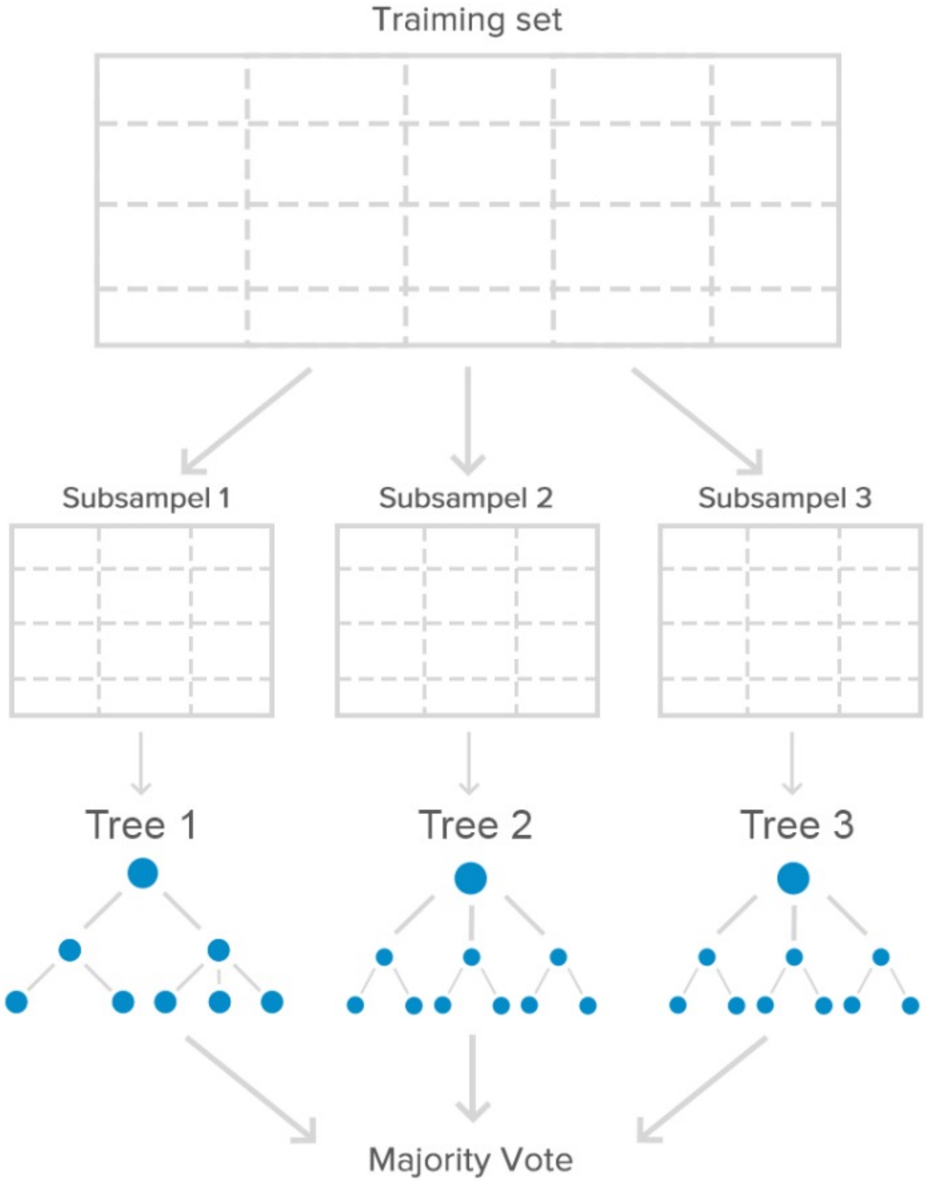
As some studies show, SVM can be inferior to random forests in credit card transactions with small datasets, but can also approach their accuracy once datasets are large enough.

✓ Random Forest

Random Forest uses a combination of decision trees to improve the results. Each decision tree checks for different conditions. They are trained on random datasets and, based on the training of the decision trees, each tree gives the probability of the transaction being ‘fraud’ and ‘non-fraud.’ Then, the model predicts the result accordingly.

Use Case: Let’s consider a scenario where a transaction is made. Now, we will see how the random forest in Machine Learning is used in fraud detection algorithms.

Random Forest or Ensemble of Decision Trees Model



Traiming set

Subsampel 1

Subsampel 2

Subsampel 3

Tree 1

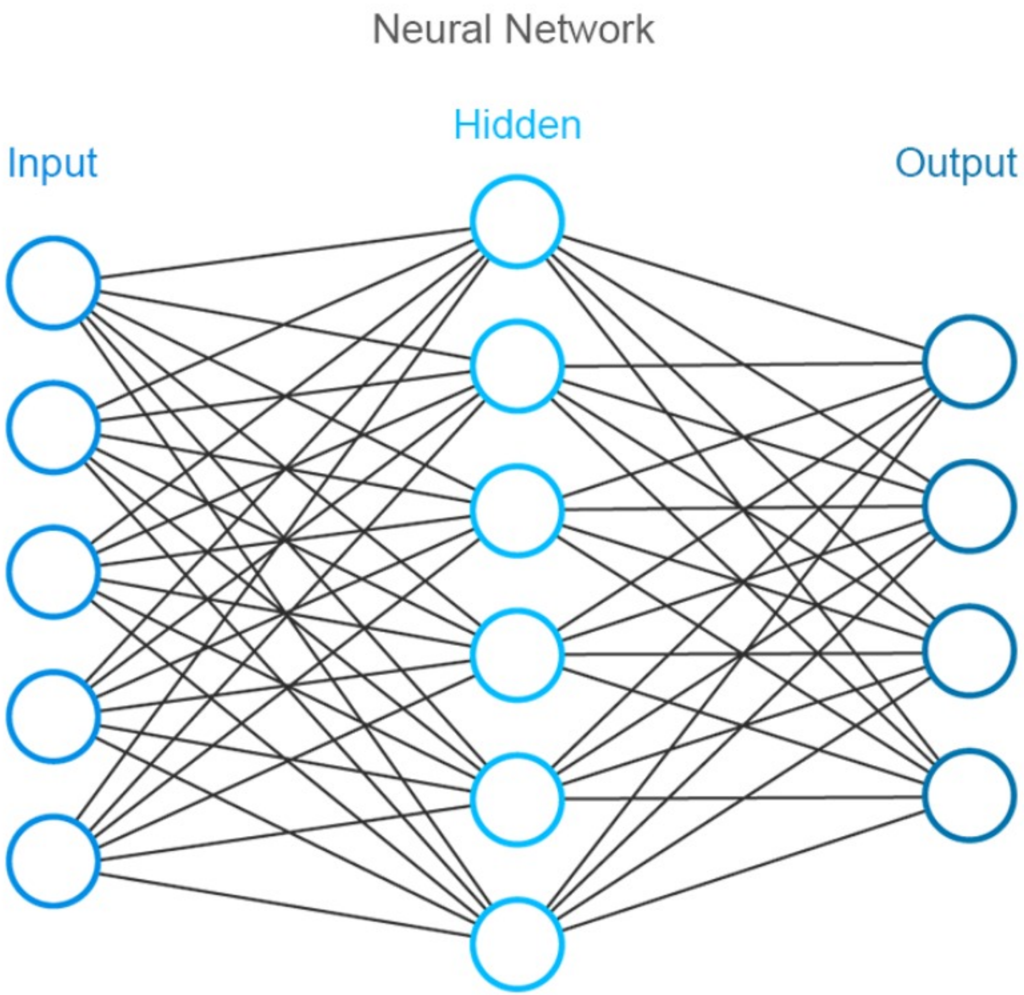
Tree 2

Tree 3

Majority Vote

✓ Neural Network

Neural Network is a model that allows for determining non-linear relations between the records. The algorithm structure is built on principles close to those of the human brain neurons. The model is trained on a labeled dataset making input data pass through several layers (i.e. sets of mathematical functions). The models of this type employ 1-2 hidden layers.

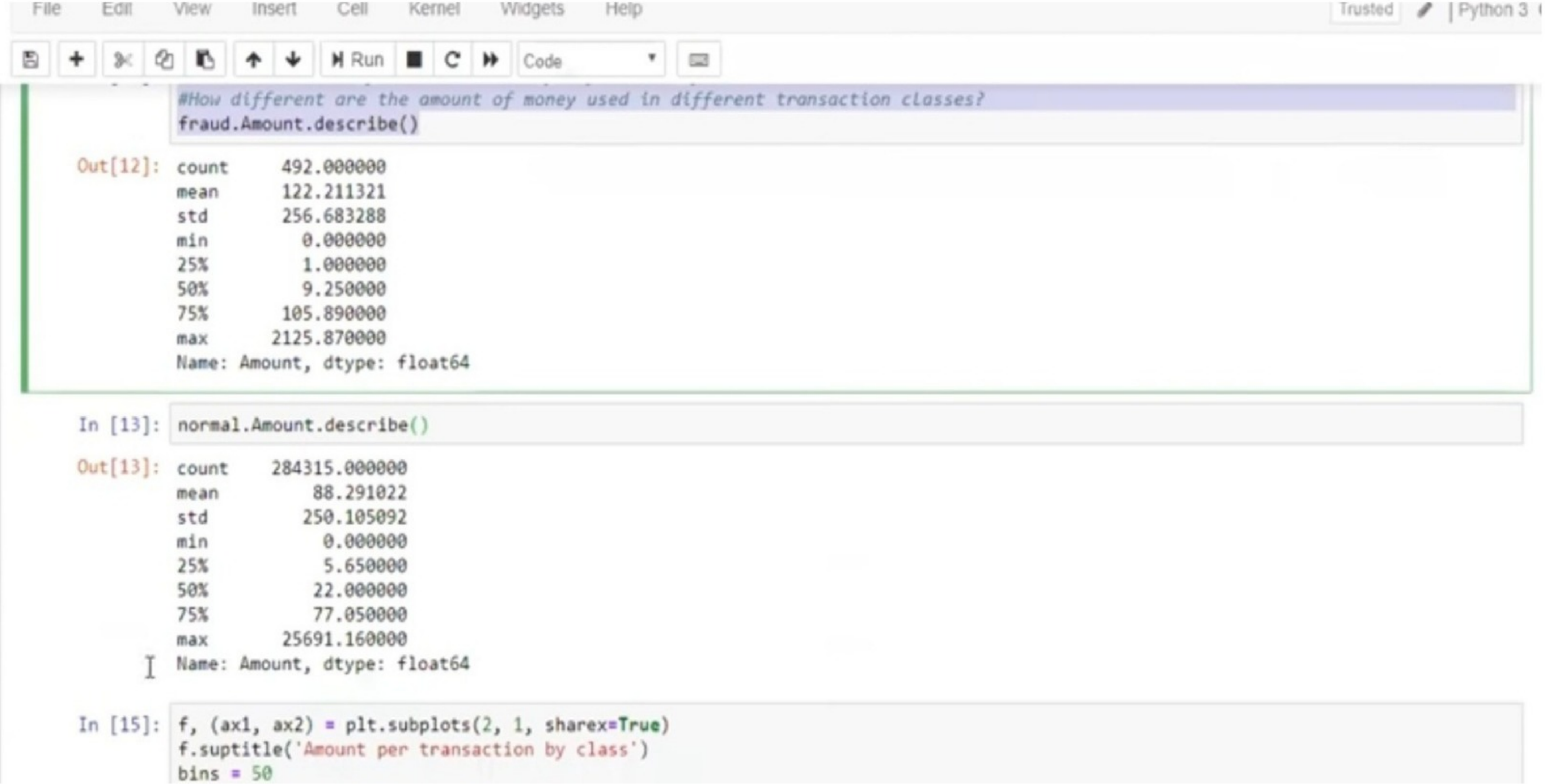


In this blog, we have seen how fraud detection algorithms work using Machine Learning techniques such as logistic regression, support vecator, random forest, and neural networks. This technology is improving day by day so that it provides us more accuracy and better results to prevent fraud.

Software Required Specification

- Operating System :
  - Windows 10
- Languages:
  - Python
- Editor:
  - Python 3.7
- Visualization Tool:
  - Visual Studio Code

Output



```
#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
```

Fig.1. These are the data of money used in different transaction classes. Fraud detection Amounts

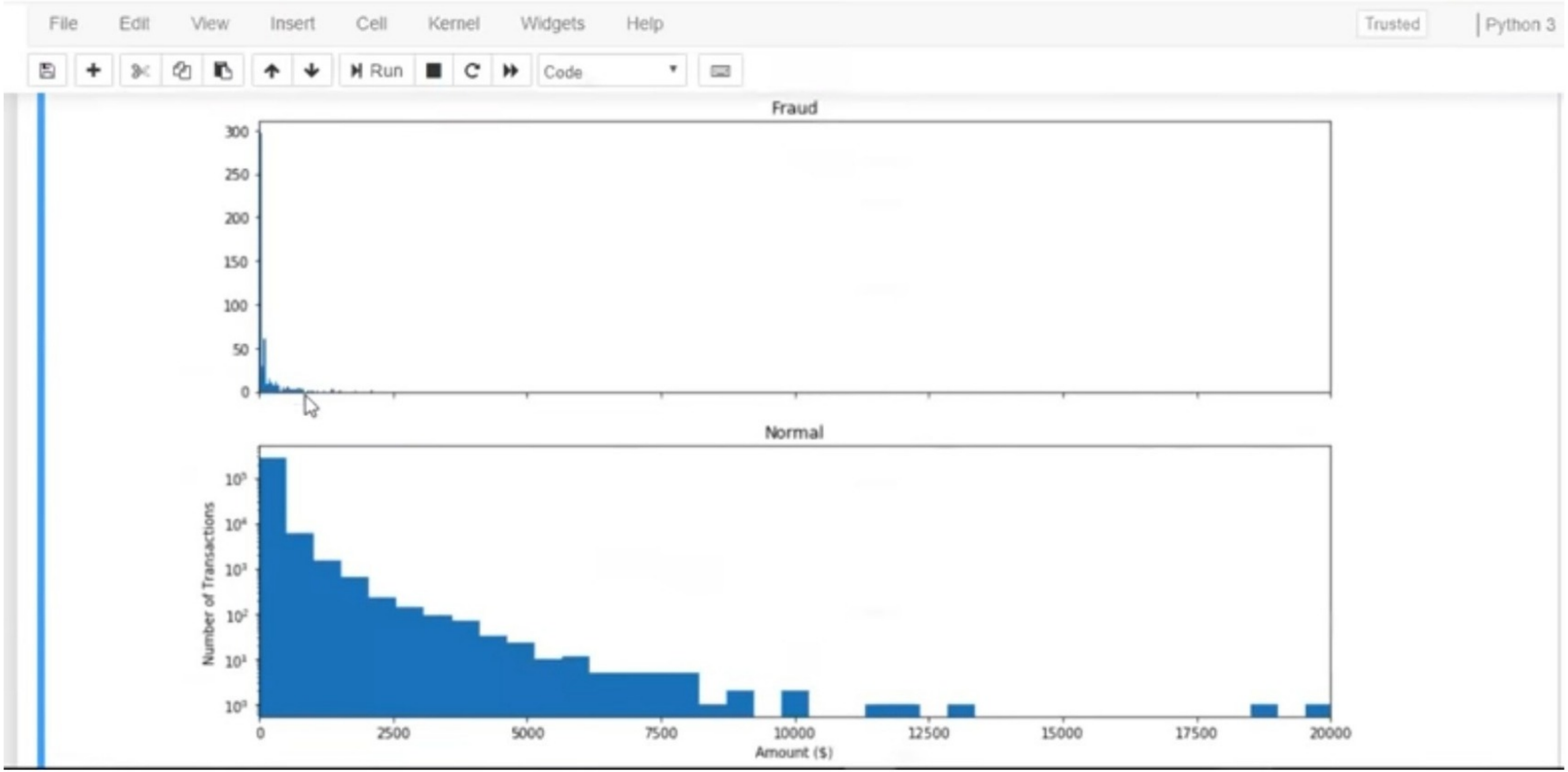
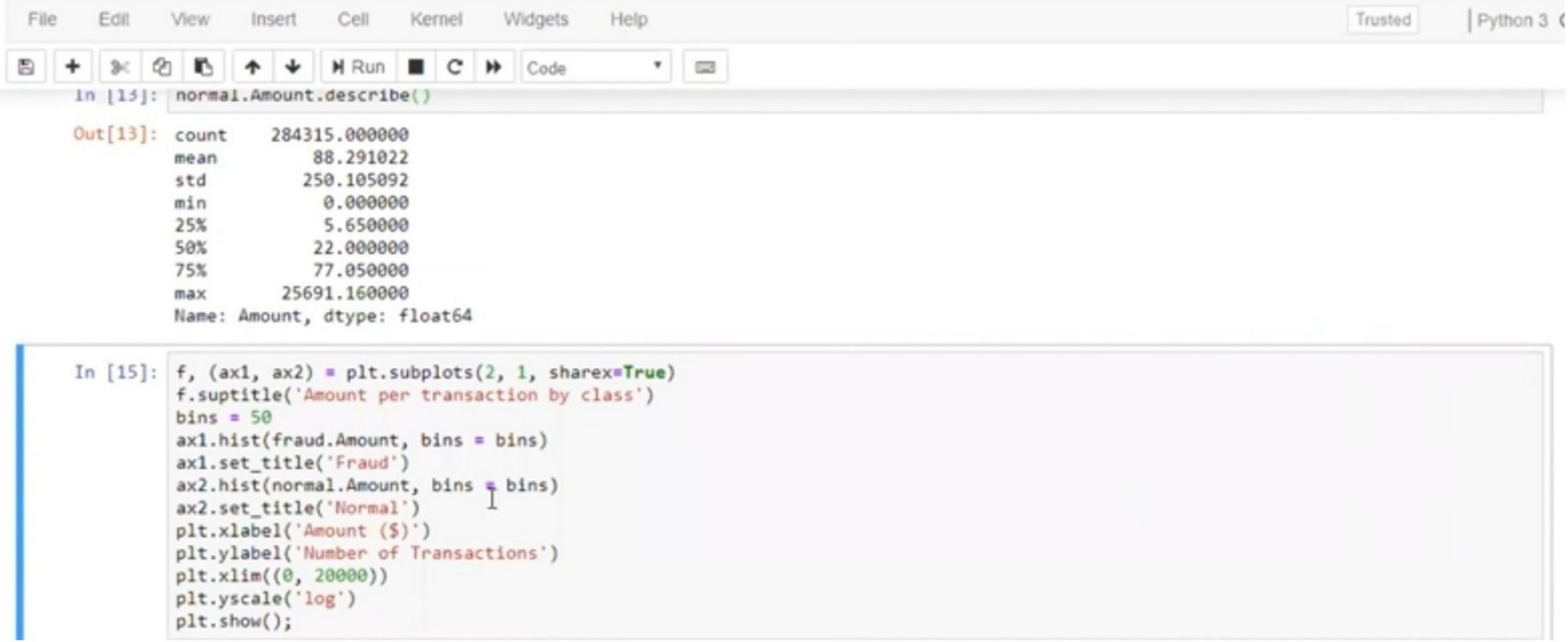


Fig.2. The output of data fraud transaction money used in different classes



```
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();
```

Fig.3.Here we see a data of different different transaction taken in terms of time.

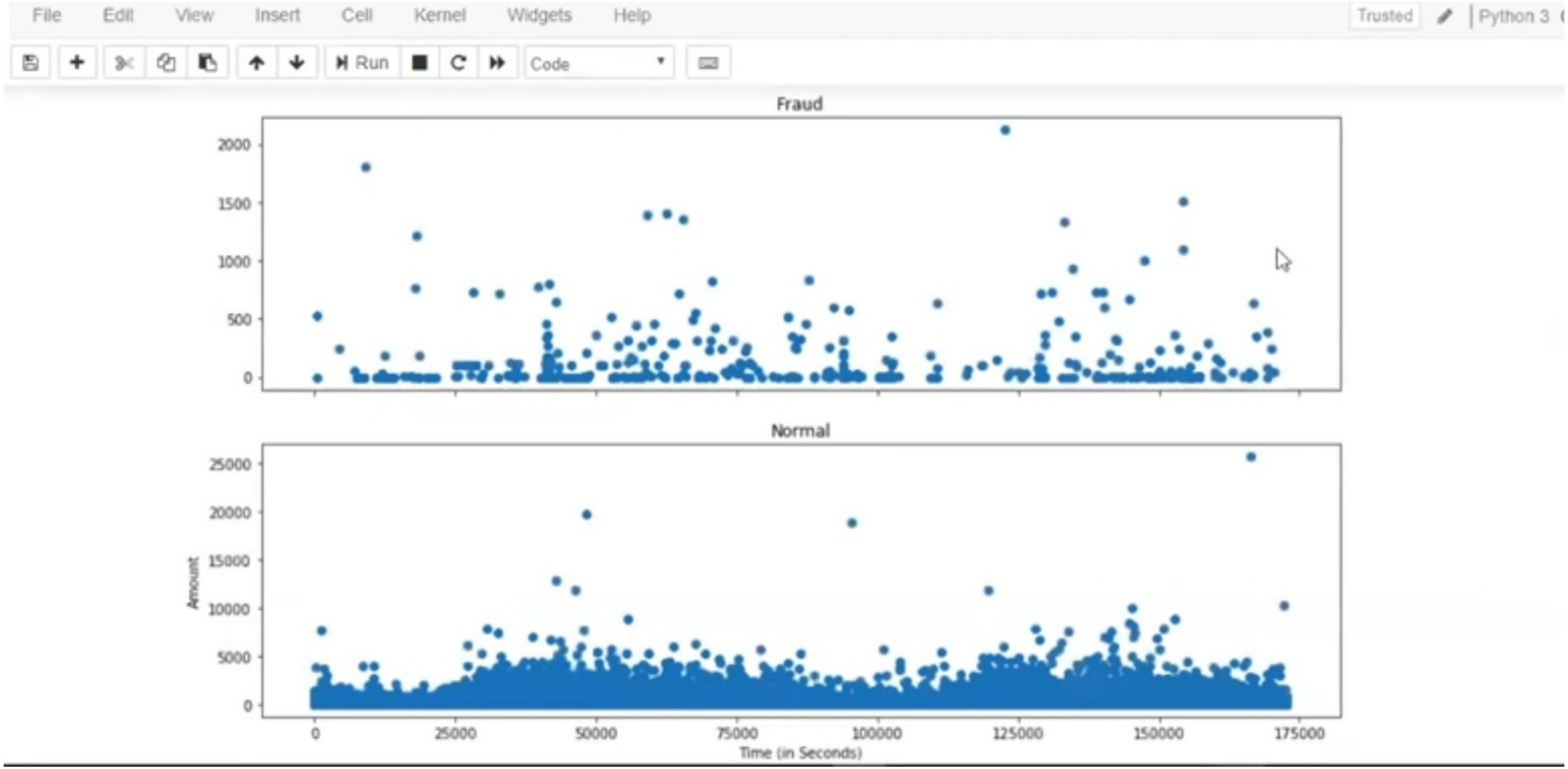


Fig.3.Here we see a data of different different transaction taken in terms of time.

Conclusion and Future Enhancement

For the chosen dataset (Paysim), we show that our proposed approaches are able to detect fraud transactions with very high accuracy and low false positives - especially for TRANSFER transactions. Fraud detection often involves a trade off between correctly detecting fraudulent samples and not misclassifying many non-fraud samples.

We can further improve our techniques by using algorithms like Decision trees to leverage categorical features associated with accounts/users in Paysim dataset. Paysim dataset can also be interpreted as time series. We can leverage this property to build time series based models using algorithms like CNN. Our current approach deals with entire set of transactions as a whole to train our models. We can create user specific models - which are based on user's previous transactional behavior - and use them to further improve our decision making process. All of these, we believe, can be very effective in improving our classification quality on this dataset.

GALGOTIAS
UNIVERSITY
(Established under Galgotias University Uttar Pradesh Act No. 14 of 2011)

SCHOOL OF COMPUTING AND SCIENCE AND ENGINEERING

BONAFIDE CERTIFICATE

Certified that this project report **“FRAUD DETECTION USING MACHINE LEARNING”** is the bonafide work of **“GAURAV SHARMA (1613112021)”** who carried out the project work under my supervision.

SIGNATURE OF GUIDE

Mr Sachin Minocha
(CS) Professor
School of Computing Science & Engineering

SIGNATURE OF SUPERVISOR

Mr Arjun KP
Astt. Professor
School of Computing Science & Engineering