**CONTENTS**

**1. Introduction**

**1.1 Introduction**

**1.2 Objectives of Research**

**1.3 Problem Statement**

**2. Review of Literature**

**3. Data Collection**

**4. Methodology**

**4.1 Exploratory Data Analysis**

**4.1.1 Figures and Tables**

**4.2 Statistical Methods and Data Visualization**

**4.3 Data Modeling Using Supervised ML Techniques**

**5. References**

**6. Conclusion**

**FRAUD DETECTION**

**1. INTRODUCTION:**

* 1. **INTRODUCTION**

Artificial Intelligence is a way of making a computer, a computer-controlled robot, or a software think intelligently, in the similar manner the intelligent humans think. AI is accomplished by studying how human brain thinks and how humans learn, decide, and work while trying to solve a problem, and then using the outcomes of this study as a basis of developing intelligent software and systems.

Machine Learning may be defined as the field of computer science, more specifically an application of artificial intelligence, which provides computer systems the ability to learn with data and improve from experience without being explicitly programmed.

Basically, the main focus of machine learning is to allow the computers learn automatically without human intervention. Now the question arises that how such learning can be started and done? It can be started with the observations of data. The data can be some examples, instruction or some direct experiences too. Then on the basis of this input, machine makes better decision by looking for some patterns in data.

* 1. **OBJECTIVES OF RESEARCH**
* To reduce the causing many financial losses for the customer and the organization.
* A successful strategy for dealing with fraud can quite literally mean millions of dollars in savings per year on operational costs.
* Techniques such as machine learning are used to detect fraudulent transactions.
* To identify the different types of credit card fraud
* Confirming whether a transaction was done by a client or a fraudster is a better option, but by phoning all card holders is cost prohibitive if it is check in all transactions. Further it might also lead to customer dissatisfaction.
* Fraud prevention by automatic fraud detections is where the well-known classification methods can be applied, where pattern 89 recognition systems play a very important role. Past data about the customer is available in huge amounts, which can be mined for useful data. This old data can be analyzed and the buying behavior of the user can be obtained.

**1.3 PROBLEM STATEMENT**

It is important that credit card companies are able to recognize fraudulent credit card transactions so that the customers are not charged for items that they did not purchase. Here we find that your credit card transaction has been fraudulent or not.

**2. REVIEW OF LITERATURE**

On doing the literature survey of various methods for fraud detection we come to the conclusion that to detect credit card fraud there are multiple approaches like

* Logistic Regression
* Decision Tree
* Random Forest
* Support Vector Machine
* KNN

Logistic regression:

Two advanced data mining approaches, support vector machines and random forests, together with the well known logistic regression as part of an attempt to better detect credit card fraud. The study is based on real-life data of transactions from an international credit card operation. It is well understood, easy to use, and remains one of the most commonly used for data-mining in practice. It thus provides a useful baseline for comparing performance of newer methods. Supervised learning methods for fraud detection face two challenges.

Decision Tree:

Decision trees are statistical data mining technique that express independent attributes and a dependent attributes logically AND in a tree shaped structure. Classification rules, extracted from decision trees, are IF-THEN expressions and all the tests have to succeed if each rule is to be generated .Decision tree usually separates the complex problem into many simple ones and resolves the sub problems through repeatedly using. Decision trees are predictive decision support tools that create mapping from observations to possible consequences. There are number of popular classifiers construct decision trees to generate class models. The work demonstrates the advantages of applying the data mining techniques including decision trees and SVMs to the credit card fraud detection problem for the purpose of reducing the bank’s risk. The results show that the proposed classifiers of C&RT and other decision tree approaches outperform SVM approaches in solving the problem under investigation.

Random Forest:

The popularity of decision tree models in data mining arises from their ease of use, flexibility in terms of handling various data attribute types, and interpretability. Single tree models, however, can be unstable and overly sensitive to specific training data. Ensemble methods seek to address this problem by developing a set of models and aggregating their predictions in determining the class label for a data point. A random forest model is an ensemble of classification (or regression) trees.

Ensembles perform well when individual members are dissimilar, and random forests obtain variation among individual trees using two sources for randomness: first, each tree is built on separate bootstrapped samples of the training data; secondly, only a randomly selected subset of data attributes is considered at each node in building the individual trees. Random forests thus combine the concepts of bagging, where individual models in an ensemble are developed through sampling with replacement from the training data, and the random subspace method, where each tree in an ensemble is built from a random subset of attributes.

Support Vector Machine:

The basic idea of SVM classification algorithm is to construct a hyper plane as the decision plane which making the distance between the positive and negative mode maximum .The strength of SVMs comes from two important properties they possess - kernel representation and margin optimization. Kernels, such as radial basis function (RBF) kernel, can be used to learn complex regions. A kernel function represents the dot product of projections of two data points in a high dimensional feature space.

KNN:

KNN can be used for both classification and regression predictive problems. However, it is more widely used in classification problems in the industry. The performance of KNN algorithm can be improved by using a genetic algorithm for optimizing the distance metric. This technique required legitimate as well as fraudulent samples of data for training.

To evaluate any technique we generally look at 3 important aspects:

1. Ease to interpret output

2. Calculation time

3. Predictive Power

**3. DATA COLLECTION**

The datasets contains transactions made by credit cards in September 2013 by european cardholders. This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.

It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, ... V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-senstive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise.

**4. METHODOLOGY**

Credit card transactions continue to grow in number, taking an ever-larger share of the US payment system and leading to a higher rate of stolen account numbers and subsequent losses by banks. Improved fraud detection thus has become essential to maintain the viability of the US payment system. Banks have used early fraud warning systems for some years. Large-scale data-mining techniques can improve on the state of the art in commercial practice. Scalable techniques to analyze massive amounts of transaction data that efficiently compute fraud detectors in a timely manner is an important problem, especially for e-commerce. Besides scalability and efficiency, the fraud-detection task exhibits technical problems that include skewed distributions of training data and non-uniform cost per error, both of which have not been widely studied in the knowledge-discovery and data mining community.

Credit card transactions have become the de facto standard for Internet and Web based e-commerce. However, the growing number of credit card transactions provides more opportunity for thieves to steal credit card numbers and subsequently commit fraud.

When banks lose money because of credit card fraud, cardholders pay for all of that loss through higher interest rates, higher fees, and reduced benefits. For many years, the credit card industry has studied computing models for automated detection systems; recently, these models have been the subject of academic research, especially with respect to e-commerce.

Credit card fraud detection has drawn a lot of research interest and a number of techniques, with special emphasis on data mining and neural networks, have been suggested Gosh and Reilly have proposed credit card fraud detection with a neural network. They have built a detection system, which is trained on a large sample of labeled credit card account transactions.

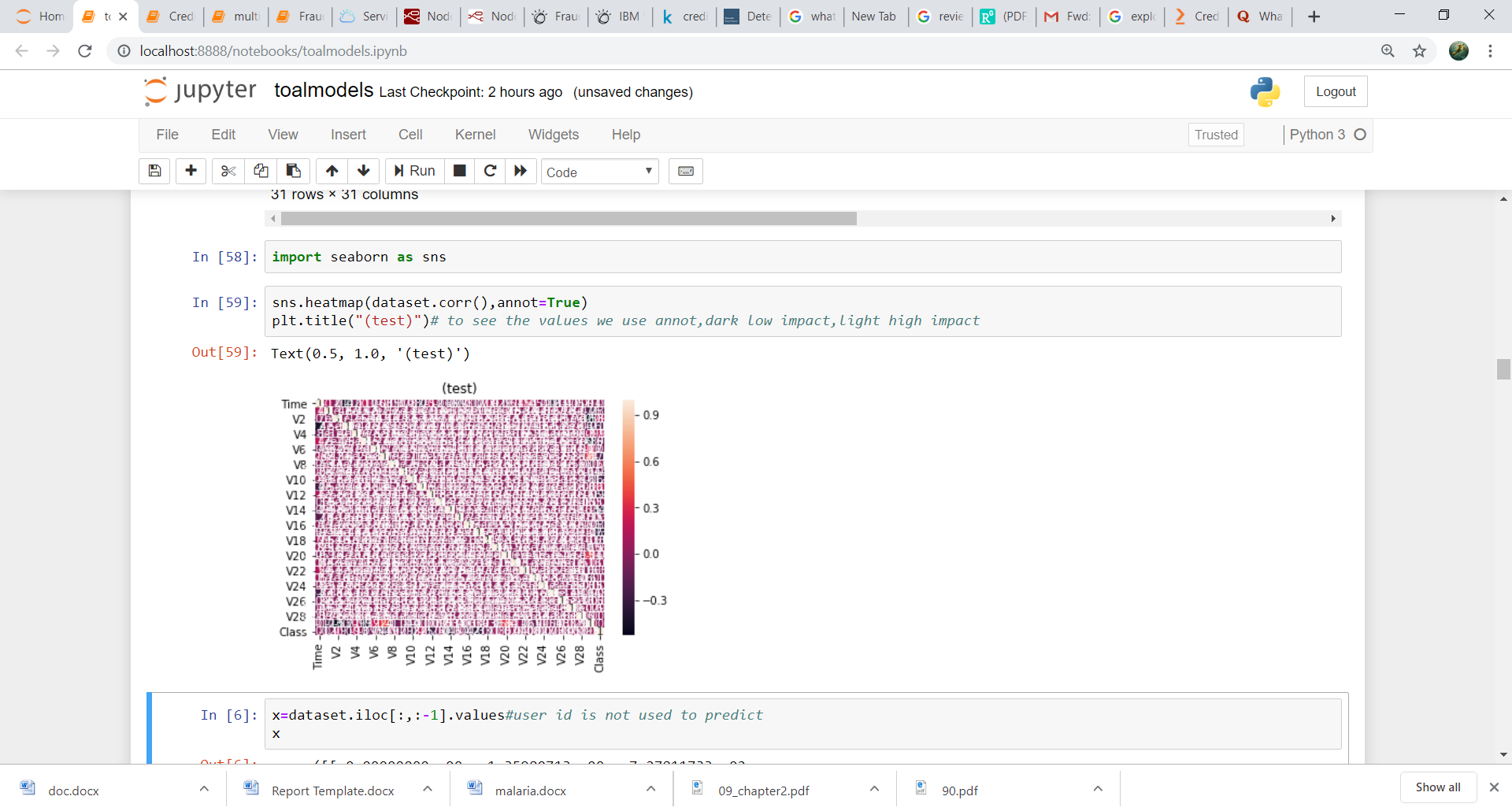
**4.1. EXPLORATORY DATA ANALYSIS**

Since nearly all predictors have been anonymized, I decided to focus on the non-anonymized predictors time and amount of the transaction during my EDA. The data set contains 284,807 transactions. The mean value of all transactions is $88.35 while the largest transaction recorded in this data set amounts to $25,691.16. The majority of transactions are relatively small and only a tiny fraction of transactions comes even close to the maximum.

The time is recorded in the number of seconds since the first transaction in the data set. Therefore, we can conclude that this data set includes all transactions recorded over the course of two days.

From the dataset we observed that most transactions are non-fraudulent. In fact, 99.83% of the transactions in this data set were not fraudulent while only 0.17% were fraudulent.

**4.1.1: TABLES AND FIGURES:**



**Figure:1 Heat Map**

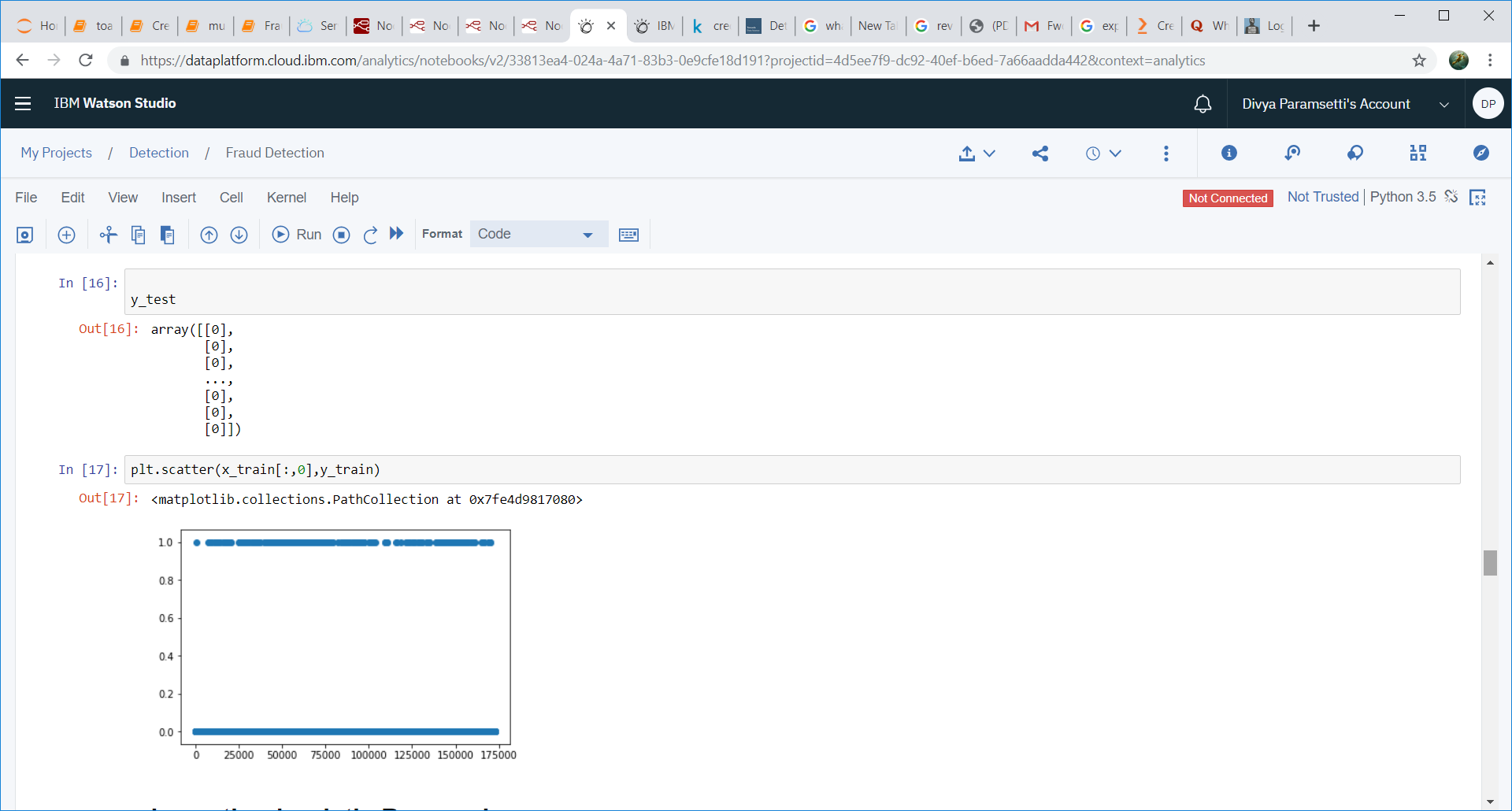


Fig 2 : Scatter plots

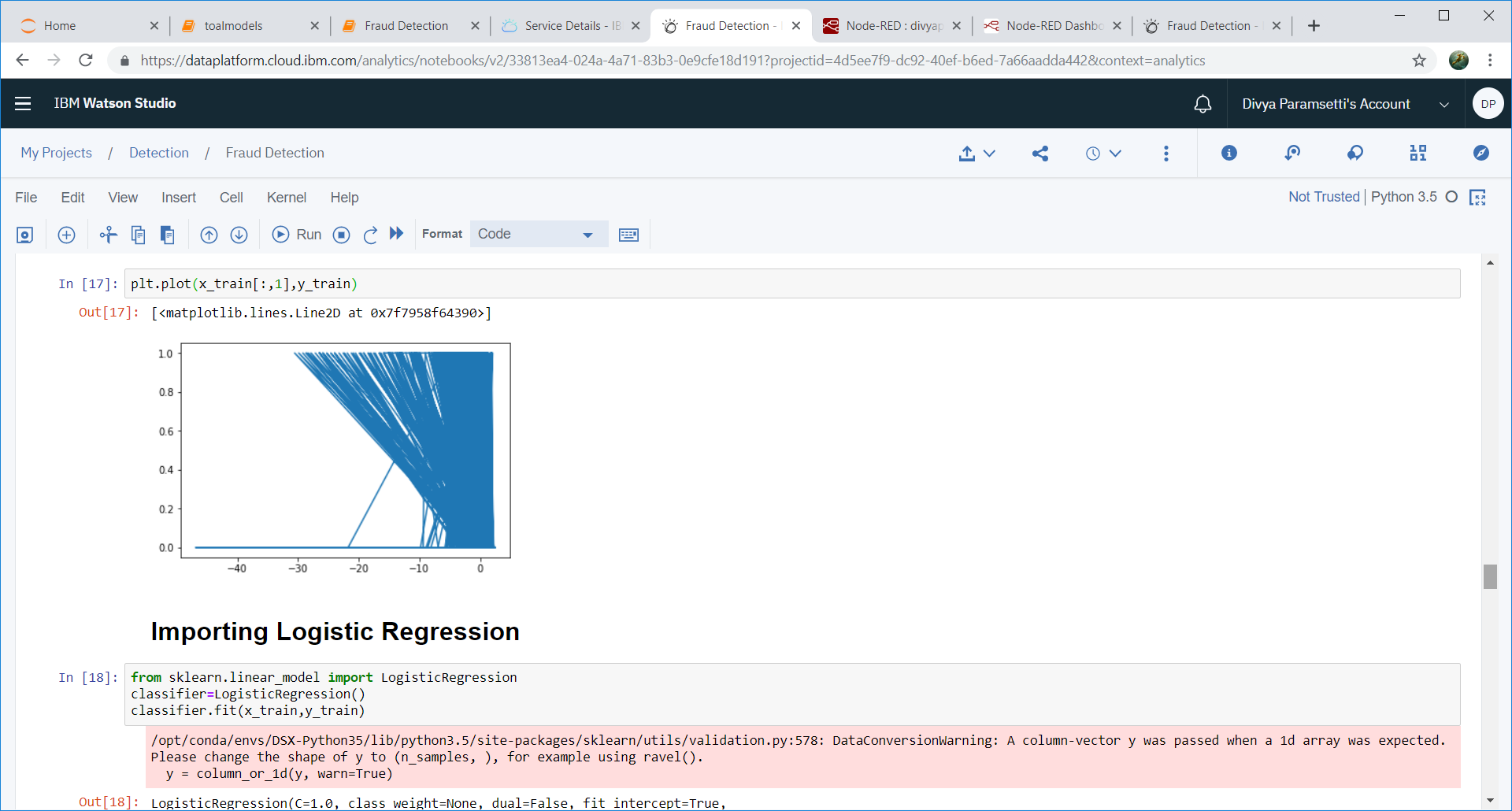


Fig 3: Plots

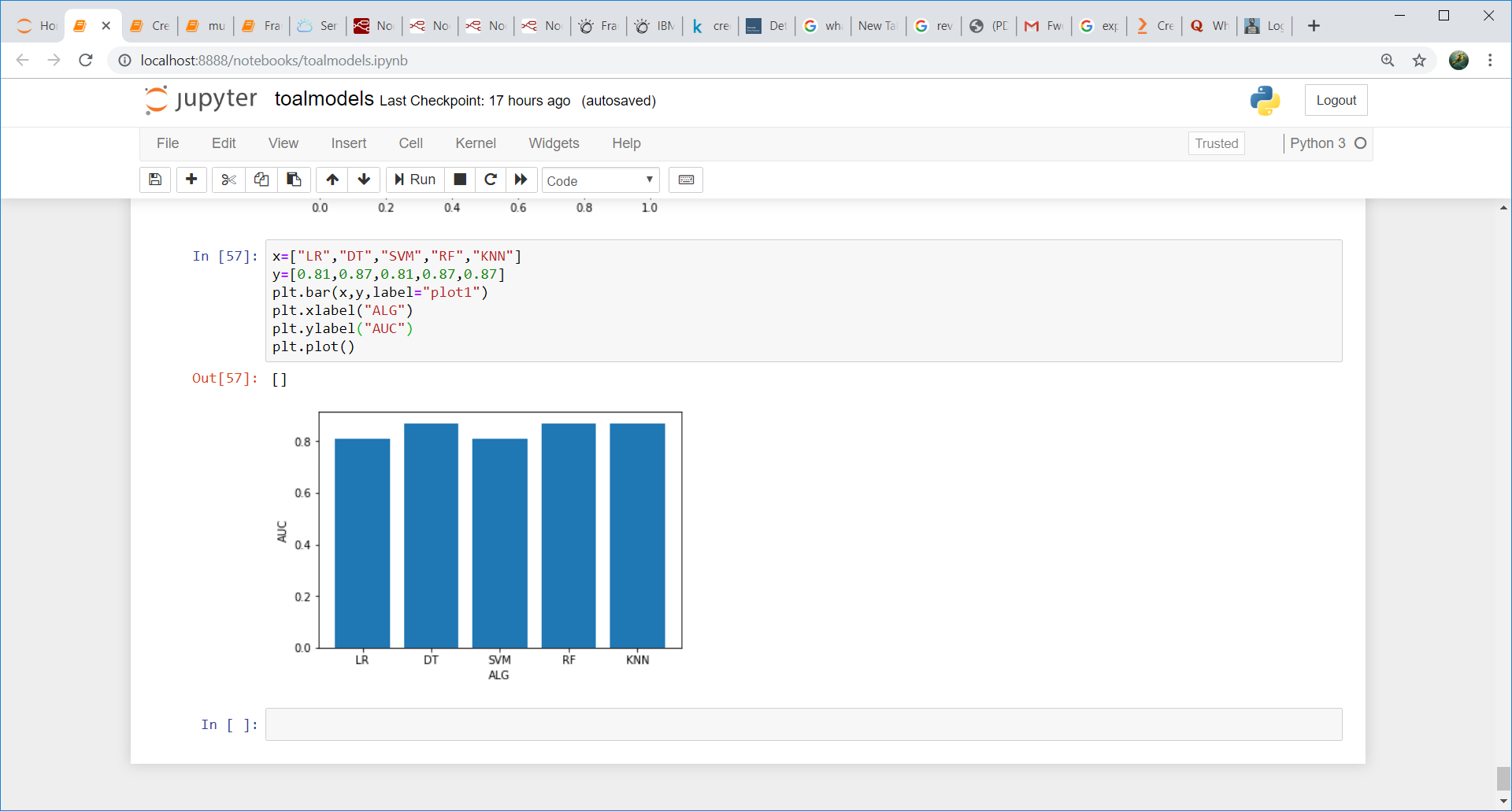


Fig 4: Bar Charts

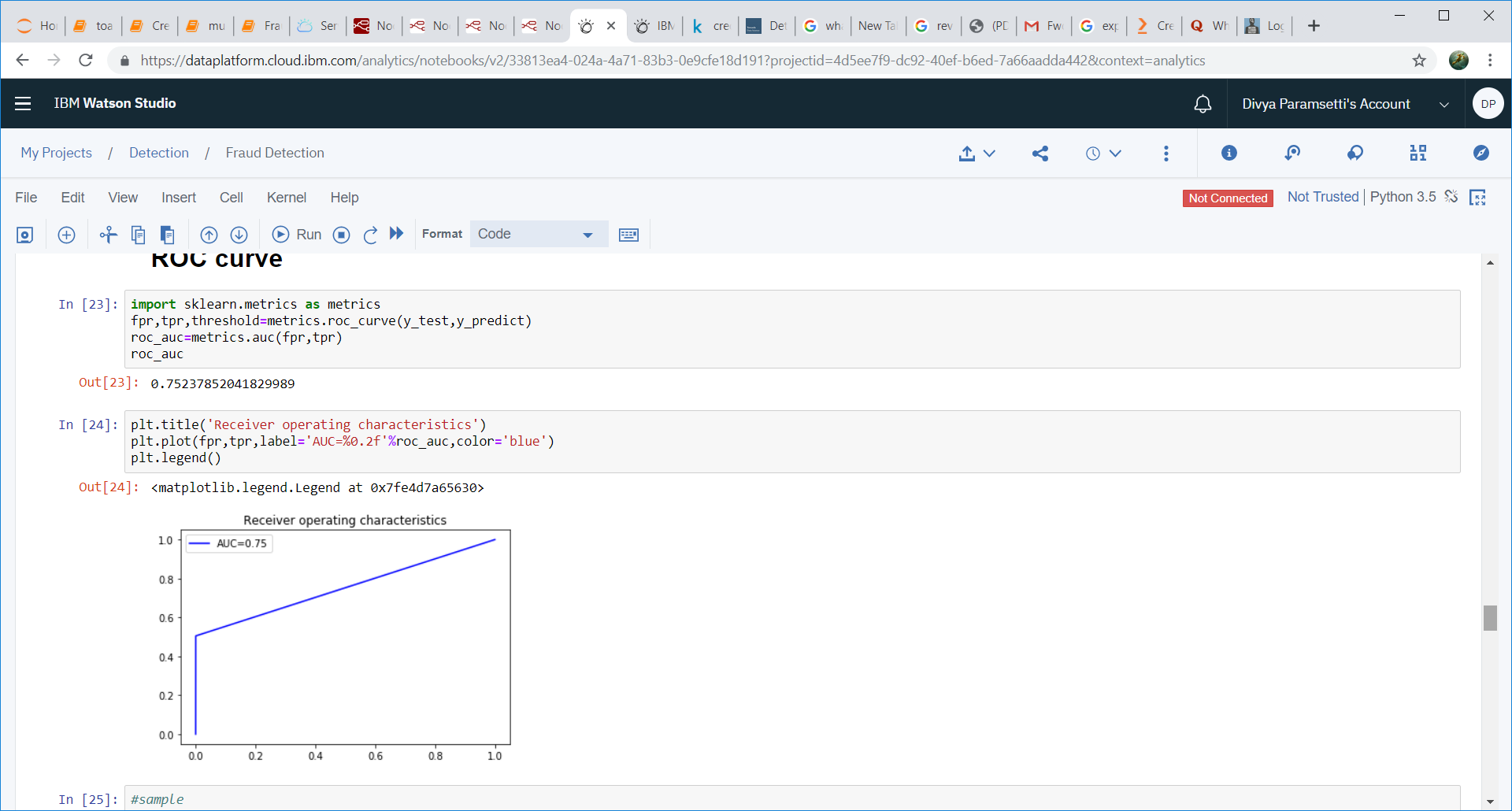


Fig 5: ROC-AUC Curve

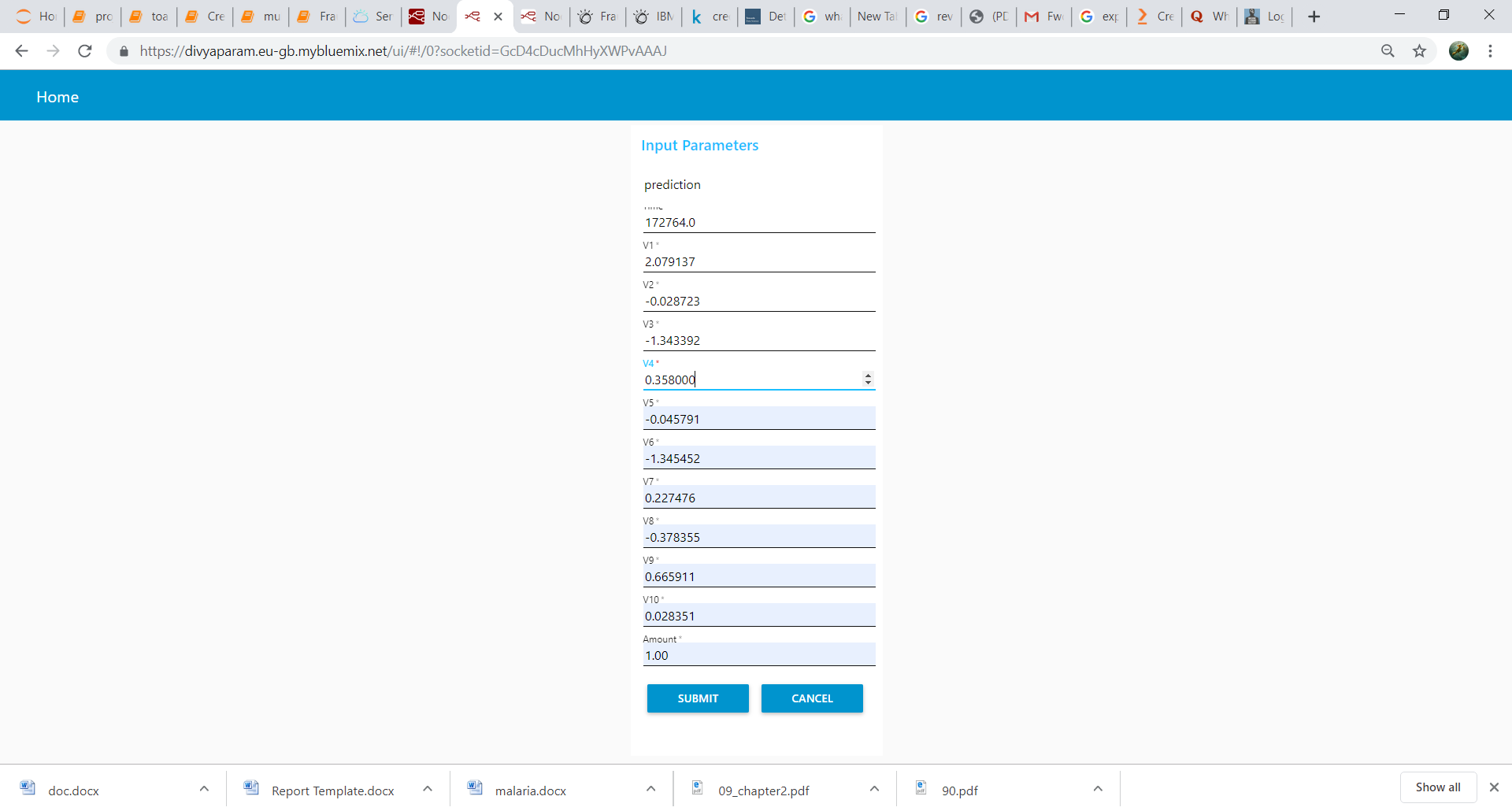


Fig 6: Predicted Values

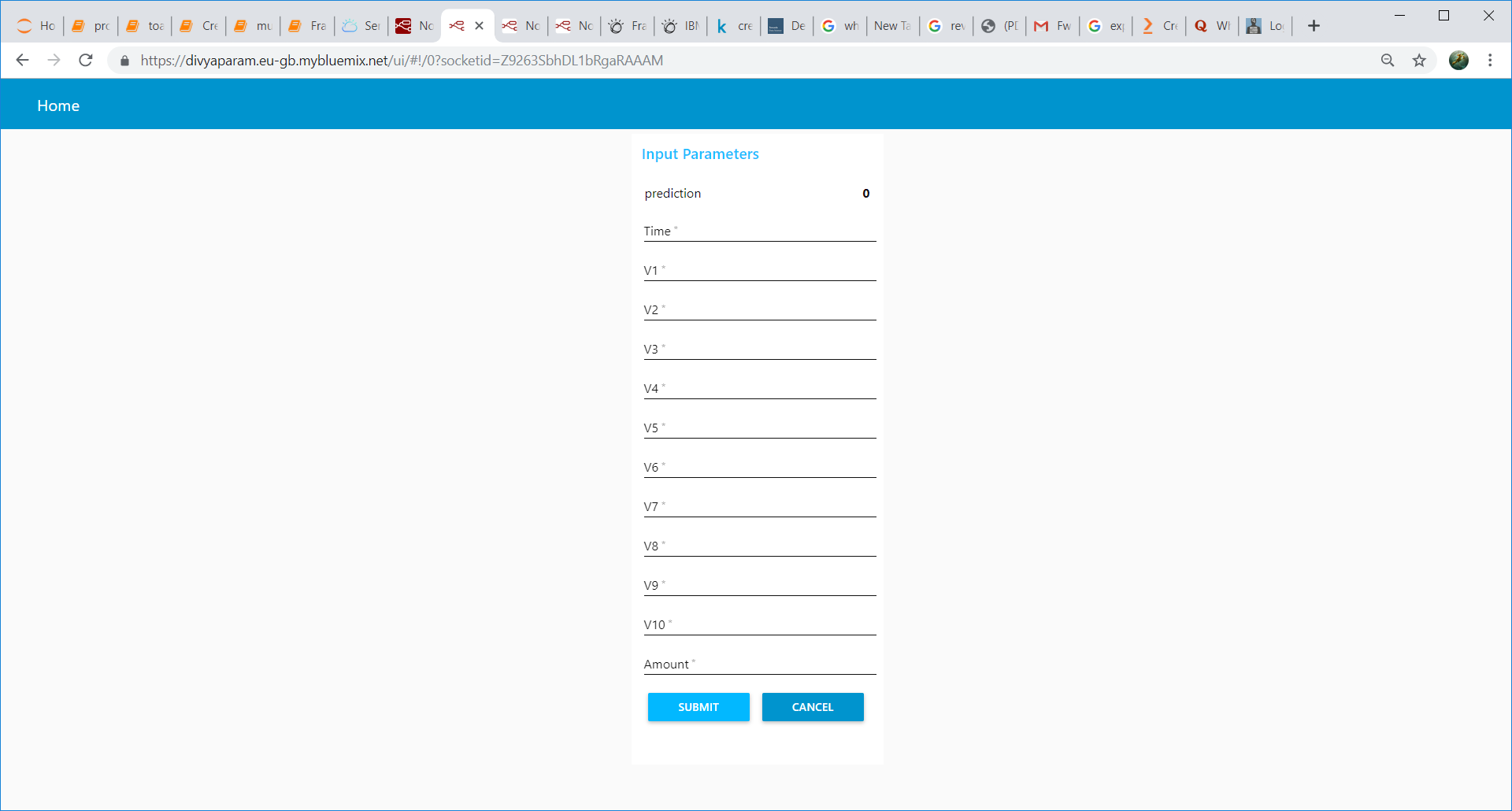


Fig 7: Predicted value

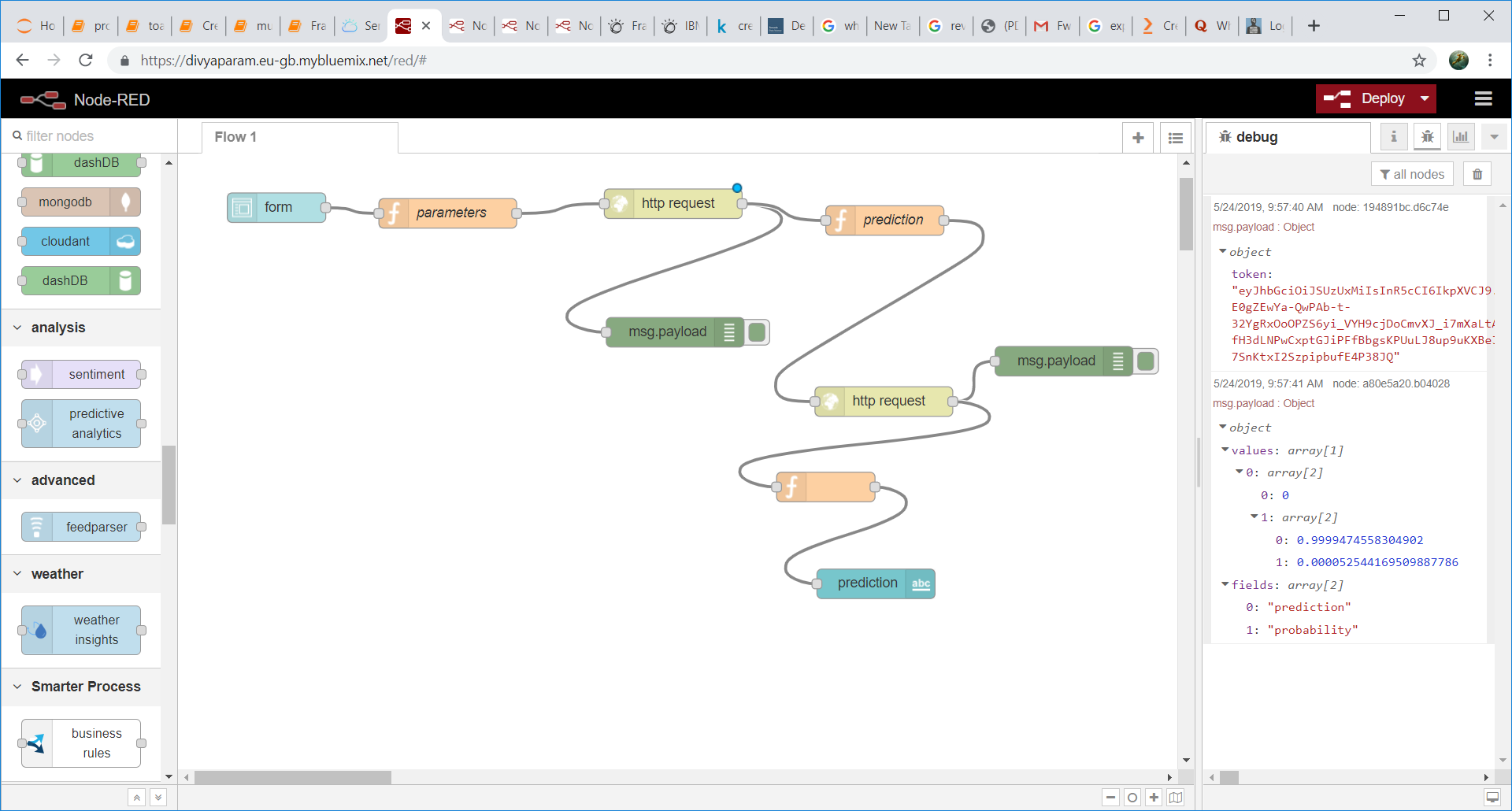


Fig 8: Flow of User Interface

**4.2 STATISTICS TECHNIQUES AND DATA VISUALIZATION:**

Credit card databases contain information on each transaction. This information includes such things as merchant code, account number, type of credit card, type of purchase, client name, size of transaction and date of transaction. Some of these data are numerical (e.g., transaction size) and others are nominal categorical (e.g., merchant code, which can have hundreds of thousands of categories) or symbolic. The mixed data types have led to the application of a wide variety of statistical, machine learning and data mining tools From our experience, the popular models to choose tend to be logistic regression, SVM(Support Vector Machine) and Random Forests.

Naturally, the actual models that we build to detect fraud are statistically based. PCA is also useful in feature engineering, allowing you to view clusters of high dimensional customer behaviour to gain an understanding of the distribution of behaviour among different groups of customers Overall, statistical techniques are extremely useful in fraud detection.

**Visualization** techniques take advantage of the human perception system and allow analysts to more easily derive insights about data. For instance, instead of exhaustively looking into tables to identify data characteristics, an analyst is able to see, explore, and understand a large amount of information by using visualization techniques. However, the efficiency of these techniques varies with respect to different tasks. The following list of visualization techniques are used in the identified fraud detection approaches:

**Bar charts:** Rectangles (bars) are used to represent different entities, where the height or the width encodes quantitative values. For example, we have an example of stacked bar charts that is being utilized for ranking. This visualization technique is well suited to represent relative differences. There are different types of bar charts such as horizontal bar charts, stacked bar charts, and range bar charts that were classified equally in this category.

**Scatter plots:** These are graphs where each sample is represented by a point or symbol. Each point or symbol position is defined according to two dimensions, or two generated features of these samples. Those graphs are useful to illustrate trends and correlations.

**Heat Maps:** This graphical representation represents values by colors. In fraud detection this technique is usually used to visually query for patterns or outliers in a large amount of data.

**4.3 DATA MODELING:**

Generally we have two types of algorithms in machine learning that are Supervised learning and Unsupervised learning. Our model is comes under supervised learning algorithm.

In supervised learning algorithm we have Classification and Regression. Our model is one of the application of Classification. Here we have some input variables to predict the output.

In this project we use the model Logistic regression. Because it gives the better accuracy when compared to other models. In logistic regression we will use the concept of probabilities. So the values will be ranging from 0 to 1.Depending on probability parameters we have to divide the classes into class ” 1” and class “0”.

We are evaluating the model performance by using Confusion matrix and AUC-ROC curves(Measurement for classification problem).Higher the AUC ,better the model is at predicting 0’s as 0’s and 1’s as 1’s.

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is binary.  Like all regression analyses, the logistic regression is a predictive analysis.  Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more independent variable.

To implement the Logistic regression model using Python , we have to import the few libraries

**from sklearn.linear\_model import LogisticRegression**

An Object is created for Logistic Regression,

**classifier=LogisticRegression()**

We have to fit the model by training data , so we using fit method,

**classifier.fit(x\_train,y\_train)**

We have to predict the values by using predict method and predicted values are stored in a variable,

**y\_predict=classifier.predict(x\_test)**

At last we find the accuracy by using accuracy\_score, confusion\_matrix and ROC-AUC Curve which is imported from skleran.metrics.

**5. REFERENCES**

1. <https://www.kaggle.com/mlg-ulb/creditcardfraud>

2. <https://towardsdatascience.com/detecting-credit-card-fraud-using-machine-learning-a3d83423d3b8>

3. <https://en.wikipedia.org/wiki/Credit_card_fraud>

4. <https://www.vantiv.com/vantage-point-fi/safer-payments/credit-card-fraud-detection-solutions>

5. <https://www.researchgate.net/publication/329412906_Learning_model_for_fraud_detection_in_credit_cards>

**6. CONCLUSION**

Although there are several fraud detection techniques available today but none is able to detect all frauds completely when they are actually happening, they usually detect it after the fraud has been committed. This happens because a very minuscule number of transactions from the total transactions are actually fraudulent in nature. So we need a technology that can detect the fraudulent transaction when it is taking place so that it can be stopped then and there and that too in a minimum cost. So the major task of today is to build an accurate, precise and fast detecting fraud detection system for credit card frauds. The major drawback of all the techniques is that they are not guaranteed to give the same results in all environments. They give better results with a particular type of dataset and poor or unsatisfactory results with other type.