# Stock Positions Analysis and Prediction using Machine Learning

Vijay Pawar
vp50395p@pace.edu
Pace University, New York

Abstract—The paper has an analysis model for predicting the stock values of various companies at the end of the day based on data visualization using matplotlib and ggmap libraries with data analysis algorithm, that is Logistics Regression. Here we predict the values of stocks for companies in various sectors around the world. In particular focusing on a stock value each day so that it improves the chance of more revenue. Sale of bad performance stocks as it would mitigate losses and would allow clients to invest into much better performing stocks. Finally, we predicted the stock performances on the bases of sectors.

Keywords—Stocks, Logistics Regression, Visualization, Matplotlib, Seaborn,

## I. INTRODUCTION

Stock market is a collection of markets and exchanges where regular activities of buying, selling, and issuance of shares of publicly-held companies take place. Such financial activities are conducted through institutionalized formal exchanges or over-the-counter (OTC) marketplaces which operate under a defined set of regulations. There can be multiple stock trading venues in a country or a region which allow transactions in stocks and other forms of securities. It is one of the oldest methods where a normal person would trade stocks, make investments and earn some money out of companies that sell a part of themselves on this platform.

While today it is possible to purchase almost everything online, there is usually a designated market for every commodity. For instance, people drive to city outskirts and farmlands to purchase Christmas trees, visit the local timber market to buy wood and other necessary material for home furniture and renovations, and go to stores like Walmart for their regular grocery supplies. Such dedicated markets serve as a platform where numerous buyers and sellers meet, interact and transact. Since the number of market participants is huge, one is assured of a fair price. A stock market is a similar designated market for trading various kinds of securities in a controlled, secure and managed the environment. Since the stock market brings together hundreds of thousands of market participants who wish to buy and sell shares, it ensures fair pricing practices and transparency in transactions.

Stock markets provide a secure and regulated environment where market participants can transact in shares and other eligible financial instruments with confidence with zero- to low-operational risk. The stock exchanges also maintain all company news, announcements, and financial reporting, which can be usually accessed on their official websites. A stock exchange also supports various other corporate-level, transaction-related activities.

Now if we try to graph the stock exchange price over the time period (say 6 months), is it really hard to predict the next outcome on the graph?

In statistics, there is a way where we look at the values and attributes of a problem in a graphs and identify the dependents and independent variables and try to establish or identify an existing relationship amongst them. This technique is known as linear regression. It is very commonly used due to its very simple and effective approach.

In machine learning we have adapted the same algorithm where we use the features to train the classifier which then predicts the value of the label with certain accuracy which can be checked while training and testing of the classifier.

## II. LITERATURE SURVEY

A thorough literature survey was performed to get a better understanding of the topic, analyze the previous models developed, note their advantages and drawbacks, and highlight the necessary developments. The methodology adopted is discussed in detail in the following section.

## III. PREDICTIOIN MODEL

*A.* Data Analysis Stage

In this stage, we shall look at the raw data available to us and study it in-order to identify suitable attributes for the prediction of our selected label. The dataset taken is in .HTML form. All the data needs to be extracted from the HTML file in order to visualize and predict the stock market values. After extraction a .csv file was created using pandas datafrane which was later imported into the PostgreSQL for further operations. From the .HTML file four .CSV files were genertaed which were used throughout the project for various operations. All this data was first loaded into the PostgreSQL database using PgAdmin4. For cleansing data, the Pandas framework of Python was used. After eliminating Missing values, the dataset was first loaded into the database. The ETL process was then completed for further implementation of Linear Regression.

*B.* Linear Regression

Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent variables. Logistic regression generates the coefficients of a formula to predict a logit transformation of the probability of presence of the characteristic of interest. In order to predict data from the created dataset firstly we need to train the data. Using Logistic Regression approach, we first need to understand the dependent and independent attributes in the dataset.

A. Dependent variable - variable whose values you want to predict. The dependent variable must be binary or dichotomous, and should only contain data coded as 0 or 1.

B. Independent variables: Select the different variables that you expect to influence the dependent variable.

```
Shape of complete dataset is  (505, 9)
Length of y_test is  3
Shape of test sample is  (3, 2)
x_input = [-0.79415228  2.10495117], predicted_output = 0
x_input = [-8.25290074 -4.71455545], predicted_output = 1
x_input = [-2.18773166  3.33352125], predicted_output = 0
```

In our Project the Dependent variable is the "Sectors" attribute which is dependent on the independent "Values" attribute of the stock holding companies. When training the dataset, the x values were taken to be "Sectors" attribute and y values were taken to be "Values" attribute.

After applying Logistic Regression model, the prediction was further made for stock values which would earn higher values at the end of the day and hence be higher at the beginning of the next day.

## IV. RELEVANCE OF THE PROJECT

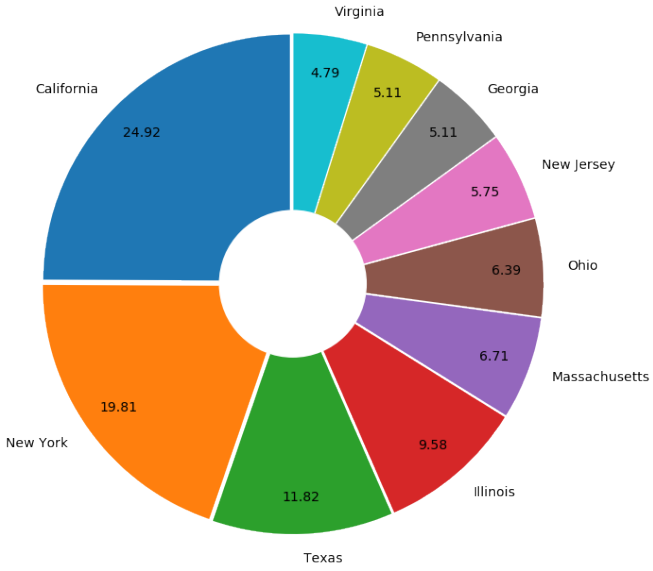
This project is quite relevant as it guides people who possess limited know-how of investments and finance into making well informed decisions regarding stock market investments. It bypasses the need for hiring investment experts who command exorbitant wages to guide our financial decisions

by providing a simple solution which can be accessed by anyone having a computer or a laptop and an internet connection. Stock market trends for a given time frame can be analyzed easily even by the uninformed. Popularizing this machine learning option provides cheap alternative to various stock market investment guidance agencies which are in vogue today. The project puts in a small effort to assist the inexperienced investors and prevent from suffering heavy capital loss.

## V. DATA VISUALIZATION

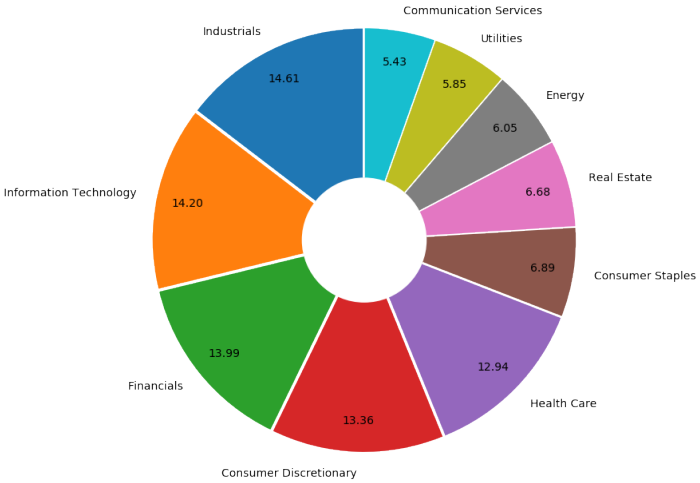
In today's world where everything is recorded digitally, right from our web surfing patterns to our medical records, we are generating and processing petabytes of data every day.

Big data will be transformative in every sphere of life. But just to process and analyze those data is not enough, human brain tends to find pattern more efficiently when data is represented visually.




[DISTRIBUTION OF TOP 10 STATES]

Almost all fields of study and practice sooner or later will confront the big-data problem. Visualization has proven effective for not only presenting essential information in vast amounts of data but also driving complex analyses. In this project, our visualization techniques indicate different levels of abstraction, understanding, or truthfulness.




[DISTRIBUTION OF TOP 10 SECTORS]

## VI. PROJECT REQUIREMENTS

A. Programming Language

Python will be the programming language utilized throughout the project.

All project code is stored in Jupyter Notebooks.

B. Machine Learning Framework and Libraries:

- Scikit Learn Library

D. Data Processing Framework and Libraries:

- Pandas
- Numpy

E. Visualization:

- Matplotlib
- Seaborn

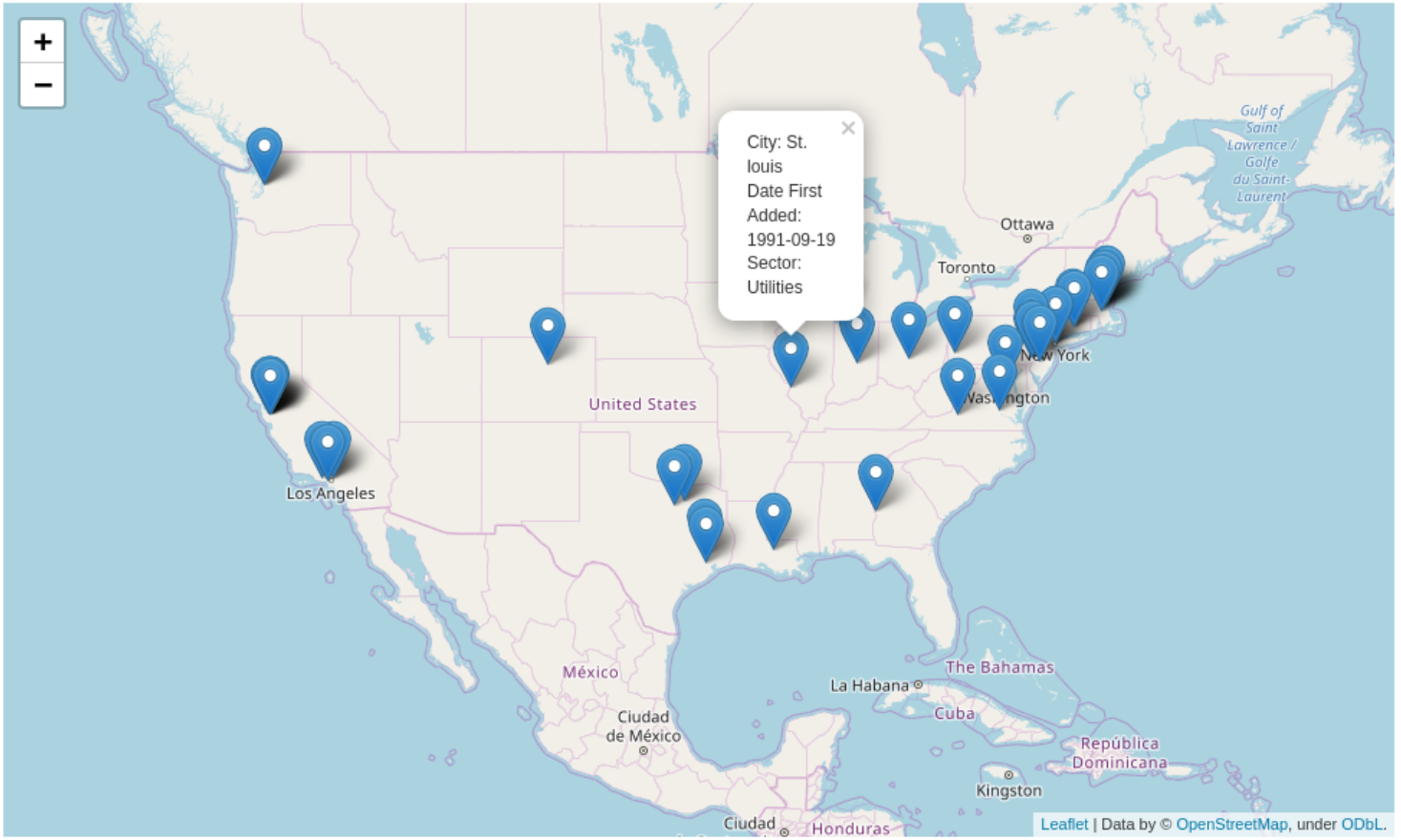
[Office headquarters of the offices plotted using the Google Geo-location API using the latitude and longitudes in the dataset.]

Edit View Insert Cell Kernel Widgets Help Trusted Python 3

Run Code

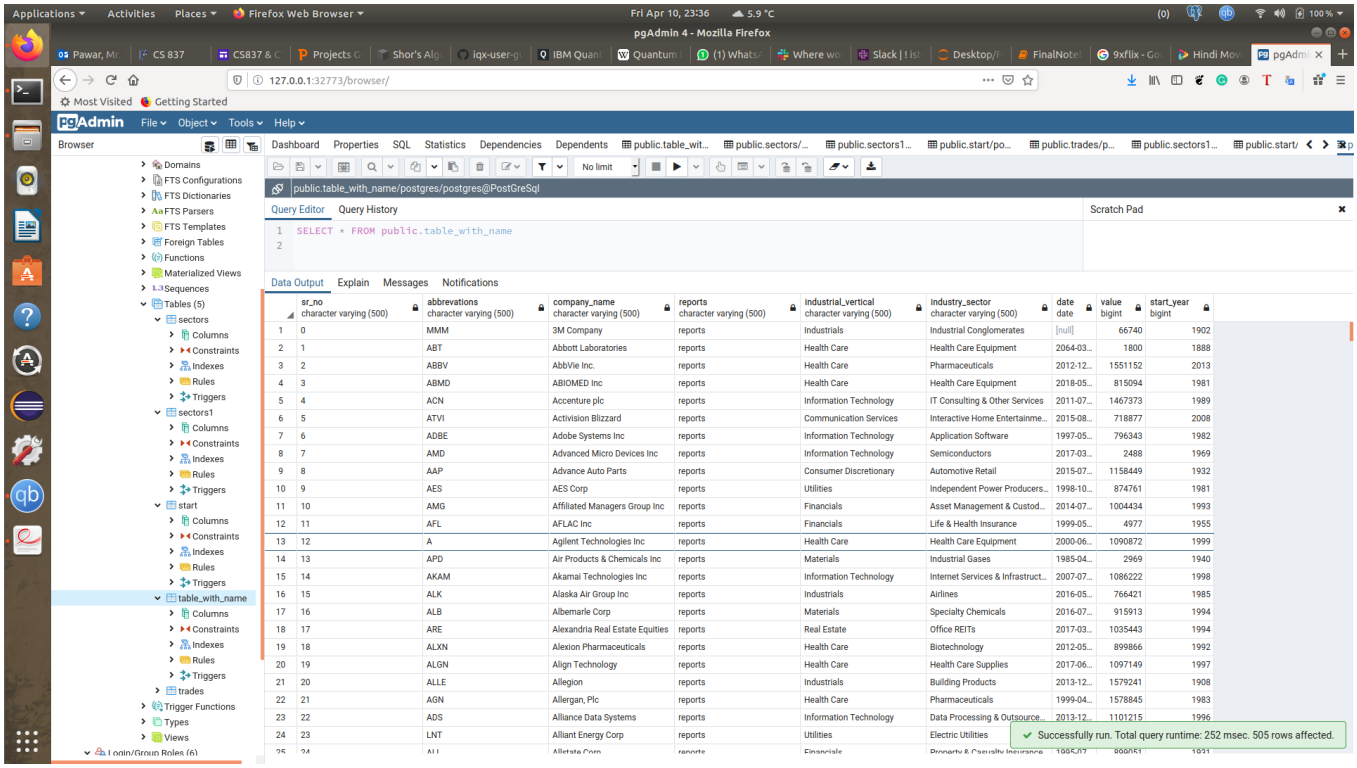
```
    plt.tight_layout()
    plt.show()
```

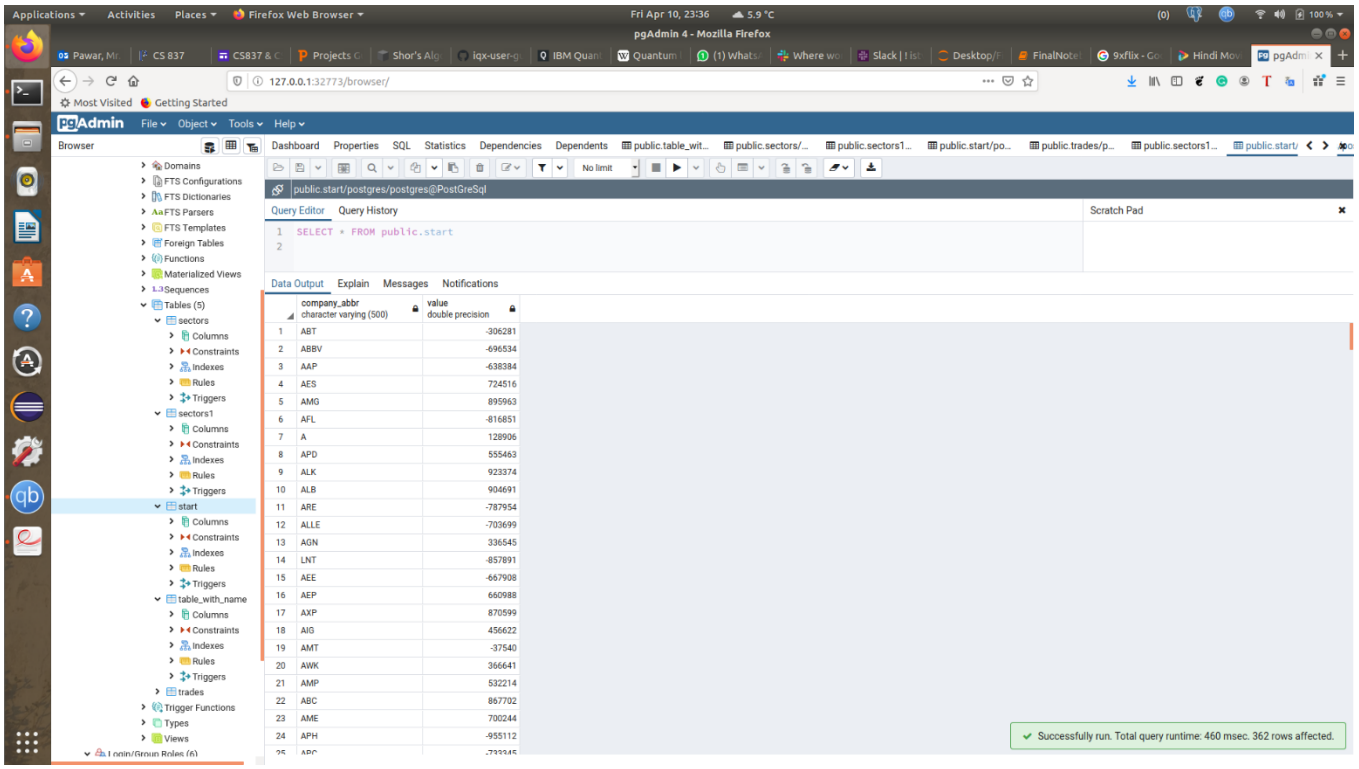
In [11]:
```
generateMap()
print("\nHTML for the map is created!\n")
```




```
HTML for the map is created!
```

[Complete dataset files were loaded in the Postgres using pgAdmin.]



## VII. CONCLUSION

The aim of our research study is to help the stock brokers and investors for investing money in the stock market. The prediction plays a very important role in stock market business which is very complicated and challenging process dynamic nature of the stock market.

## VIII. REFERENCES


1) A SHETA, "SOFTWARE EFFORT ESTIMATION AND STOCK MARKET PREDICTION USING TAKAGI-SUGENO FUZZY MODELS", IN PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS, PP.171- 178, VANCOUVER, BC, 2006.

2) A SHETA, "SOFTWARE EFFORT ESTIMATION AND STOCK MARKET PREDICTION USING TAKAGI-SUGENO FUZZY MODELS", IN PROCEEDINGS OF THE IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS, PP.171- 178, VANCOUVER, BC, 2006.

3) M.H. FAZELZARANDI, B. REZAEE, I.B. TURKSEN AND E. NESHAT, "A TYPE-2 FUZZY RULE-BASED EXPERT SYSTEM MODEL FORSTOCK PRICE ANALYSIS", EXPERT SYSTEMS WITH APPLICATIONS, VOL.36, NO.1, PP. 139-154, JANUARY 2009.

4) ROBERT K. LAI, CHIN-YUAN FAN, WEI-HSIU HUANG AND PEI-CHANN CHANG, "EVOLVING AND CLUSTERING FUZZY DECISIONTREE FOR FINANCIAL TIME SERIES DATA FORECASTING", AN INTERNATIONAL JOURNAL OF EXPERT

5) SYSTEMS WITH APPLICATIONS,VOL.36, NO.2, PP. 3761-3773, MARCH 2009.

6) SHYI-MING CHEN AND YU-CHUAN CHANG, "MULTI-VARIABLE FUZZY FORECASTING BASED ON FUZZY CLUSTERING AND FUZZY RULE INTERPOLATION TECHNIQUES", INFORMATION SCIENCES, VOL.180, NO.24, PP. 4772-4783, 2010.