Restaurant Content Based Recommendation System

Submitted for the Summer Internship

on

Machine Learning and Deep Learning

(from 8th June, 2021 to 31st July, 2021)

Organised by

DST Centre of Excellence – Artificial Intelligence, IGDTUW

IGDTUW-Anveshan Foundation

Department of AI and Data Sciences, IGDTUW

By

Indu Rani
M.Tech CSE(AI)
2020-2022

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN
(Established by Govt. of Delhi vide Act 09 of 2012)
Kashmere Gate, Delhi-110006

Index

Certificate

INDIRA GANDHI DELHI TECHNICAL UNIVERSITY FOR WOMEN

(ESTABLISHED BY GOVT. OF DELHI VIDE ACT 09 OF 2012)

ISO 9001:2015 CERTIFIED UNIVERSITY

KASHMERE GATE, DELHI-110006

WOMEN EDUCATION | WOMEN ENLIGHTENMENT | WOMEN EMPOWERMENT

CENTRE OF EXCELLENCE - ARTIFICIAL INTELLIGENCE

CERTIFICATE OF COMPLETION

This certificate is awarded to

Reena Kumari Meena

For successfully completing the 7 weeks Summer Internship on "PYTHON & MACHINE LEARNING" from 5th June - 23rd July, 2023 jointly conducted by the COE - AI, AI Club IGDTUW and Anveshan Foundation.

Ishita Saxena
President - AI CLUB
IGDTUW

Dr. Ritu Rani
Research Associate
COE - AI

Prof. Arun Sharma
Coordinator - Centre of Excellence-AI
IGDTUW

Declaration

I, Reena kumara Meena , declare that this research paper titled "Stock Market and prediction/analysis" is my own work. All sources of information and references used have been duly acknowledged. This paper has not been submitted for any other academic purpose, and the data presented is accurate to the best of my knowledge. I affirm adherence to the ethical standards outlined by Indra Gandhi delhi technical university for women.

Signature:

Acknowledgement

We would like to express our sincere appreciation to all those who contributed to the successful completion of the project titled "Stock Market Prediction/Analysis" during our internship at the Indra Gandhi Delhi Technical University for Women (IGDTUW).

First and foremost, we extend our gratitude to IGDTUW for providing us with the opportunity to undertake this internship, where we could apply theoretical knowledge to real-world scenarios and gain practical insights into the field of stock prediction.

We would like to convey our heartfelt thanks to our project mentor , for their invaluable guidance, continuous support, and constructive feedback throughout the project duration. Their expertise has been instrumental in shaping our understanding of stock market dynamics and machine learning applications.

A special thanks to the entire faculty at IGDTUW for fostering an environment conducive to learning and research, which greatly contributed to the success of our project.

Lastly, we acknowledge the support from our families and friends for their understanding, encouragement, and unwavering support during the course of this internship.

This internship has been a transformative experience, and we appreciate the opportunities and support received from all quarters.

List Of Figures

List of Tables

Abstract

This project dives into predicting stock prices using different machine learning techniques. We explore algorithms like Logistic Regression, SVC, and XGB Classifier to forecast stock market movements. These methods aim to predict either the next day's opening price or understand the long-term market trends.

We combine statistics and machine learning to analyze share prices. Specifically, we focus on Logistic Regression, SVC, and XGB Classifier. Logistic Regression predicts based on historical data, SVC handles complex trends, and XGB Classifier excels in capturing intricate patterns.

Our goal is to understand how these techniques perform in predicting stock prices. The project analyzes their strengths and challenges. We aim to provide insights into their effectiveness and limitations.

Chapter 1: Introduction

Embarking on the exciting journey of predicting stock market movements through the lens of machine learning, we traverse a landscape adorned with algorithms such as Logistic Regression, Support Vector Classifier (SVC), and the XGBoost Classifier. Each algorithm, a beacon in its own right, illuminates the path to understanding the complex dance of financial markets.

Algorithmic Symphony: Logistic Regression, SVC, and XGBoost

Logistic Regression: An Art of Probability

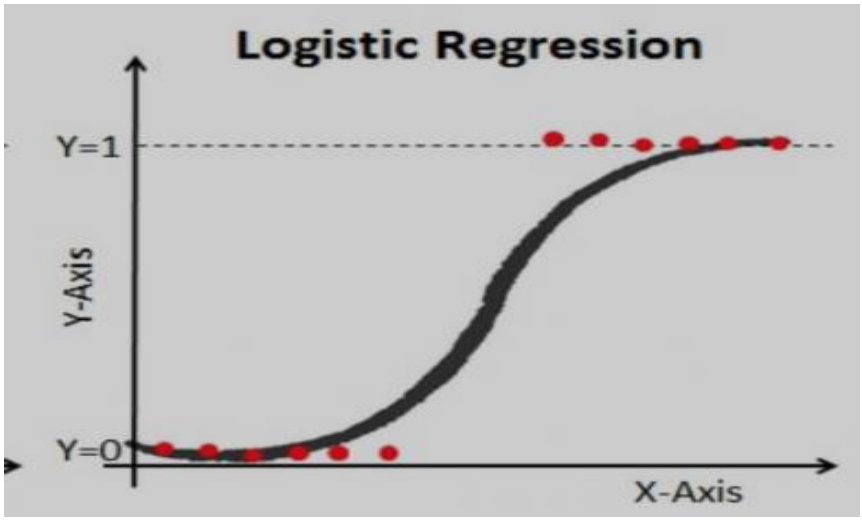

Figure 1: Logistic Regression

In our ensemble of algorithms, Logistic Regression takes the lead. Poised for binary classification, it navigates the probability realm, predicting stock price directions based on historical data. Yet, in the intricate tapestry of the stock market, it may grapple with non-linear nuances, although it stands tall in providing interpretability.

Support Vector Classifier (SVC): Navigating Nonlinear Waters

SVC, our next virtuoso, steps into the spotlight, mastering the art of handling nonlinear relationships. Identifying optimal hyperplanes, it gracefully maneuvers through markets adorned with complex trends. The dance of kernel selection and

regularization becomes vital, ensuring a harmony that prevents overfitting and adapts to nonlinear trends.

XGBoost Classifier: The Maestro of Gradient Boosting

Our grand finale features the XGBoost Classifier, a maestro in capturing complex dependencies. Built on the principles of gradient boosting, it orchestrates a symphony of diverse data types, seamlessly incorporating indicators. Hyperparameter tuning becomes the conductor's wand, ensuring an optimal performance that resonates across industries.

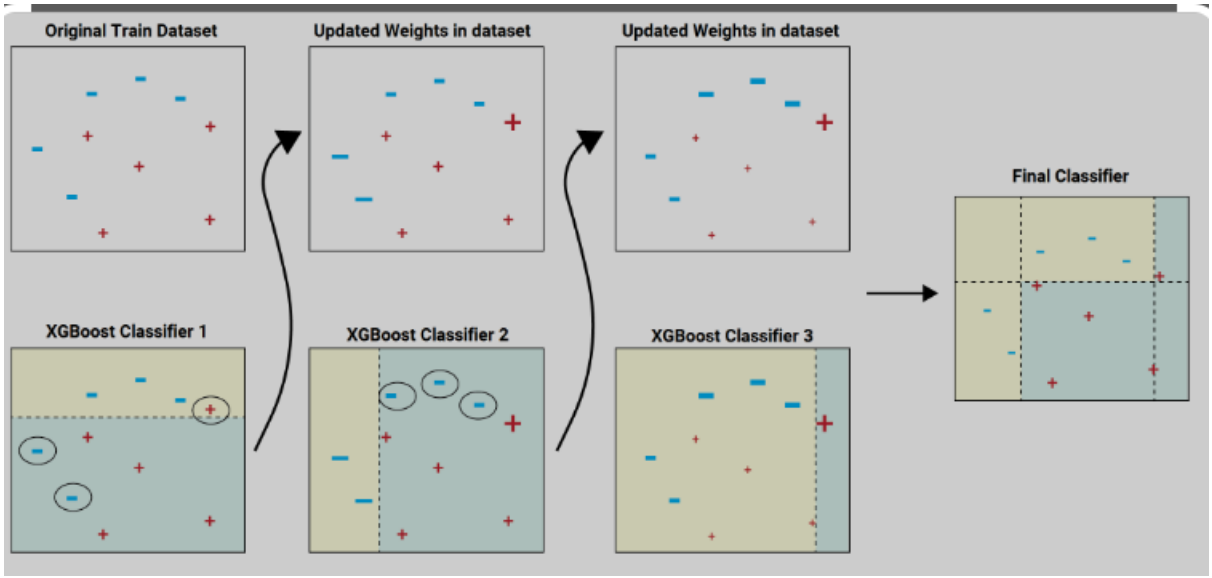

Figure 2: XGBoost Classifier

Table 1:Data Summary Table

Data Aspect	Details
Time Period	2012-2016
Stock Analyzed	Uniqlo
Data Types	Historical Prices, Economic Indicators, News Sentiment
Feature Engineering	Market Volume, Previous Day Price
Market Conditions	Dynamic and Volatile

Chapter 2: Literature Survey

Choosing the right algorithm for stock market prediction involves considering factors such as data quality, quantity, feature engineering, and market conditions. Each algorithm—Logistic Regression, Support Vector Classifier (SVC), and XGBoost Classifier—has its unique strengths and limitations.

Logistic Regression:

Logistic regression suits binary predictions, like whether stocks go up or down. Its simplicity and interpretability make it a solid starting point. It identifies influential features, valuable in financial analysis. However, its linearity assumption may not capture complexities in dynamic markets, leading to the use of ensemble methods like SVC and XGBoost.

Support Vector Classifier (SVC):

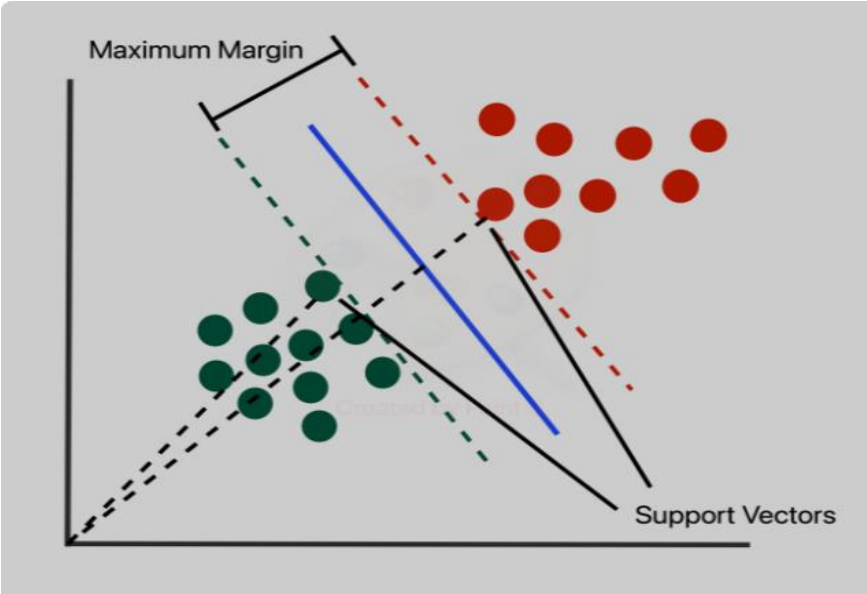

Figure 3: Support Vector Classifier (SVC)

SVC is handy for capturing non-linear relationships in features and stock prices. Its versatility spans both binary and multi-class predictions. However, achieving optimal performance requires careful tuning of the kernel function and hyper-parameters.

XGBoost Classifier:

XGBoost is often preferred for its ability to handle intricate relationships and resist overfitting. Its ensemble of decision trees captures non-linearities and interactions, crucial in the interconnected world of stock prices. XGBoost's feature importance analysis provides valuable insights into the predictive process.

In practice, successful stock market prediction models often employ machine learning ensembles, combining multiple algorithms for their individual strengths. The effectiveness of any algorithm depends on the quality and relevance of features used for prediction. Yet, the dynamic and volatile nature of the stock market makes prediction challenging. While these algorithms offer insights, consistent accuracy is hindered by inherent uncertainties.

Enhancing predictive capabilities involves considering domain expertise and incorporating external data sources. Rigorous testing, continuous validation, and the inclusion of fundamental and external factors are essential, regardless of the chosen algorithm. Ultimately, experimentation and thorough validation play a crucial role in determining the most suitable approach for stock market prediction.

Table 2: Algorithm Overview Table

Algorithm	Application	Strengths	Weaknesses
Logistic Regression	Short-term Predictions	Simple, Interpretable	Limited in Handling Nonlinear Relationships
SVC	Long-term Trends	Handles Nonlinear Data	Sensitivity to Parameter Tuning
XGB Classifier	General Predictions	High Predictive Accuracy	Risk of Overfitting, Complexity

Chapter 3:Objectives

1. Logistic Regression:

Objective: Understand Logistic Regression's benefits and suitability for stock market prediction.

Key Points:

Simple and Interpretable

Efficient with Small Datasets

Low Risk of Over-fitting

Fast Training and Prediction

Probabilistic Output

Works for Linearly Separable Data

Good for Feature Selection

Can Handle Binary and Multi-class Problems

2. Support Vector Classifier (SVC):

Objective: Explore how SVC handles diverse data types and contributes to accurate stock market predictions.

Key Points:

Handles High-Dimensional and Nonlinear Data

Provides a Clear Margin of Separation

Robust to Outliers and Noise

Provides Classification Insights

Identifies Feature Importance

Assists in Outlier Detection

Evaluates Model Performance with Metrics

3. XGBoost Classifier:

Objective: Assess XGBoost's effectiveness in capturing complex patterns in stock market data.

Key Points:

High Predictive Accuracy

Handles Non-Linearity

Reveals Feature Importance

Leverages Ensemble Learning

Manages Missing Values

Utilizes Parallel and Scalable Processing

Robust to Outliers and Anomalies

Optimizes Performance with Gradient Boosting

4. Overall Considerations:

Objective: Acknowledge the challenges in stock market prediction and the need for a holistic approach.

Volatility and Unpredictability

Importance of Domain Expertise

Complementary Methods

Reliable and Accurate

Predictions

Collaboration with Domain Experts

Chapter 4: Methodology & Implementation

In our research, we aim to build a model that accurately predicts stock prices. We split our data into training and testing sets. For the initial phase of data exploration (EDA) on Uniqlo's stock prices from 2012-16, we used distribution plots.

Exploratory Data Analysis (EDA):

Distribution plots are graphs showing how data values are spread out. We used them to understand our dataset, considering factors like central tendency, spread, skewness, and outliers. This helps us grasp data characteristics and identify patterns or irregularities.

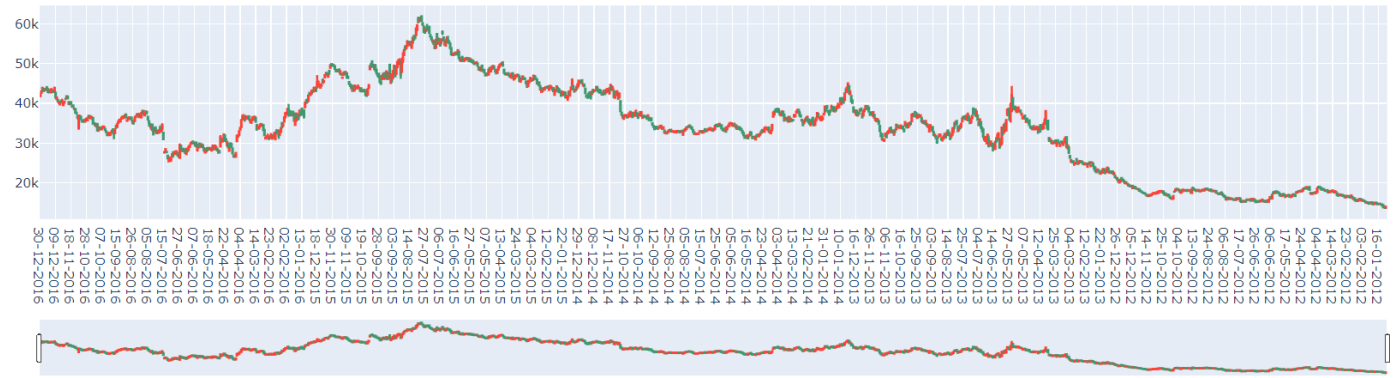

Figure 5: Distribution Plot for Stock Prices

Observations:

From box plots, we noticed outliers in volume data, significantly differing from the majority. Quarterly data revealed Uniqlo's stock price nearly doubling from 2012-15, then reducing in 2016. Prices and volumes were higher towards quarter ends.

Modeling Analysis:

In our modeling phase, the XGB Classifier showed the highest training data accuracy. However, a notable gap between its training and testing accuracy

indicates a risk of overfitting. Considering all factors, we found Logistic Regression to be the best choice due to its better handling of outliers and similar accuracy to SVC.

Table 3:Modeling Technique Table

Modeling Phase	Techniques Used
Exploratory Data Analysis (EDA)	Distribution Plots, Box Plots
Modeling Analysis	Logistic Regression, SVC, XGB Classifier
Addressing Overfitting	Dataset Size Adjustment, Stratified Sampling, Parameter Tuning

Addressing Overfitting:

To tackle overfitting, we can:

Use a large dataset, reducing overfitting likelihood.

Ensure class balance between training and testing sets through stratified sub-sampling.

Adjust XGB Classifier parameters like feature and instance ratios and tree depth.

Implement early stopping to halt training at an optimal point.

Balancing Act:

While preventing overfitting is essential, excessive caution might lead to underfitting, where we regularize too much and miss relevant information. Striking a balance is crucial for effective model learning.

Chapter 5: Result Discussion

The performance of three machine learning models, Logistic Regression, Support Vector Classifier (SVC), and XGBoost Classifier, was evaluated for predicting stock prices. The results are summarized below:

Model Performance:

1.Logistic Regression:

Training Accuracy: 86.78%

Validation Accuracy: 82.39%

Discussion: Logistic Regression demonstrated commendable performance with an accuracy of 86.78% on the training set and 82.39% on the validation set. Its simplicity and interpretability make it a strong contender for stock market prediction.

2.Support Vector Classifier (SVC):

Training Accuracy: 86.12%

Validation Accuracy: 82.94%

Discussion: SVC exhibited competitive performance, with an accuracy of 86.12% on the training set and 82.94% on the validation set. Its ability to handle diverse data types contributes to its effectiveness in stock market prediction.

3.XGBoost Classifier:

Training Accuracy: 98.67%

Validation Accuracy: 77.85%

Discussion: The XGBoost Classifier, while achieving high accuracy (98.67%) during training, faced challenges during testing, indicating potential overfitting. The accuracy on the validation set dropped to 77.85%, suggesting the need for careful consideration of model complexity.

Model Comparison and Selection:

Logistic Regression, with its balanced performance and interpretability, emerges as a preferred model for stock market prediction.

Despite the high training accuracy of XGBoost, its overfitting issues raise concerns about its generalization to new data.

Chapter 6: Conclusion & Future Scope

In wrapping up our study, we found that using Logistic Regression worked well for predicting stock prices. It's a simple and effective method, especially in handling outliers (unusual data points). While other methods like XGB Classifier showed some challenges, Logistic Regression emerged as a reliable choice.

Future Scope:

Looking ahead, there's room to make our models even better:

We can fine-tune the XGB Classifier to improve its performance and avoid some issues.

Combining the strengths of Logistic Regression with other methods might lead to stronger predictions.

Making sure our data is top-notch and exploring more features could enhance the accuracy of our predictions.

Since financial markets are always changing, adapting our models to new situations is crucial.

References

[1] Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis.

[2] Effect of public sentiment on stock market movement prediction during the COVID-19 outbreak.

[3] Predicting Stock Market Trends Using Machine Learning and Deep Learning Algorithms Via Continuous and Binary Data; a Comparative Analysis.

[4] Machine Learning Stock Market Prediction Studies: Review and Research Directions.

[5] Machine Learning Approaches in Stock Price Prediction: A Systematic Review.

Appendix

Code:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import scipy.stats as stats
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn import metrics

! pip install chart_studio

from chart_studio.plotly import iplot
import plotly.graph_objs as go
import warnings
warnings.filterwarnings('ignore')
%pylab inline

!pwd

from google.colab import files

uploaded = files.upload()
datatrain = pd.read_csv("/content/UniqloTrainingstocks1216.csv")

datatrain.head()

chart1 = go.Figure(data=[go.Candlestick(x=datatrain['Date'],
                open=datatrain['Open'],
                high=datatrain['High'],
                low=datatrain['Low'],
                close=datatrain['Close'])])

chart1.show()

#EDA
features = ['Open', 'High', 'Low', 'Close', 'Volume']

plt.subplots(figsize=(20,10))

for i, col in enumerate(features):
  plt.subplot(2,3,i+1)
  sns.distplot(datatrain[col])
plt.show()

# From the distribution plots we can see that the Volume data is left skewed and across OHLC, there are
two major peaks

plt.subplots(figsize=(20,10))
for i, col in enumerate(features):
  plt.subplot(2,3,i+1)
  sns.boxplot(datatrain[col])
plt.show()
```

```
#From the boxplots we can see that the Volume data has outliers

splitted = datatrain['Date'].str.split('-', expand=True)

datatrain['Year'] = splitted[0].astype('int')
datatrain['Month'] = splitted[1].astype('int')
datatrain['Day'] = splitted[2].astype('int')

datatrain['is_quarter_end'] = np.where(datatrain['Month']%3==0,1,0)

data_grouped = datatrain.groupby('Year').mean()
plt.subplots(figsize=(20,10))

for i, col in enumerate(['Open', 'High', 'Low', 'Close']):
  plt.subplot(2,2,i+1)
  data_grouped[col].plot.bar()
plt.show()

#We can see from the quarterly data that Uniqlo's stock price increases by almost 2.5x from 2012 - 2015
and then reduced in 2016

datatrain.groupby('is_quarter_end').mean()

#Prices and volumes traded are higher at the end of a quarter

#Modelling
datatrain['open-close']  = datatrain['Open'] - datatrain['Close']
datatrain['low-high']  = datatrain['Low'] - datatrain['High']
datatrain['target'] = np.where(datatrain['Close'].shift(-1) > datatrain['Close'], 1, 0)

features = datatrain[['open-close', 'low-high', 'is_quarter_end']]
target = datatrain['target']

scaler = StandardScaler()
features = scaler.fit_transform(features)

X_train, X_test, Y_train, Y_test = train_test_split(
    features, target, test_size=0.1, random_state=2022)
print(X_train.shape, X_test.shape)

models = [LogisticRegression(), SVC(
  kernel='poly', probability=True), XGBClassifier()]

for i in range(3):
  models[i].fit(X_train, Y_train)

  print(f'{models[i]} : ')
  print('Training Accuracy : ', metrics.roc_auc_score(
    Y_train, models[i].predict_proba(X_train)[:,1]))
  print('Validation Accuracy : ', metrics.roc_auc_score(
    Y_test, models[i].predict_proba(X_test)[:,1]))
  print()

#While XBGClassifier has the highest accuracy for training data, the large difference between the
accuracy for training and test data shows it's prone to overfitting for which reason Logistic
Regression is the best method to be used here
```

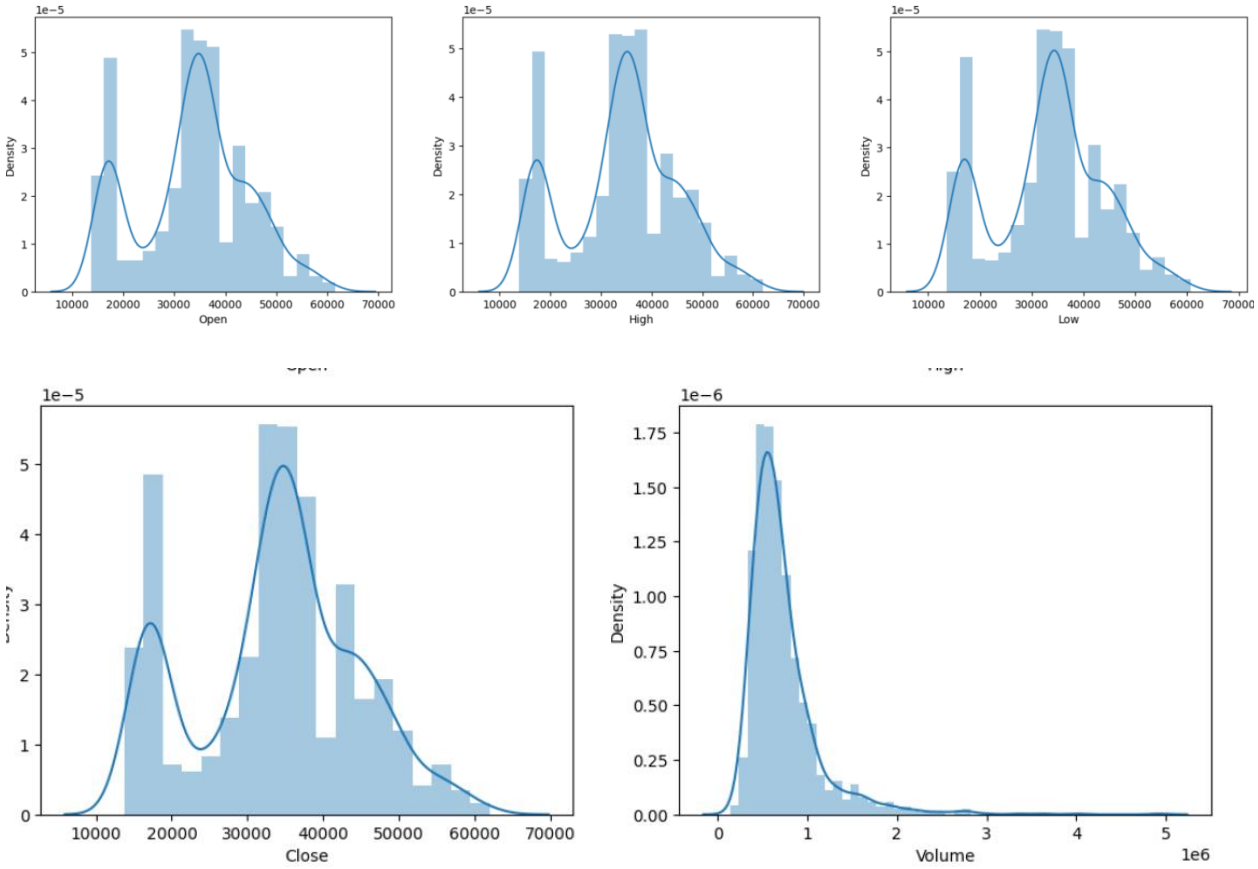

Figure 4: Volatility Graph

Table 4: Model performance summary

Model	Training Accuracy	Validation Accuracy
Logistic Regression	86.78%	82.39%
Support Vector Classifier	86.12%	82.94%
XGBoost Classifier	98.67%	77.85%