PrognoStock - A stock forecasting model

Aayush Pradhan, Manas Maskar, Naga Satya Dheeraj Anumala, Tanvi Shinde (Department Name)

KEYWORDS — Machine Learning, Forecasting, Prediction, Gradient Boost, Random Forest Regressor

I. DATA FETCH

The yahoo finance API has a wide range of data like Cryptocurrencies, regular currencies, stocks and bonds, fundamental and options data, and market analysis and news. We intented to use the official API provided by Yahoo, but seems like this was shutdown in 2017. The next option was to scrape the data from their website or to use pre exisitng APIs. We explored two popular APIS in python - RapidAPI, yahoo_finance and yfinance. We decide to go forward with yfinance API.
Data from Yahoo Finance API is sourced from various financial markets, including stock exchanges, commodities markets, and currency exchanges. Yahoo Finance, a subsidiary of Verizon Media, provides this data. The raw data from Yahoo Finance API is typically in JSON format. It contains a variety of information such as stock prices, market indices, company fundamentals, historical data, and news related to financial markets.

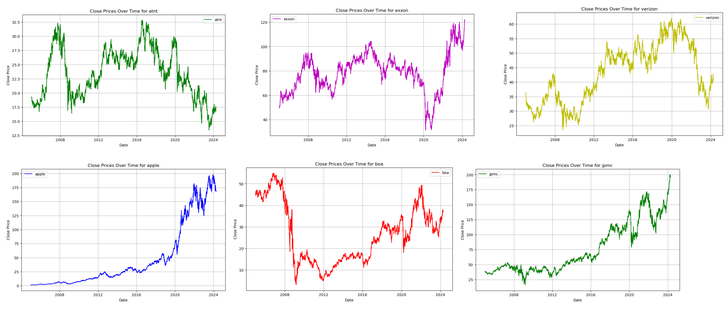
	company_name	Data	Open	High	Low	Close	Adj Close	Volume	MACD	Signal Line
0	apple	2005-01-03	1.156786	1.162679	1.117857	1.130179	0.966809	499992000	0.000000	0.000000
1	apple	2005-01-04	1.139107	1.159107	1.134464	1.147786	0.996503	1096853400	0.002220	0.000722
2	apple	2005-01-05	1.151071	1.165179	1.143750	1.159786	0.979101	680433600	0.005543	0.000296
3	apple	2005-01-06	1.154821	1.159107	1.130893	1.152679	0.975658	705559200	0.008704	0.000434
4	apple	2005-01-07	1.160714	1.243393	1.156250	1.236607	1.046911	2227450400	0.003604	0.001871
...	...	...	...	...	...	...	...	...	...	...
66037	amt	2024-04-04	17.620001	17.879999	17.559999	17.570000	17.286842	31850300	0.137127	0.098125
66038	amt	2024-04-05	17.520000	17.540001	17.280001	17.420000	17.139259	37757200	0.126414	0.104185
66039	amt	2024-04-08	17.400000	17.459999	17.219999	17.250000	16.972000	27372200	0.106782	0.104705
66040	amt	2024-04-09	17.000000	17.080000	16.809999	16.900000	16.900000	32802500	0.082873	0.100330
66041	amt	2024-04-10	16.770000	16.790001	16.610000	16.730000	16.730000	39046800	0.049835	0.090198

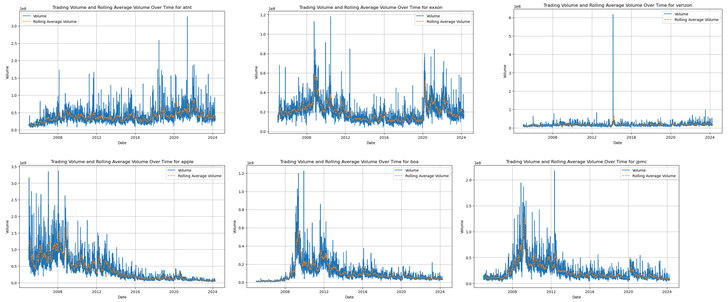
	Mean	Median	Standard Deviation
Open	6.372299e+01	4.631250e+01	5.654971e+01
High	6.434341e+01	4.676000e+01	5.714018e+01
Low	6.309623e+01	4.589347e+01	5.597350e+01
Close	6.373471e+01	4.631000e+01	5.657969e+01
Adj Close	5.281323e+01	3.471290e+01	5.672724e+01
Volume	6.353079e+07	2.262917e+07	1.548333e+08
MACD	-1.240202e+00	6.566909e-02	7.471101e+00
Signal Line	-1.239155e+00	6.460009e-02	7.448345e+00

Combined Clean Dataframe.

The data covers a wide range of financial instruments and is updated frequently, often in real-time or with minimal delay. This includes data on stocks, bonds, commodities, currencies, and indices. Updates can occur multiple times per minute during trading hours and less frequently during non-trading hours.

II. VISUALISATIONS





III. TIME SERIES ANALYSIS

In Time Series Analysis, we first group by the data by sector and then we add df normalization to remove outliers to avoid problems such as oversampling and undersampling. Make 2 layers of LSTM of 50 each. Trained the model on 5 epochs and in batch size of 32. After getting the results we find “residuals”.

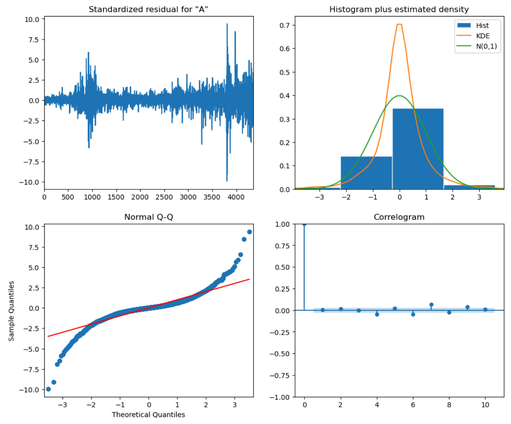


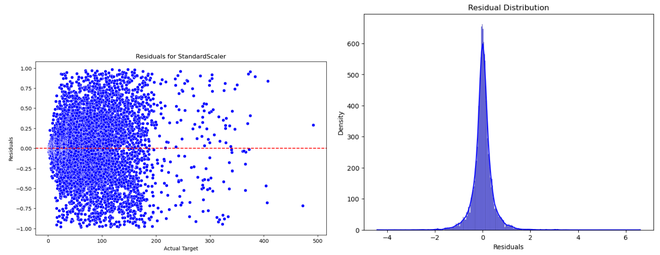
Figure 1: PLotting Diagnostic Graphs

.

.

IV. MACHINE LEARNING MODELS

In ML Model implementation we mainly select Gradient Boosting Predictor as it performs quite well as compared to every other model we implemented. The hyperparameters we selected are –> a. Loss, b. Learning_Rate, c. n_estimators, d. max_depth, e. random_State. Calculate MSE and residuals to verify the accuracy and precision.



V. CONCLUSION

In this project, we applied Time Series Analysis (TSA) and Gradient Boosting, two advanced machine learning techniques, to predict stock prices. Time Series Analysis helped us understand and model the sequential patterns inherent in historical stock price data, providing insights into trends, seasonal variations, and cyclic behavior. On the other hand, Gradient Boosting leveraged an ensemble of decision trees to capture complex nonlinear relationships in the data, offering robust predictions even in the face of volatile market conditions.
The combination of these methods allowed for a nuanced approach to stock price forecasting. TSA laid the groundwork by elucidating the temporal dynamics, while Gradient Boosting built on this foundation to enhance prediction accuracy through powerful modeling of data intricacies. The results demonstrated that integrating these methodologies can significantly improve predictive performance compared to using traditional linear models or each technique in isolation. This project underscores the potential of machine learning in financial analytics, particularly in harnessing diverse algorithms to forecast market movements more reliably and effectively.