Data Cowboys

Stock Prediction

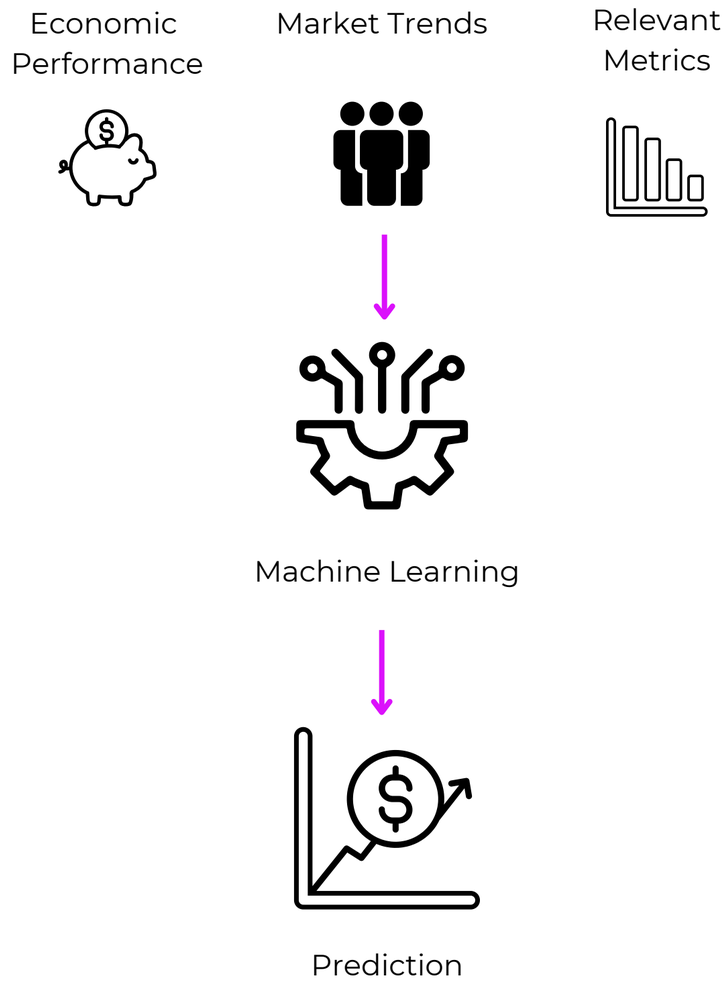
Task 4 of Data Intensive Computing 2023 course

Introduction

The stock market is a **dynamic** realm, influenced by a myriad of factors,

Big Data ML analytics are a powerful tool to help decision-making process

Accurate predictions are pivotal for informed investment decisions



Challenges

Influences

Stock prices are influenced by a vast array of interconnected factors over long periods

Complexity

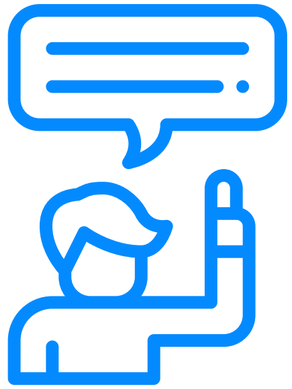
Relationships between predictors and target are complex and continuously evolving

Toolkit

Subtle patterns are not evident with basic tools, and extensive data information is required.

Objectives

FORECASTING
“Close” Price of Stocks



EXPLAINING
Model’s decision-making process

Project Tools and Data

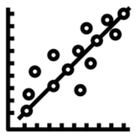
Distributed processing, exploration and modeling

Data Lake Scalability

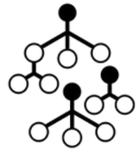
Stock and macroeconomics data

Our Solutions

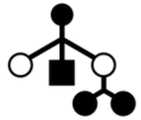
Machine learning (ML) and deep (DL) techniques were leveraged to effectively capture patterns and produce coherent stock movement forecasts



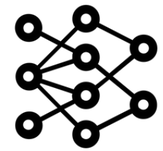
Linear Regression



Random Forests

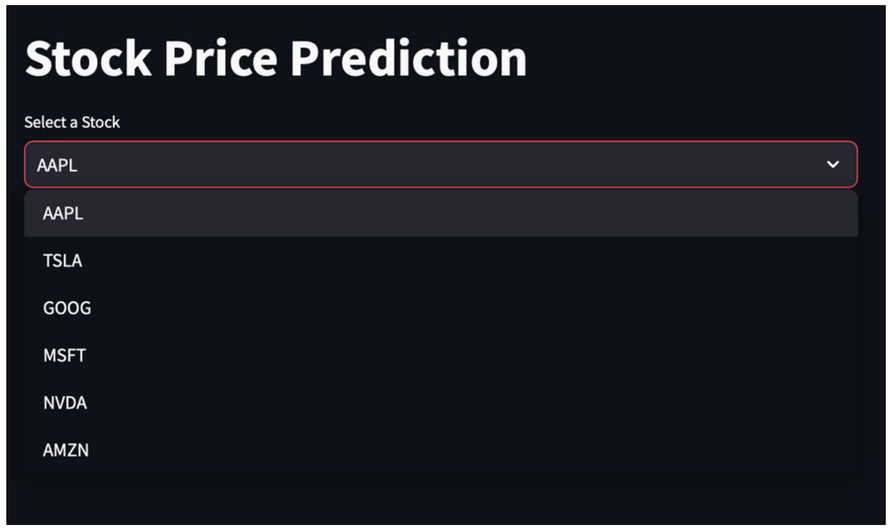


Gradient Boosted Trees

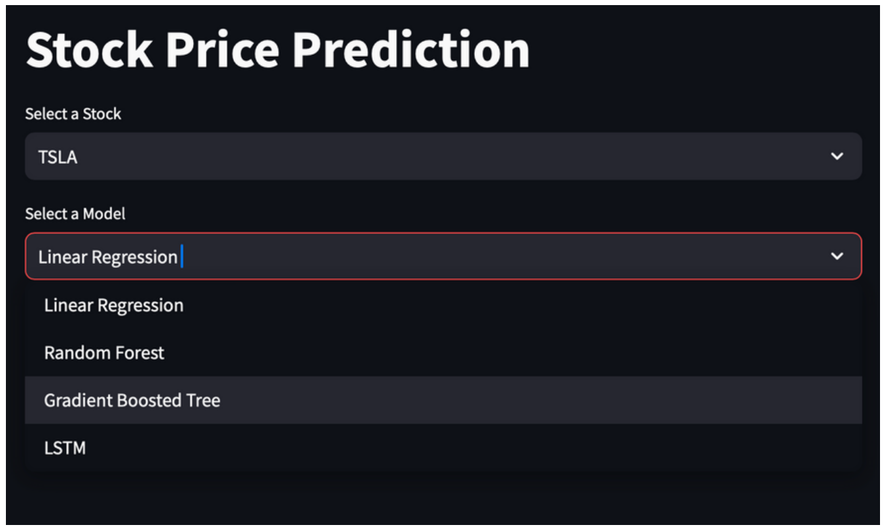


LSTM

Stock & Model Selection



The first step consists of selecting the stock to predict. **Different data options** are available.



Then, the **model** for predictions can be chosen

TSLA STOCK AS BASELINE

Data Collection & Preprocessing

Data

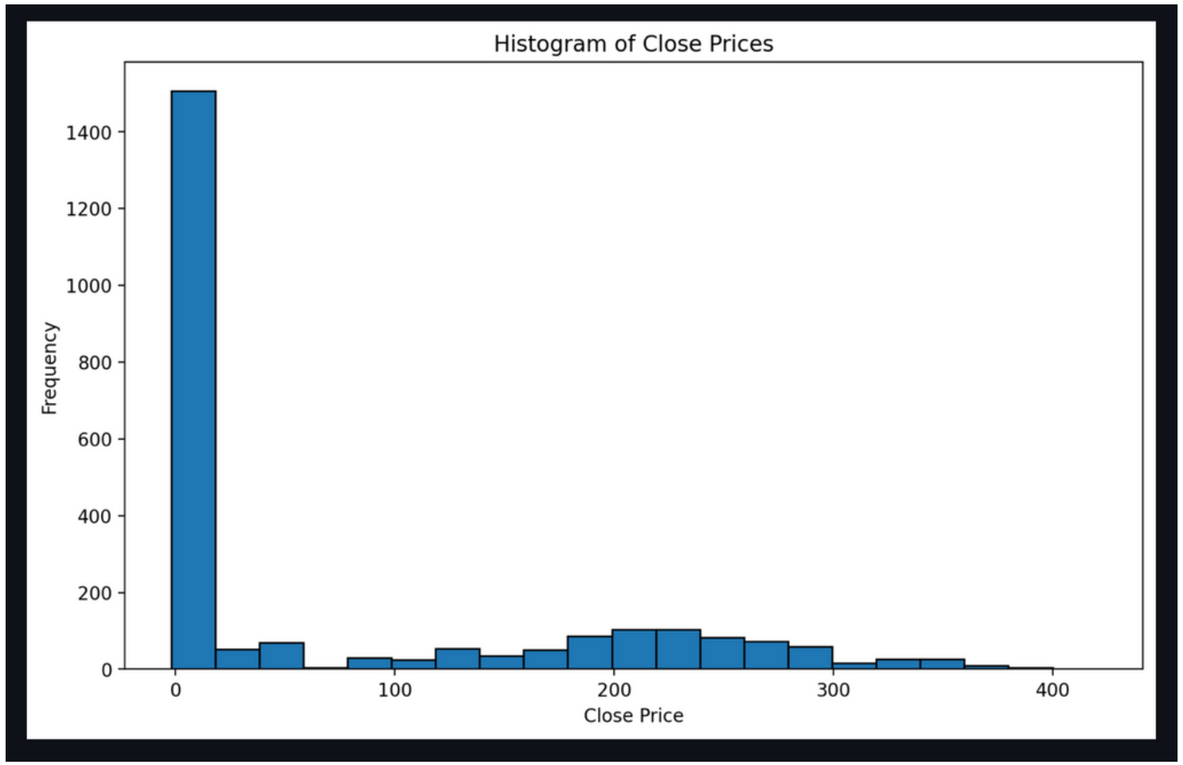
Data is taken from stock open, close, high, lows, and FRED indicators

Preprocessing

Includes data cleaning, merging, NULL filling, train split, scaling and vectorization (LSTM)

Data Exploration

The generated histograms visualize the distribution of closing prices to identify **dominant price ranges and outliers**



Performing exploration of Close Prices of TSLA stock...

Minimum Close

8.033332824707031

Maximum Close

409.9700012207031

Average Close

88.24294112473356

Standard Deviation of Close

106.36414651513999

ML Predictions

ML models are **still limited** in fully extracting the temporal patterns inherent in stock market fluctuations. This provided motivation to also explore models with stronger capabilities for sequential data like LSTM networks.

Linear Regression

RMSE: 3.812

Random Forest

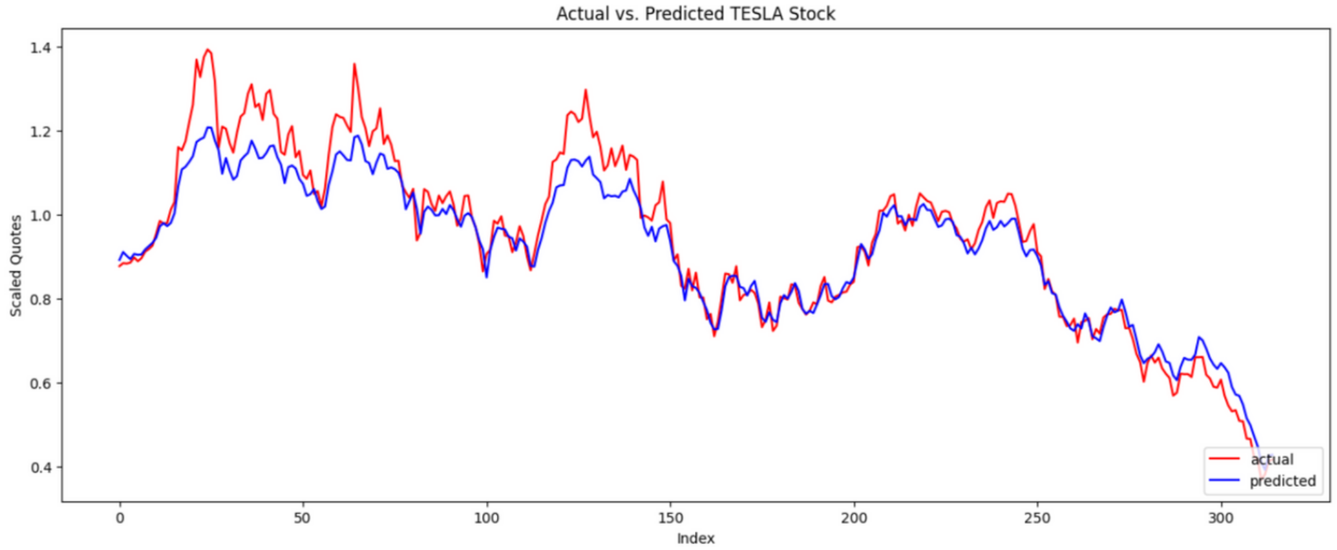
RMSE: 108.762

GB Tree

RMSE: 50.452

DL Predictions

LSTM network capable of learning from sequence data through internal memory gates. This allows it to capture **complex temporal patterns** between past stock/indicator values and future prices.



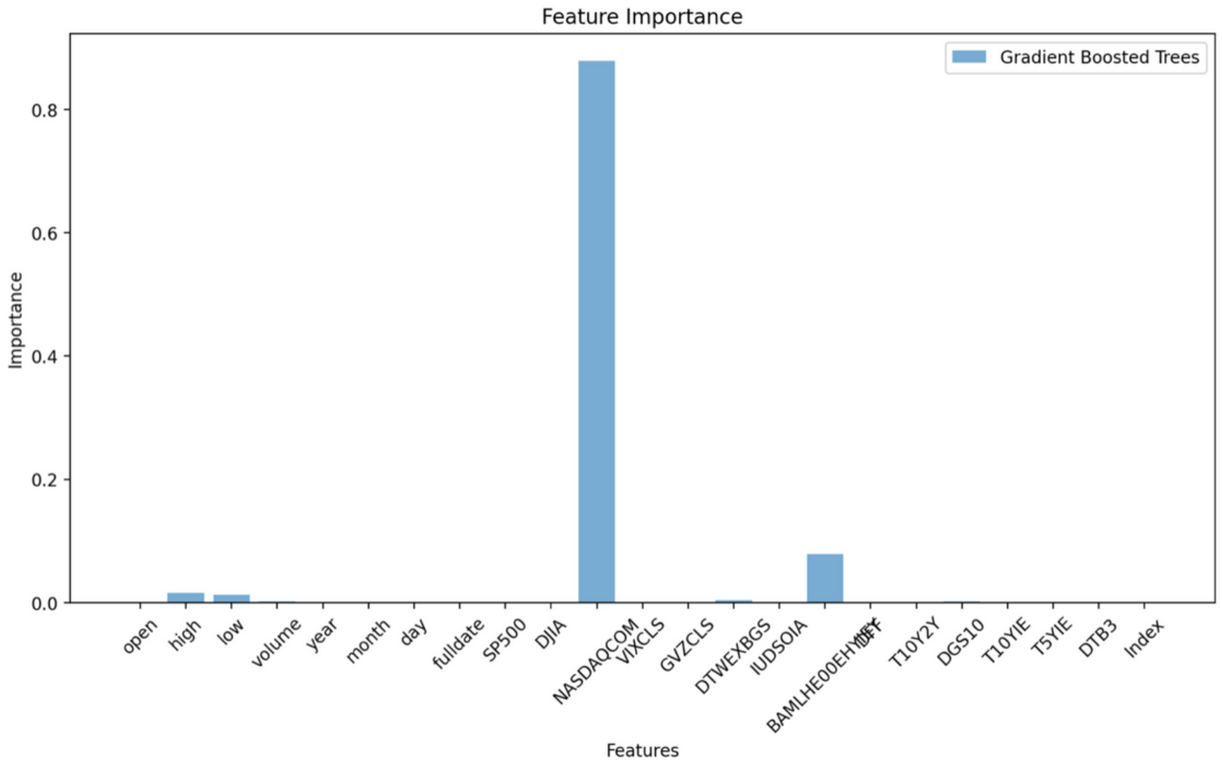
LSTM

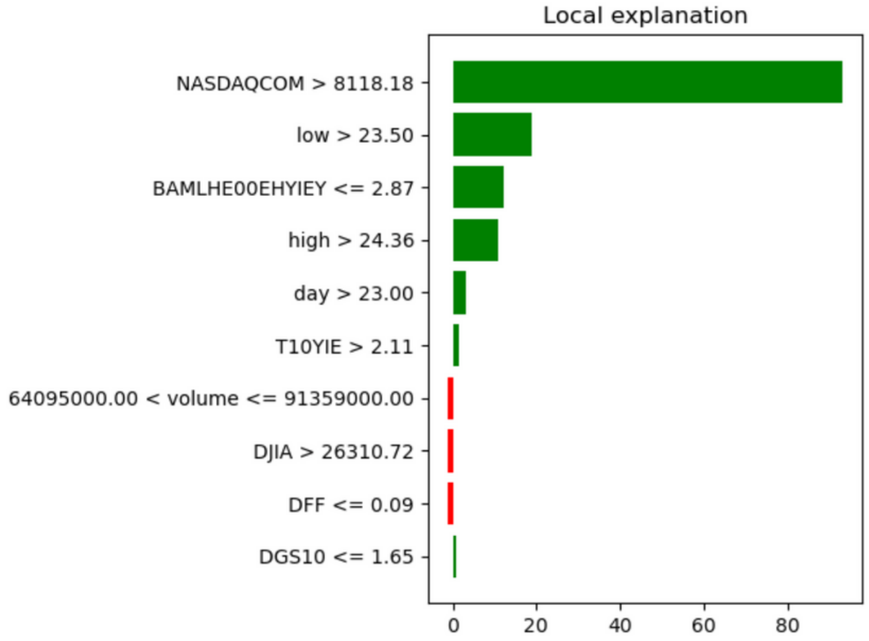
RMSE: 0.057

BEST PERFORMANCE ON EXAMPLE DATASET

Model Explanation

An example of analysis of **which factors most influenced** gradient-boosted tree decisions revealed "NASDAQCOM" index as the top signal, hence the possible performance impacts.





LIME: Revealed feature weights to better understand model reasoning.

Conclusions

This data-driven approach shows promise as an adaptive solution to the complex problem of distributed stock forecasting.

LSTM outperformed all algorithms, demonstrating value of recurrent networks for temporal data.

Feature analysis and LIME gave a **deeper understanding** beyond numbers.

The methodology showed **consistent performances** over different stock datasets.

Our Team

Matteo Pancini

Beatrice Insalata

Samuele Peri


38.962
18.158
7.58
78.561
43.997
25.886
7.352
14.229
10.552
51.298
2.345
22.743
99.537
0.34
98.773
7.58
7.58
8.668
7.352
8.668
43.997

Thank you