STOCK PRICE PREDICTION USING MACHINE LEARNING

Submitted By [Group 2]

Chris Lazarus - 673773993
Deepak Singhal - 672190946
Sanjay Madesha - 662505955

MOTIVATION

- The rapid spread of the unprecedented COVID-19 pandemic put the world in jeopardy and changed the global outlook unexpectedly.
- The US Stock Market saw its sharpest dip since the 2008 Great Depression. At the same time, the elections and fiscal rescue package by the government led to premature spikes in the stock market.
- The GameStop story that unfolded highlights the uncertainty that existed in the market during this time. While some people made profits others incurred huge losses.
- Our project aims at finding a solution to this problem by exploring the use of various machine learning techniques to predict stock prices.

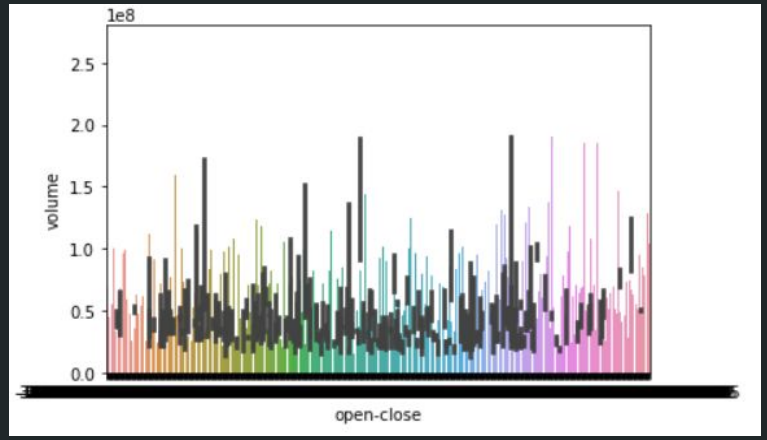
INTRODUCTION

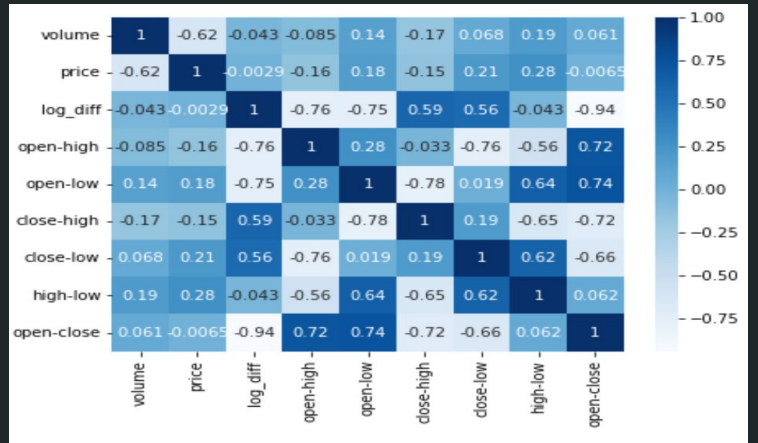
- **Dataset:**
  - Stock prices of stocks listed in the S&P 500 list for the past 5 years.
  - Features include daily high, low, opening, and closing prices listed along with their volumes.
- **Models used for Prediction:**
  - Linear Regression (Baseline Model)
  - K-Nearest Neighbours(**KNN**)
  - Support Vector Regression (**SVR**)
  - Long Short Term Memory (**LSTM**)
- **Model Evaluation Metric**
  - Root Mean Square Error (**RMSE**)
  - R-Squared Value (**R^2**)

UIC

EXPLORATORY DATA ANALYSIS

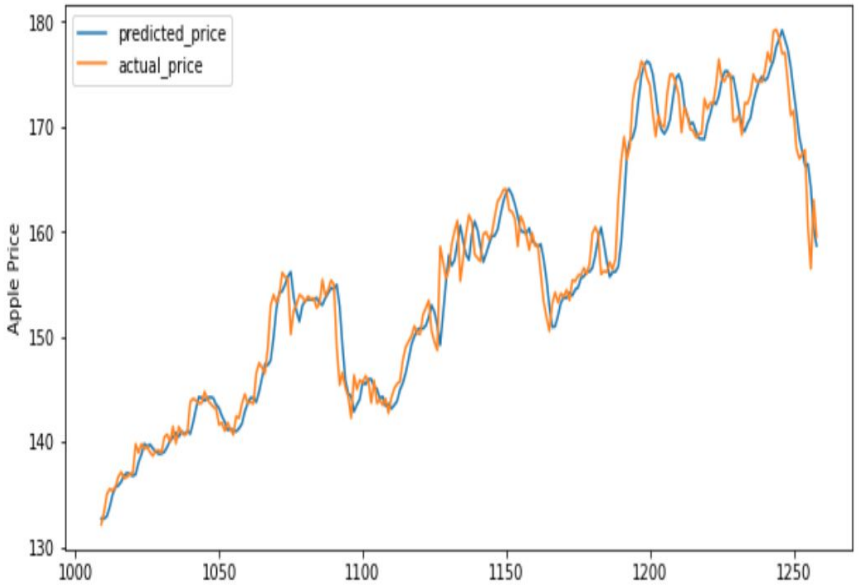
For EDA we started out by detecting null values for which we used the Missingno library. For observing the overall trend and looking at the outliers. We use histogram and line graph. We also calculated several measures like log difference and differences among parameters to observe the factors like symmetry, correlation, and influence of variables on the close price prediction.





Linear Regression

Baseline Model

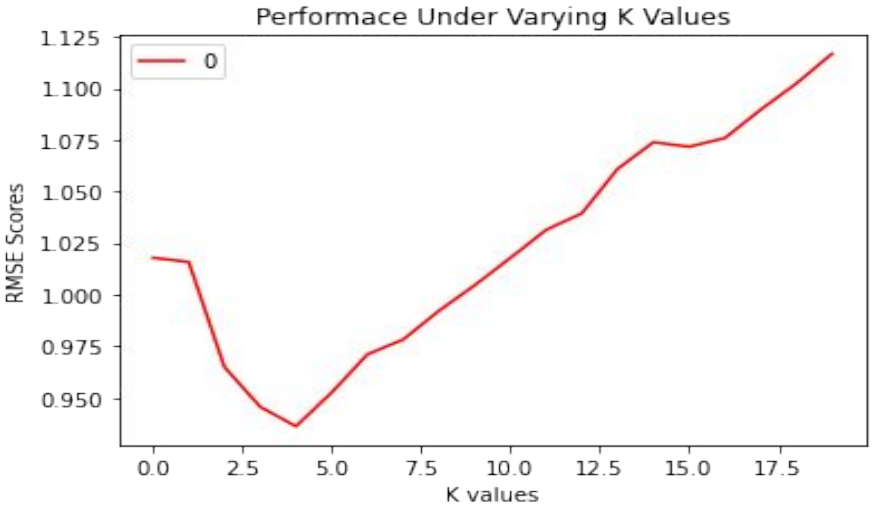


- Simple and easy to implement model to build for regression problems.
- Used Moving averages across the past 3 and 9 days.
- Evaluation Metrics Value
  - RMSE = 2.248
  - $R^2 = 0.96$

UIC

K-Nearest Neighbors (KNN)

First Main Model

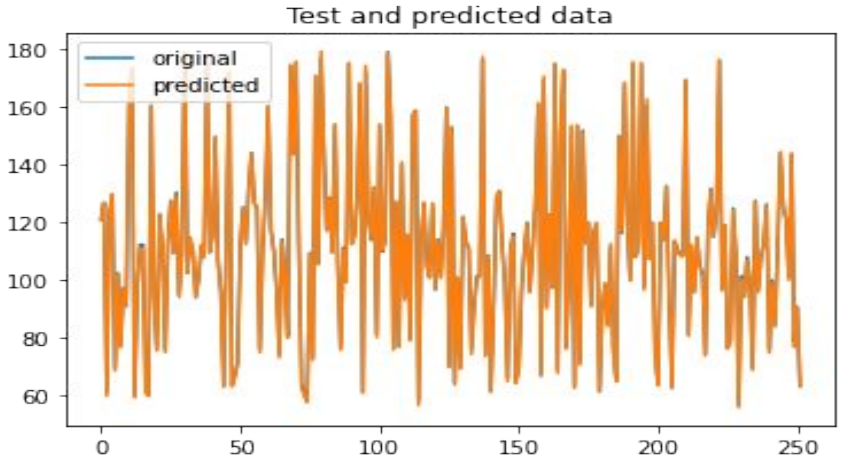


- Employes the lazy approach of using information from K nearest points,calculated using distance, to predict the output label.
- Used Elbow curve to determine the best value of K
  - K = 3
- Evaluation Metrics Value
  - RMSE = 0.93
  - R^2 = 0.99

UIC

Support Vector Regression (SVR)

Second Main Model

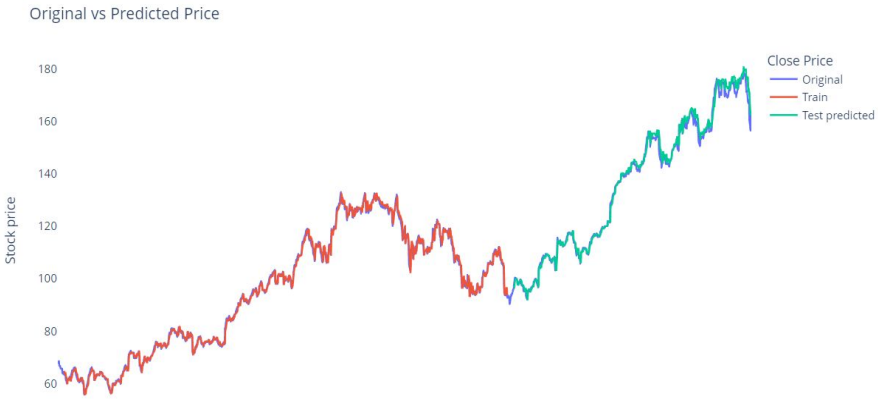


- Gives flexibility of defining how much error is acceptable in our model using C.
- Used GridSearchCV with cross validation to find the optimal Kernel and other hyperparameters
  - Kernel = Linear
  - C = 1000
  - Gamma = 0.001
- Evaluation Metrics Value
  - RMSE = 0.73
  - $R^2 = 0.99$

UIC

Long Term Short Memory (LSTM)

Third Main Model



- A Deep learning Model using Recurrent Neural Network (RNN) based architecture, capable of handling long-term dependencies.
- Works best on time series data.
- Model parameters
  - Hidden layers: 2
  - Epochs: 100
- Evaluation Metrics Value
  - RMSE = 1.78
  - $R^2 = 0.99$

Comparisons, Results & Conclusions:

MODEL	RMSE VALUE	R^2 VALUE
Linear Regression	2.248	0.96
KNN	0.93	0.99
SVR	0.73	0.99
LSTM	1.78	0.99

- Using **RMSE** and **R^2 Value**, we can conclude **SVR** as the most accurate model when compared to Linear, KNN and LSTM.
- Due to limited dataset LSTM does not perform as good as expected.
- R^2 parameter does not measure the predictive capacity of the obtained fit, it only measure proportion of variation explained by the posed predictor, where as RMSE is the square root of the average of squared errors.
- Both parameters are independent of each other. A combination of the two is used to select the best model.

Thank You!

Questions?