

CS-483
Group Project 1
Group: JET = NaN
Members:
Julian Keller
Tianyi Wang
Eric Gioe

Stock Price Prediction Summary

The historical timeframe chosen by the team was determined to be one year of data (253 points) for the training set, and one month (24 points) for the prediction dataset. The reason behind the single year for the training dataset was to eliminate the seasonality of the stock market. Every year there is a similar repeatable pattern from November to April, with the conclusion of the fourth quarter of the fiscal year and the beginning of the new fiscal year. It was our goal to only look at the latest year of data, so that we eliminate this phenomenon. One month of data was chosen for the prediction dataset in order to provide a sufficient amount of data for the model to perform well. It is difficult for regression models to perform complex data prediction without enough data.

It was decided that each of the group members was to complete a different model to predict stock data. The goal of this was to try three different methods to determine the most effective in predicting the stocks. The three models that were chosen were Long Short-Term Memory (LSTM), K-Nearest Neighbors (KNN) with Linear Regression, and Autoregressive Integrated Moving Average (ARIMA). The methods of predicting the data were compared using an error measurement for the same feature of the same stock that was chosen. The method with the least error was chosen for analyzing all of the data. The Long Short-Term Memory was the model that had the least amount of error in the data prediction.

Long Shot Term Memory Model Logic

LSTM uses a recurrent neural network architecture.

1. Load the data into a Pandas Dataframe and select a column to use for training the model.
2. Use Scikit-Learn's MinMaxScalar to scale the dataset between 0 and 1. This normalizes the data and maximizes performance.
3. The data must be transformed into the correct shape so it can be consumed by the LSTM model for training and predictions.
4. Add layers to the LSTM model. The layers includes the number of neurons/nodes to use in the layer. There is also a dropout layer which is used to prevent over-fitting our results.
5. The model is compiled and uses a mean squared error as a loss function.
6. The model is trained with the data, which can take some time.
7. Now the model is ready to use. We pass it some best guess data to use as a prediction and output a plot. This data is loaded and reformatted just like the training data is. The plot contains the actual stock price in grey, the test data prediction in orange, and the future forecasted prediction in purple.

Stocks Chosen:

Adidas, ADP, Bitcoin, Costco, FireEye, GoPro, Honeywell, Medtronic, S&P 500, Tesla

Team Responsibilities:

Each member of the team chose three stocks and downloaded the raw data. One stock was chosen as a team.

Julian - Long Short-Term Memory, Project Submission

Tianyi - K-Nearest Neighbors with Linear Regression, Setup Github Repository

Eric - Autoregressive Integrated Moving Average, Project Summary