Ryan Senoune

260989415

Deliverable 1: Data Selection Proposal

Project Idea:

As my final project, I will attempt to predict the price of a stock market index fund. My model will use past data of the S&P500 to attempt to predict future prices.

Dataset

Kaggle dataset: https://www.kaggle.com/camnugent/sandp500

This dataset has 5 years historical data on the 500 stocks included in the S&P500 Index fund. The information included are opening price, highest price, lowest price, closing price and volume for each day. If this data is not enough to make predictions, I may try to correlate the stock prices with additional data from other datasets. I could integrate other technical indicators/information about the stock such as moving average or volatility.

Methodology

Data Preprocessing

I don't think there should be much data preprocessing. I will not need the column for the date, instead I will use the row index to list prices from day 1 to day 1825. All the other columns are useful to make prediction about the price.

Machine Learning Model

With the data, I would like to estimate the future price of the S&P500 index fund. The most straightforward model seems like linear regression. However, doing research on past projects, it seems like LSTM would be the best model for such predictions. This model counters the problem of long-term dependencies.

Evaluation Metric

I will calculate the mean squared error to see how much the predict price differs from the actual price.

Final Conceptualization

I would like to build a webapp where the predicted graph of prices is displayed along with the accuracy.