Course Project Midterm Report

Clayton Tucker
COSC 325
Ctrl Alt Delete
Knoxville, US
ctucke24@vols.utk.edu

Todd Van Meter
COSC 325
Ctrl Alt Delete
Knoxville, US
tvanmete@vols.utk.edu

Danyil Chuprynov
COSC 325
Ctrl Alt Delete
Knoxvile, US
dchupryn@vols.utk.edu

Ian Henson
COSC 325
Ctrl Alt Delete
Knoxvile, US
tkd995@vols.utk.edu

***Abstract*—The purpose of this article is to find a way to accurately predict future stock prices. This can be done by developing a machine learning program that trains itself on previous stock prices. There exist many different learning models to predict the future of stocks, though many of them are not considered efficient enough. The aim of this project is to take an existing learning model and further improve it using learning techniques such as ARIMA, Linear Regression, or Random forests. For this article, we will be using an ARIMA with Long Short Term Based Memory. We hope to improve it by shortening the data cut-off or perhaps combining multiple learning models.**

***Index Terms*—ARIMA, Linear Regression, Random forest, Stock Price Prediction**

I. Introduction

The prediction of stock prices has long been a subject of interest for investors, financial analysts, and researchers. Accurate prediction of stock prices can provide valuable information for decision-making in trading and investment strategies. However, stock price prediction is inherently complex due to the volatile and dynamic nature of financial markets. This project focuses on time series analysis to predict stock prices, specifically using historical stock data from Berkshire Hathaway Inc.

A. Problem Overview

Stock prices are influenced by a multitude of factors, including market trends, economic indicators, investor sentiment, and company performance. Traditional methods for predicting stock prices, such as fundamental and technical analysis, have limitations, as they rely on historical patterns and market assumptions. With advances in machine learning and deep learning, data-driven models have emerged as powerful tools for time series forecasting.

Subject to change.

In this project, our objective is to predict stock price movements using historical data consisting of features such as open price, close price, high and low prices, and trading volume. By analyzing these patterns over time, we seek to develop a model that can identify trends and potential future stock price fluctuations for Berkshire Hathaway. The unpredictability of the stock market poses challenges, but through proper analysis and appropriate modeling techniques, we can aim to improve prediction accuracy and offer valuable insight.

B. Importance of Solution

Accurate stock price prediction has significant implications for both individual and institutional investors. A reliable forecasting model can help investors make informed decisions, optimize portfolio management, and reduce financial risks. In addition, stock prediction models can contribute to the broader financial industry by enhancing risk assessment strategies and improving the trading algorithms used by hedge funds and financial institutions.

Berkshire Hathaway is a major player in the global financial market with a diverse portfolio of subsidiaries in multiple industries. Understanding and predicting the stock price movements of such a company can provide insight into broader economic trends and market behavior. Given Berkshire Hathaway's substantial market capitalization and its influence on investment decisions around the world, an effective predictive model could have far-reaching applications. Investors, traders, and financial analysts could use such a model to make strategic decisions that align with market movements and economic conditions.

Moreover, stock price prediction is not only about maximizing financial gains, but also about mitigating losses. Market volatility can lead to substantial financial risks, and an improved prediction model could help investors develop risk management strategies to protect their investments. By reducing uncertainty, investors can make data-driven decisions rather than relying on speculation or intuition.

C. Project Goal

The primary objective of this project is to develop a stock price prediction model using historical stock data from Berkshire Hathaway. Our goal is to explore various time series forecasting techniques, including statistical models and machine learning approaches, to determine the most effective method for predicting stock price trends. The analysis will focus on leveraging features such as open, close, high, low prices, and trading volume to train the model.

Through this research, we aim to achieve the following:

- Investigate the effectiveness of different time series prediction models for stock price forecasting.
- Evaluate the impact of various stock price features on prediction accuracy.

- Develop a predictive model that can provide insights into future stock price movements with a reasonable degree of accuracy.
- Provide an in-depth analysis of the results and discuss potential applications and limitations of the model.
- Assess the feasibility of using machine learning techniques in financial forecasting and their implications for the stock market.

By accomplishing these goals, we hope to contribute to the growing field of financial analytics and offer valuable insights for investors and analysts looking to enhance their decision-making processes in stock trading. This project will not only serve as an academic exploration of time series forecasting but also provide practical implications for real-world financial markets.

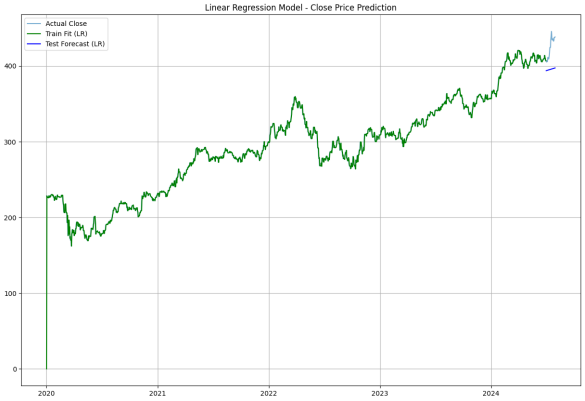
In summary, the study of stock price prediction using historical data is an essential aspect of financial analytics. With the increasing role of technology in market analysis, leveraging advanced machine learning and statistical techniques can enhance our ability to forecast market trends accurately. This research will help bridge the gap between traditional financial analysis and modern predictive modeling, offering valuable contributions to the field of financial forecasting and investment strategy.

II. Baseline Model

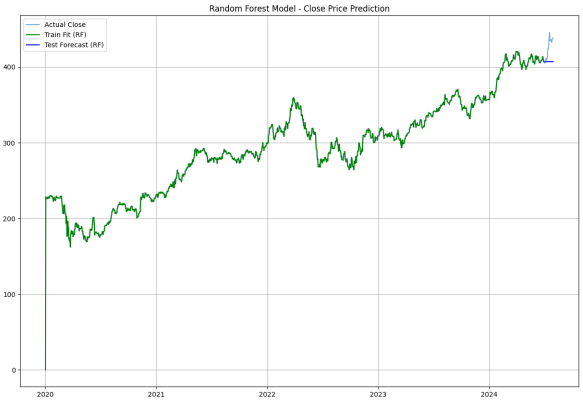
Once we have cleaned up our data, we now need to choose a baseline model to work on. Below are some of the models and the data they generated for us:

A. Existing Solutions:

- Linear Regression: Used for predicting continuous values given independent values. Such factors we can consider using for this model would be low, open, high, close, or volume. There is a problem using linear regression, as the data in the prediction of stocks is usually non-linear [1]. Based on our data, we found that linear regression would not be a suitable model for stock price predictions.



- Random Forest: Uses trees trained on different data randomness in feature selection. Capable of analyzing from multiple perspectives and is quickly at calculating results[2]. This model provided us with the best results compared to the others; however, they tend to overfit, resulting in accurate predictions.



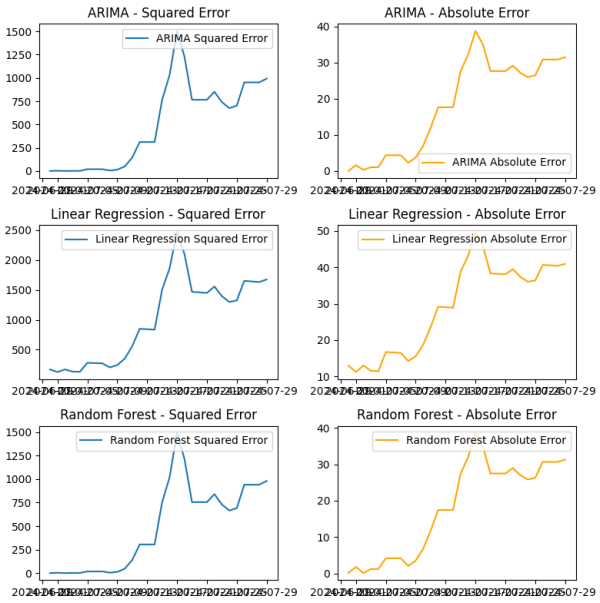
- ARIMA: Uses previous data to make predictions based on past events. Requires three different parameters to work. This model provided results similar to the random forest model and typically has the highest accuracy compared to other models[3]. The biggest downside of using this model is that the data must be stationary.

B. Selection of Baseline

We chose to use an ARIMA model as our baseline for this project as it provided the best data and is known for being accurate among other models.

C. Baseline Performance

Here is the data of the ARIMA, Linear Regression, and Random Forest baseline with squared and absolute error data.



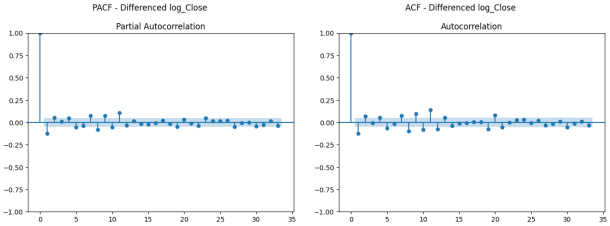
D. implementation

First, our data was made stationary and cleaned by removing empty columns. We then had to choose the parameters (p, d, q) for the ARIMA, which are the AR Term(p), obtained from PACF, I Term(d), obtained from finding the number of times data is differenced to become stationary, and MA Term(q), obtained from the ACF. Both the order of the AR and MA Terms are equal to the lags that can cross a significance limit. After the parameters have been chosen, we can then fit our model based on the parameters. Lastly, we evaluated our model as seen in Baseline Performance give different models.

III. Data Exploration

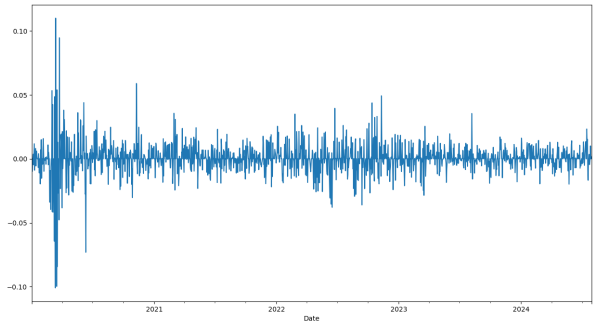
A. Cleaning and Normalization

To clean up our data, we only used prices that were after 2020. We also removed all empty rows of data and found the log of the closed columns to make the data easier to use.



B. Trends and Insights

To make sure ARIMA works properly, we remove trends and make the data more stationary. The following is a graph of our data after is has been changed to be stationed.



IV. Improvements

Some ideas to improve the data accuracy is to use simple moving averages, as currently the model is using the last 2 years of data. By using only recent data such as 10 - 50 days of data, we can create more accurate predictions. We also considered the idea of combining some models, mainly random forests and linear regression, as if we can take only the pros of both and disregard the cons, our predictions become more accurate.

References

[1] I. Parmar et al., "Stock Market Prediction Using Machine Learning," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), Dec. 2018, doi: https://doi.org/10.1109/icsccc.2018.8703332.

[2] S. Du, J. Qiu, and W. Ding, "Research on Decision Tree in Price Prediction of Low Priced Stocks," Proceedings of the 2nd International Seminar on Artificial Intelligence, Networking and Information Technology, pp. 386–390, 2023, doi: https://doi.org/10.5220/0012284500003807.

[3] T. Phuoc, P. T. K. Anh, P. H. Tam, and C. V. Nguyen, "Applying machine learning algorithms to predict the stock price trend in the stock market – The case of Vietnam," Humanities and Social Sciences Communications, vol. 11, no. 1, pp. 1–18, Mar. 2024, doi: https://doi.org/10.1057/s41599-024-02807-x.