Using Machine Learning to Predict Berkshire Hathaway Stock

Team 17: Triton Eden, Faithful Odoi, and Gabe Lapham

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Abstract

The accurate prediction of stock prices is vital for financial planning and decision-making. This project leverages historical data of Berkshire Hathaway's stock prices to explore machine learning models for forecasting future price movements. A baseline model is established using existing approaches, followed by enhancements to improve prediction accuracy. The dataset is preprocessed through data cleaning, normalization, and exploratory data analysis (EDA) to uncover insights about trends, patterns, and correlations. The project aims to deliver a robust model that can capture both short-term fluctuations and long-term trends in stock price.

Motivation

- Develop a model that captures both short-term fluctuations and long-term trends in stock prices.

- Improve upon existing prediction approaches by establishing a baseline model and enhancing it.

Through exploratory data analysis (EDA),gain valuable insights about trends, patterns, and correlations in Berkshire Hathaway's historical stock data.

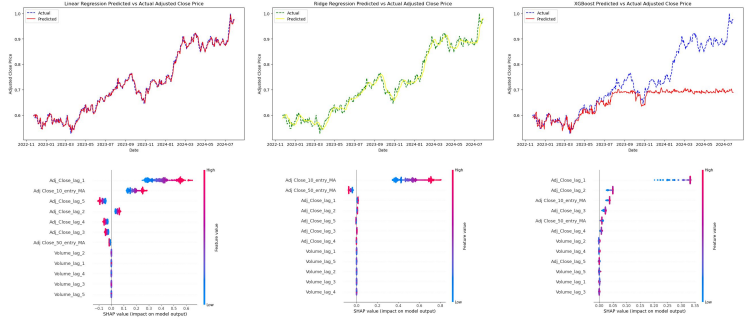
Methodology

We began with a simple linear regression model as a baseline to understand how well basic statistical methods could predict adjusted closing prices. This model features were the last five entries of the adjusted closing price and the volume of shares traded that day and a 10-day and a 50-day moving average of the adjusted closing price. To improve on the baseline we applied ridge regularization (L2) to our linear regression model and tried using more advanced machine learning models like a random forest regressor and an XGBoost regressor. Models were trained using a time-sensitive 80/20 train-test split and mean squared error (MSE), R^2 score, and bias-variance analysis were used to evaluate models.

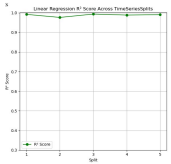
Model	MSE	R^2 Score	Observations
Linear Regression	8.615 x e^-05	0.994	While these results are encouraging the model seems prone to overfitting.
Ridge Regression	0.0002	0.919	Should have a lower tendency to overfit than linear regression.
XGBoost Regressor	0.012	0.176	The model still struggled to extrapolating trends into the future.

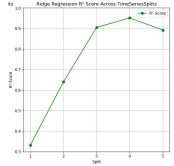
Results

The linear models performed better than the tree-based models. They might be more prone to overfitting but they are much better are extrapolating trends into the future.



Lessons Learned





- Time-series data works much differently than other data. First of all you need to do your train-test split without a random shuffle to prevent data leakage. Also I needed to have a 50 entry buffer between my train and test sets to prevent data leakage because of my 50 entry moving average model feature. We also learned that you can't use K-fold cross-validation on time-series data so we used a time series split for our validation.

- Sometimes ridge regularization doesn't make a linear regression model better. If the data doesn't suffer from overfitting or multicollinearity, ridge regression could actually hurt a model's performance.

Selected References

1. Berkshire Hathaway Stock Price Data
2. Bias and Variance in Machine Learning - geeksforgeeks
3. Predicting Time Series Data with Machine Learning, Generative AI, and Deep Learning - Medium

Acknowledgments

Heavily influenced by Dr. Santos's feedback on our midterm report to influence our feature selection, visualizations, prevention of data leakage, use of cross validation and XAI techniques. Used the create_lag_features() function from reference #3.