Future Stock Market Price Prediction with Supervised Machine Learning Techniques

Team 32: Shanna Wallace, John Paul Saia, Riley Taylor

THE UNIVERSITY OF TENNESSEE KNOXVILLE

Abstract

This poster details the progress made in the development of a machine learning model to predict future stock market closing price. Using a set of stock data from Berkshire Hathaway, we first created a simple baseline linear regression model, using closing stock price to predict the closing price seven days into the future. We then created a more complex ensemble random forest model using all the features in our data set. The methodology used to create the models is discussed, and their performance is compared.

Motivation

- Simplify process of analyzing market trends.
- Make it easier for companies and individuals alike to predict future stock market trends and make educated trading decisions.

To accomplish these goals, we have built a supervised machine learning model that aims to accurately predict the closing stock price seven days into the future.

Methodology

The Data Set

We trained our model using Berkshire Hathaway stock data (2015–2024) sourced from Kaggle.
The dataset includes the date, the trading volume, and the opening, closing, lowest, and highest prices for that day.
We split the data into training and testing sets, with 80% being used for training. To avoid data leakage, we made sure that our data points in the training set came from dates before the data points in the testing set.

Feature Selection

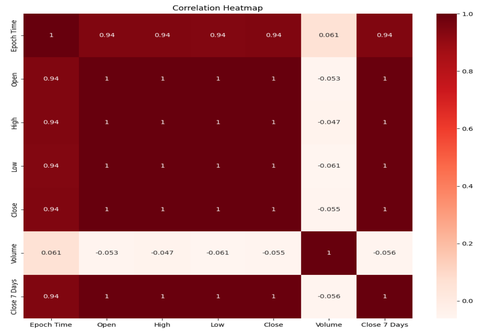
To evaluate the relationships between our features and our target, we first created a correlation matrix. We observed a strongly correlated linear relationship between our price variables and target in our dataset, as well as between the features themselves. When creating our random forest model, we used mean decrease in impurity, MDI, to evaluate the relative importance of the features.

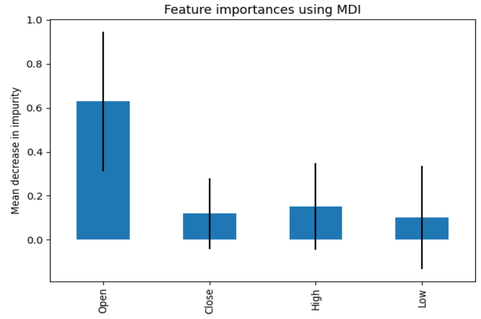
Baseline Model – Linear Regression

Since we observed such a strong linear relationship between our features and target, we decided that linear regression would be the most appropriate choice for our baseline model. Linear regression models are relatively simple to implement and easily interpretable, making them a good starting point for our project. We evaluated several linear regression methods, including polynomial expansion, OLS, and gradient descent, and found that Scikit Learn's default linear regression library, which uses OLS to fit the model, worked the best for us. Since highly correlated features can negatively impact model performance, we chose to work with only one feature for our simple model, the closing price.

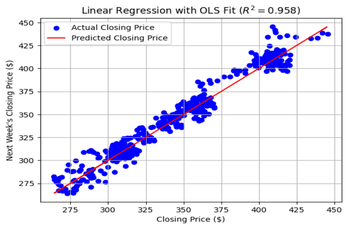
Ensemble Model – Random Forest

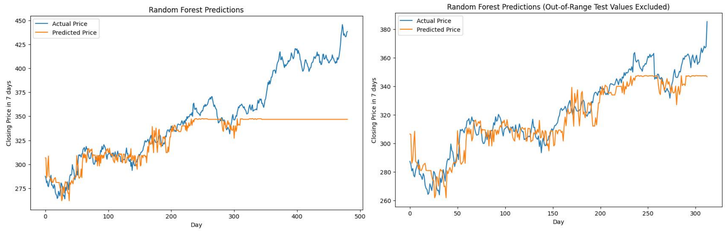
To capture more complex relationships between the features in our data, we created a random forest model. Using K-folds Cross Validation, we determined that six was the optimal max depth for the decision trees in our forest. We trained the model using the full data set of all features. MDI showed that Open price was the most important feature.





Results





Baseline Model – Linear Regression

R2: 0.9579 **RMSE:** $9.1153 **MAE:** $7.3144
Confidence Interval: [327.9427, 363.8392]

Our R2 value of 0.9579 indicates that over 95% of changes in our dependent variable, Cost in 7 Days, can be explained by changes in our independent variable, Cost. This indicates a high goodness-of-fit and strong relationship between our input data and model output. Our MAE of 7.3144 means that on average, our model's predictions differed from the actual values by about $7.31. Our RMSE, which also measures the average difference between actual values and our model output while giving higher penalty to outliers than MAE, was $9.12.

Ensemble Model – Random Forest

R2: 0.9886 **RMSE:** $34.1382 **MAE:** $23.3328
Confidence Interval: [215.0355, 244.1674]

Our random forest model has a similar R2 value to our baseline linear regression model. However, the RMSE and MAE are significantly higher.

Ensemble Model – Random Forest (Filtered)

R2: 0.9886 **RMSE:** $10.5194 **MAE:** $8.6036
Confidence Interval: [296.8732, 338.4019]

When we remove values in the testing data that are outside the range of values encountered in the training data, model performance improves significantly, coming closer to the performance of the baseline linear regression model.

Discussion

Our simple linear regression model shows superior performance compared to our ensemble random forest method, giving the most accurate predictions with lowest average error. This is unsurprising since our features had such a strong linear relationship with each other, as well as with the target. A linear model is the most appropriate choice to model this type of data.

Since our data is chronological, it was split so that data points in the test set are from dates later than data points in the training set. This can lead to poor performance with a random forest if the later dates happen to contain closing prices higher than any date in the training set, since the random forest model's prediction is limited to values it saw in the training data.

Unfortunately, this was the case with our data set. When dates that contain prices higher than the maximum price in the training set are filtered out of the testing data, model performance improves significantly. This demonstrates how negatively our random forest model's accuracy is impacted when it encounters unseen data values that are outside the range of data points it saw in the training data.

Lessons Learned

A major takeaway from this project is that sometimes, simpler is better. Introducing complexity into our model negatively impacted our results. We achieved our most accurate predictions using a simple linear regression model with only one feature.

Another takeaway is the importance of splitting data correctly. Our initial baseline model was inaccurate due to data leaks introduced by splitting the data randomly instead of chronologically, and our random forest model performs poorly due to the limited range of values in the training set. Issues such as these must be considered carefully to ensure the model is producing accurate results.

References

[1]https://www.kaggle.com/datasets/umerhaddii/berkshire-hathaway-stock-price-dataUmer Haddii “Berkshire Hathaway Stock Price Data,” *Kaggle*, Aug. 02, 2024.
[2] https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LinearRegression.html
[3] https://docs.aws.amazon.com/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html
[4] https://thedatascientist.com/performance-measures-rmse-mae/

Acknowledgments

We printed the poster at FedEx and received the dataset from Kaggle. We would also like to highlight Shanna's exceptional contribution for leading our group.