

CS 210 Project Phase 3

Data Verification

Evaluation of Models

Both models were evaluated on the test set using MSE.

Model Training

Two models were trained: kNN and Random Forest regressors.

Data Splitting:

The data was split into training and testing sets using a 80-20 ratio.

Applying Machine Learning Models

K Nearest Neighbor Model

I applied with a different number of neighbors and I find the most effective one is 5. I got the following mean squared error 156.45 which very high this indicates there is problem could be caused by non-linearity because python search is not rising as also expected Microsoft_Price is increasing much higher than python searches. Also, there should be insufficient features.

Random Forest Model

I apply different number of number of estimators. But the result has not effectively changed so I decided to make it 100. Again, the result is in very large numbers. Also random state is 42 selected.

Conclusion

In the context of using Microsoft price data to predict Python data, the slightly better performance of the KNN model could be due to its ability to capture local, non-linear patterns. The Microsoft price data and Python data may exhibit complex relationships that are not entirely linear. These relationships may not follow a simple linear pattern and could vary across different time periods or market conditions. KNN's ability to capture local patterns makes it suitable for modeling such non-linear relationships. This approach allows KNN to adapt to the local structure of the data, which can be beneficial when dealing with complex relationships. Therefore, in the context of predicting Python data using Microsoft price data, the KNN model's better performance may be attributed to its ability to capture the non-linear relationships and handle the noise present in the financial data.