CS210 Project Step 3 – Explanatory PDF

Teoman Arabul – 32283

This project investigates the impact of the COVID-19 pandemic, specifically the government-imposed Stringency Index, on the stock prices of Netflix. We will do that using two machine learning techniques: K-Nearest Neighbours (KNN) and Random Forest Regression.

2. Data & Preprocessing

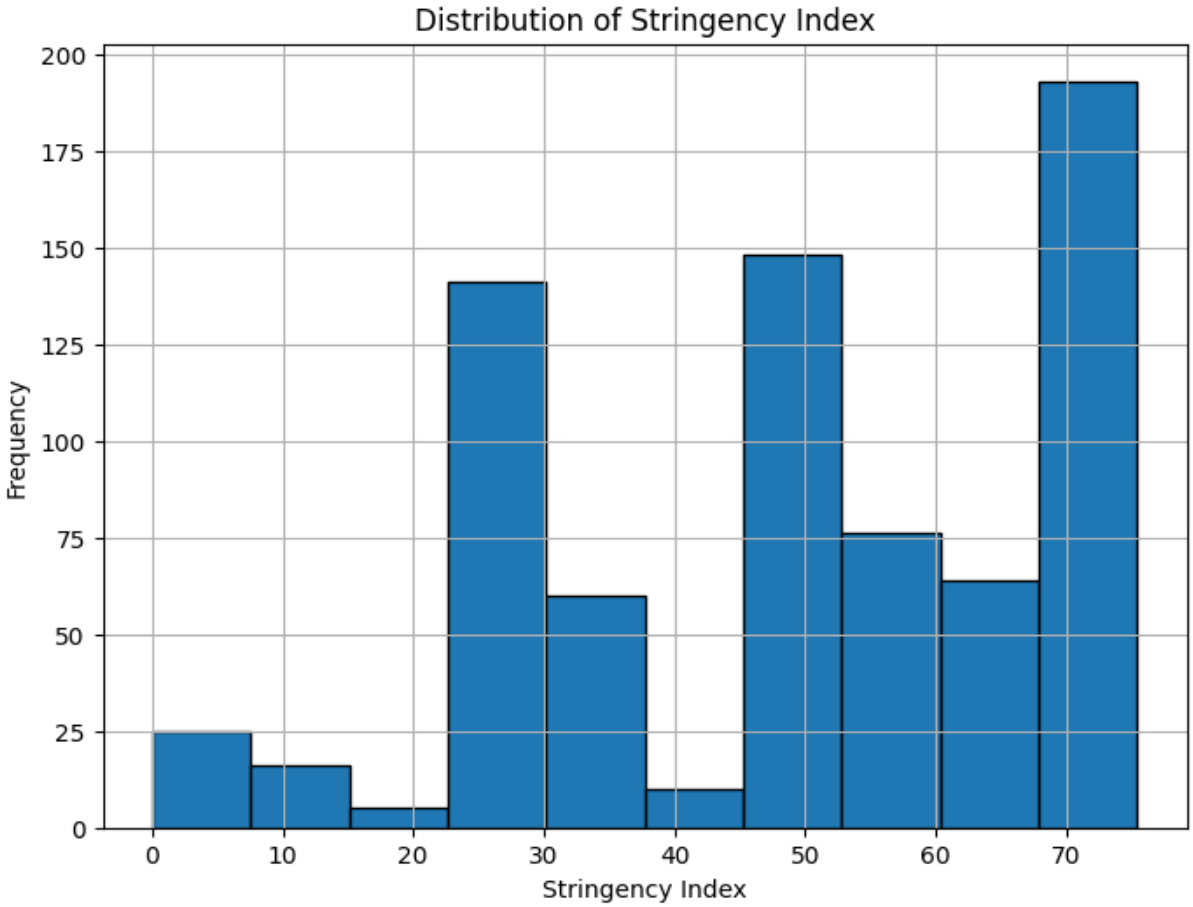
We use two datasets:
- "owid-covid-data.csv": Contains daily data on the Stringency Index for various countries.
- "Stock Market Dataset.csv": Contains historical stock prices for several US companies, including Netflix.

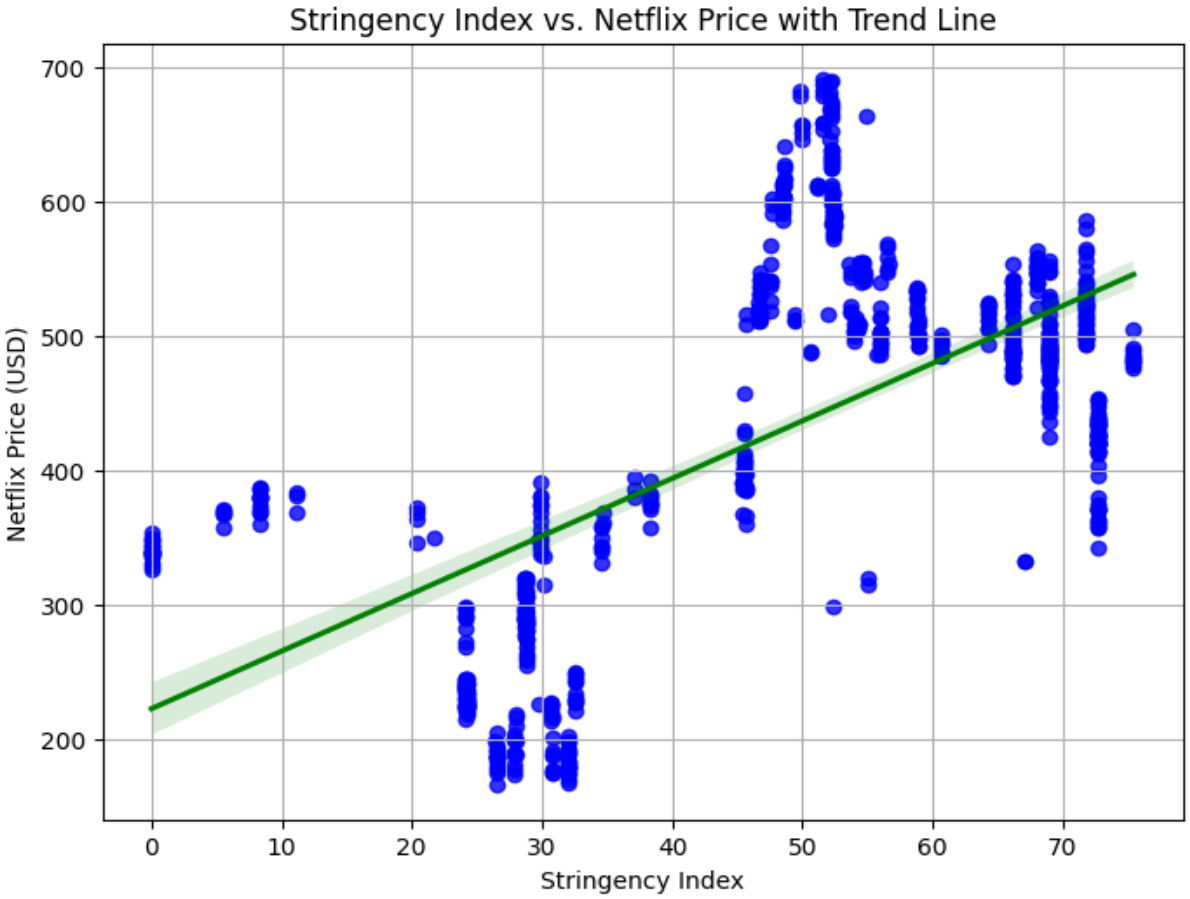
Preprocessing steps:
- **Data Cleaning:** Handling missing values, converting data types to appropriate formats.
- **Filtering:** Isolating US data from the Stringency Index dataset and merging it with the stock prices based on the date.
- **Normalization:** Using MinMaxScaler to normalize both Stringency Index and Netflix price to a 0-1 range for improved model performance.

3. Exploratory Data Analysis (EDA)

Before applying ML techniques, we performed EDA:
- Descriptive statistics of the Stringency Index were calculated.
- A histogram visualized the distribution of the Stringency Index.
- A scatter plot with a regression line illustrated the relationship between Stringency Index and Netflix stock price, suggesting a potential positive correlation.
- Pearson's correlation coefficient and p-value were calculated to confirm a statistically significant positive correlation.

Look for the statistics in the next page →


Distribution of Stringency Index
Frequency
Stringency Index
0
25
50
75
100
125
150
175
200
0
10
20
30
40
50
60
70


Stringency Index vs. Netflix Price with Trend Line
Netflix Price (USD)
Stringency Index
200
300
400
500
600
700
0
10
20
30
40
50
60
70

Machine Learning Models

1 K-Nearest Neighbours (KNN) Regression

KNN predicts a data point's value based on the average of its 'k' nearest neighbours in the training set. We used cross-validation to determine the optimal 'k' value, which yielded the lowest Root Mean Squared Error (RMSE).

2 Random Forest Regression

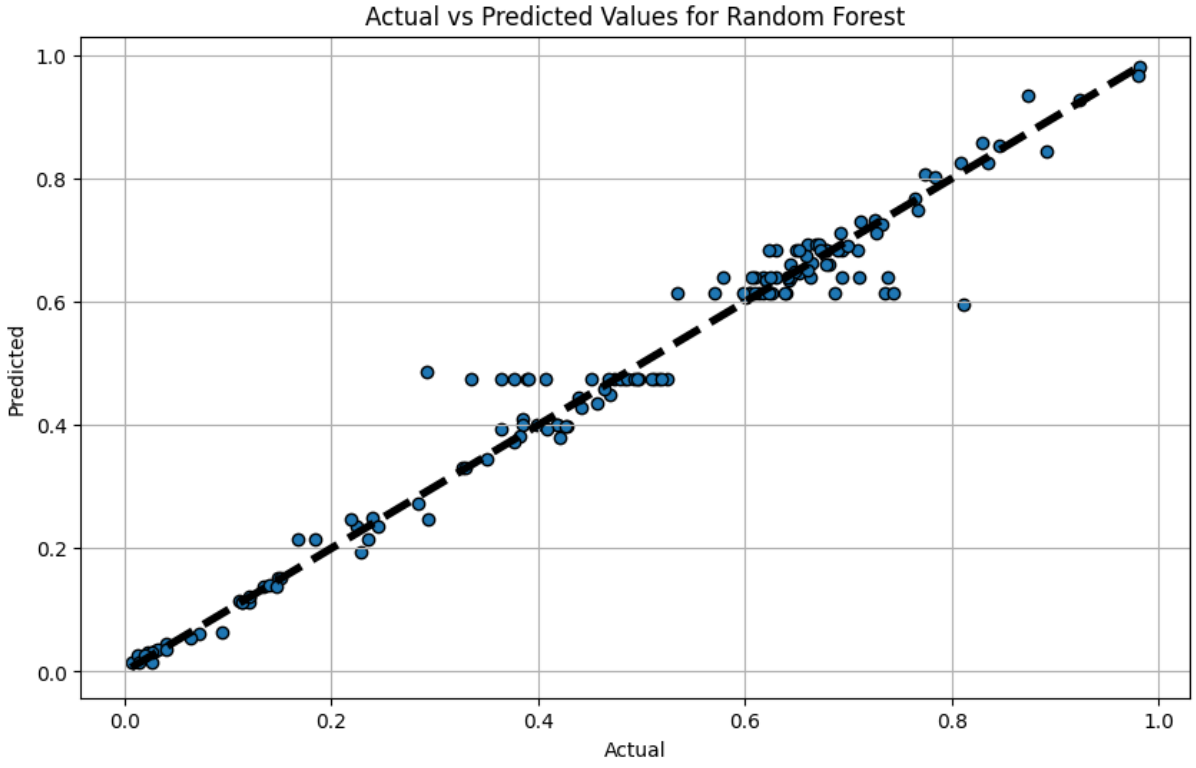
Random Forest constructs multiple decision trees during training and outputs the average prediction from all the trees. It's more robust and less prone to overfitting than a single decision tree. We used GridSearchCV to fine-tune hyperparameters like the number of trees, tree depth, etc., minimizing RMSE on the test set.

Model Evaluation and Comparison

We evaluated both models using RMSE as the primary metric. A lower RMSE indicates better predictive accuracy.

Results:

Model	RMSE
KNN Regression	0.05708781281185579
Random Forest Regression	0.0424969944383886



- **Random Forest consistently outperformed KNN** in our tests.
- This suggests that the relationship between Stringency Index and Netflix stock price might be **non-linear**, which Random Forest can model better.
- Random Forest's ensemble nature further improves its generalization ability by reducing the variance and risk of overfitting.

Feature Importance

- **KNN:** Doesn't provide an inherent mechanism for feature importance ranking. It treats all features equally based on distance calculations.
- **Random Forest:** Allows us to extract feature importances, indicating the relative contribution of each feature to the model's predictions. In our case, the Stringency Index has a feature importance score of 1.0. This means that the Random Forest model relies entirely on the Stringency Index to make predictions about Netflix stock prices.

In conclusion, this project demonstrates that machine learning techniques, particularly Random Forest Regression, show promise in predicting Netflix stock prices using the Stringency Index as a feature. However, relying solely on this index is insufficient for accurate prediction in real-world scenarios.