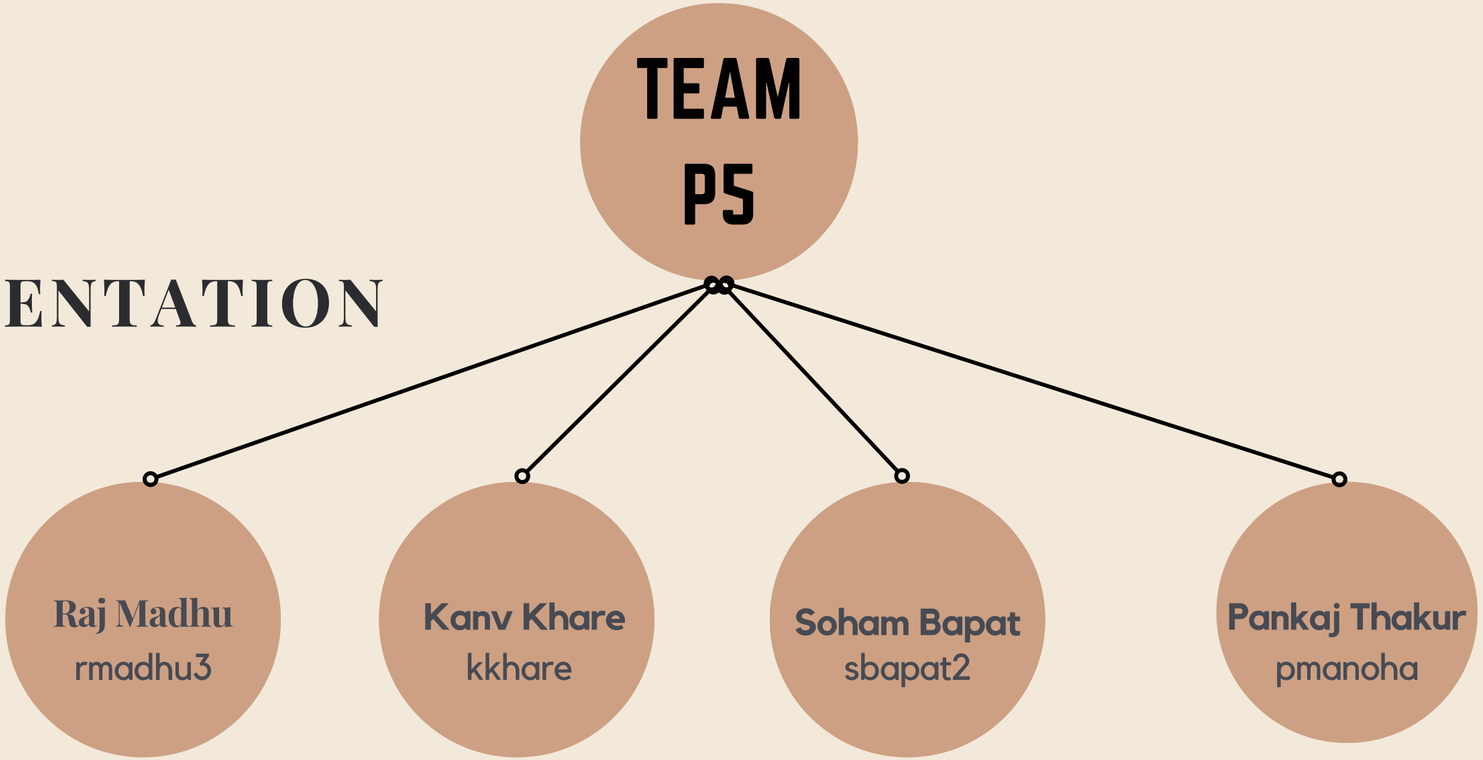
Features Affecting Oil Prediction

CSC522: ALDA FINAL PRESENTATION
DR. MIN CHI



Today's Agenda

1 Introduction

2 Data gathering and processing

3 Feature Selection and Elimination

4 Ensemble Model

5 Result & Conclusion

1. INTRODUCTION

- Predict West Texas Intermediate crude oil prices
- Do research and experimentation to identify features affecting oil prices due to various factors like wars, crime rate, precious metal prices, fuel reserves, economic conditions, etc.
- Feature selection and elimination on collected data
- Using Ensemble model to predict oil prices based on selected features

2.1 DATA GATHERING

- Reviewed various research papers to identify features influencing oil price prediction.
- Observed that most papers used only oil prices and applied time series analysis with previous day oil prices as input.
- Conducted Google research to identify 12 potential features correlated with oil prices.
- Developed a Python program to fetch feature prices from Yahoo Finance, saving data in CSV format.
- Merged CSV files to create a comprehensive dataset.
- The dataset comprises 4704 rows.
- Data spans from December 1, 2003, to December 31, 2021.

2.1 DATA OVERVIEW

WTI Oil price

WTI Crude Oil Spot Price is the price for immediate delivery of West Texas Intermediate grade oil, also known as Texas light sweet. It, along with Brent Spot Price, is one of the major benchmarks used in pricing oil. WTI in particular is useful for pricing any oil produce in the Americas.

Copper Price

The close price of copper

DJI Index

The Dow Jones Industrial Average, Dow Jones, or simply the Dow, is a stock market index of 30 prominent companies listed on stock exchanges in the United States

Henry Hub Natural Gas

The Henry Hub Natural Gas Spot Price measures the price in US Dollar per 1 Million Btu. The price of Natural gas notably spiked in February of 2003 when there was a large shortage in natural gas.

Euro

Euro price in comparison to dollar

Gold Price

The close price of gold.

Rubble

Rubble price in comparison to dollars

Silver Price

The close price of silver

[1,2,3]

2.1 DATA OVERVIEW

NASDAQ

The NASDAQ Composite is a stock market index that includes all companies listed on the NASDAQ stock exchange, featuring technology and various other sectors. It serves as a benchmark for the overall performance of the NASDAQ market.

SP500

The S&P 500, or Standard & Poor's 500, is a prominent stock market index in the United States, tracking the performance of 500 of the largest publicly traded companies. It's a key indicator of the U.S. stock market and economic conditions

Heat

Heating oil futures are derivative contracts that allow traders and investors to speculate on or hedge against the future price movements of heating oil.

Corn price

Corn futures are financial contracts that allow traders to buy or sell a specified quantity of corn at a predetermined price on a future date. These futures contracts are traded on commodity exchanges, such as the Chicago Board of Trade (CBOT) in the United States

Palladium price

The close price of Palladium

2.2 DATA PROCESSING

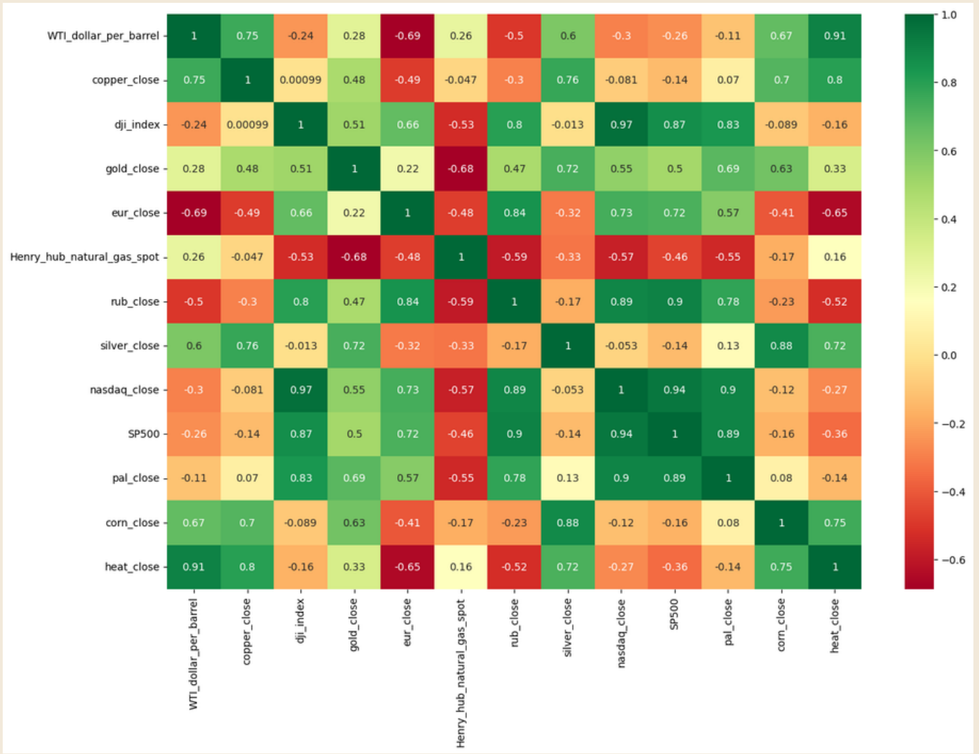
- Thoroughly examined the data for any null or missing values
- Employed bfill() method from Pandas library to handle missing data
- Substracted “Date” and “Value” columns from each individual dataset and performed a merge operation based on “Date”
- Concequently, our dataset comprises of 13 features along with “Date”
- These 13 features will be used for feature selection and eleminiation in the next steps
- Normalized the data using MinMaxScaler() function for our machine learning model

3.1 FEATURE SELECTION

- Definition: Choosing a subset of relevant and significant features (variables or attributes) from a larger set of features
- Reduces overfitting, improve generalization, and enhance the model's predictive performance.
- We are using Heatmap, XGBoost, Genetic Algorithms for feature selection
- Once features are selected, we perform feature elimination

3.1.1 HEATMAP

- Definition: Square matrix displaying pairwise correlations between variables in a dataset.
- Calculation: Uses the Pearson correlation coefficient (r) to measure the strength and direction of linear relationships.
- Correlation Values: Numeric values of 'r' range from -1 to 1.
  - Negative value near -1 suggests a strong negative correlation.
  - Value close to 0 indicates no correlation.
  - Positive value near 1 suggests a strong positive correlation.
- Identification of Correlated Features with WTI oil prices:
  - heat_close correlation: 0.91
  - copper_close correlation: +0.75
  - eur_close correlation: -0.69



3.1.2 XGBOOST(EXTREME GRADIENT BOOSTING)

- Definition: Popular machine learning algorithm for classification and regression. Utilizes ensemble learning with decision trees as base learners.
- Feature Importance in XGBoost: Determines the significance of each feature in making predictions. Helps understand the model and potentially improving performance.
- Gradient Boosting Framework: XGBoost is based on the gradient boosting framework.
- Boosting Technique: Combines weak learners to create a strong learner.
- Sequential Tree Building: Builds an ensemble of decision trees sequentially.
- Calculation of Feature Importance:
  - Weighted Importance: Calculated based on the number of times a feature is used to split data across all trees in the ensemble.
  - Relative Frequency: Measures how frequently a feature appears in tree nodes as a split criterion.
- Top Features with High Correlation with WTI Oil Prices:
  - Henry_hub_natural_gas_spot
  - Copper_close
  - Heat_close

[5]

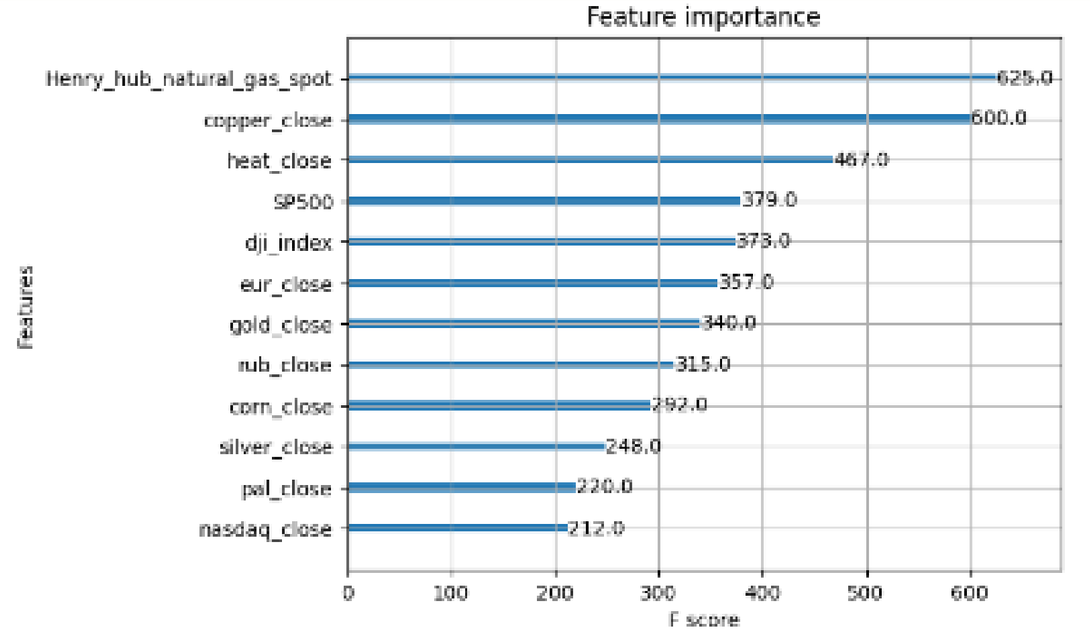


Figure 2: XGBoost Feature Importance

3.1.3 GENETIC ALGORHTMS

Main idea: Use genetic evolution to select features

Start with a random population of features (n=30)

Evaluate fitness of population (three-fold CV, MSE)

Mutate for genetic diversity (20% probability)

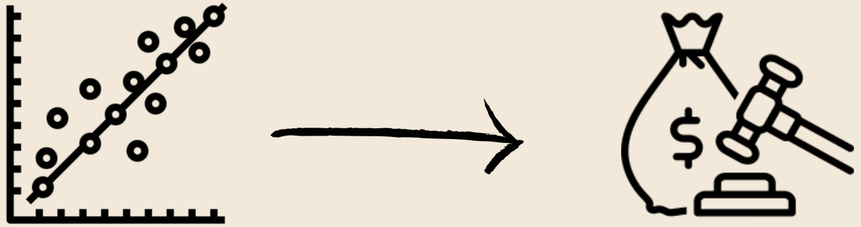
Encourage crossovers (60% probability)

Evaluate fitness of population and move on to the next generation

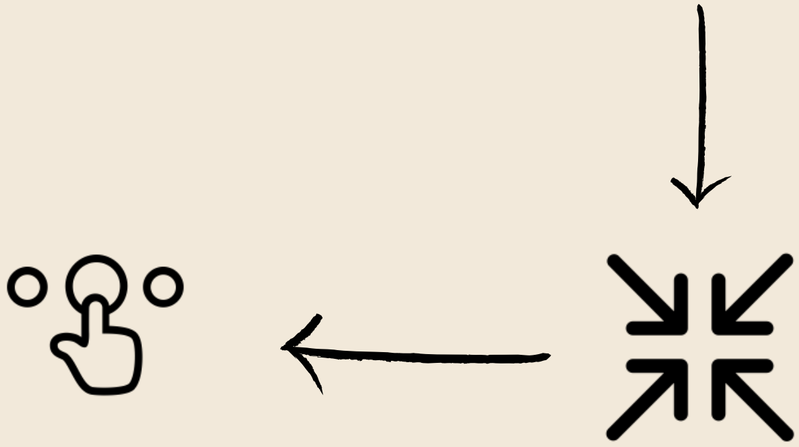
3.1.4 Least Absolute Shrinkage and Selection Operator (Lasso) Method

Overview: Lasso combines Linear Regression, Regularization and Automatic Feature Selection

Methodology



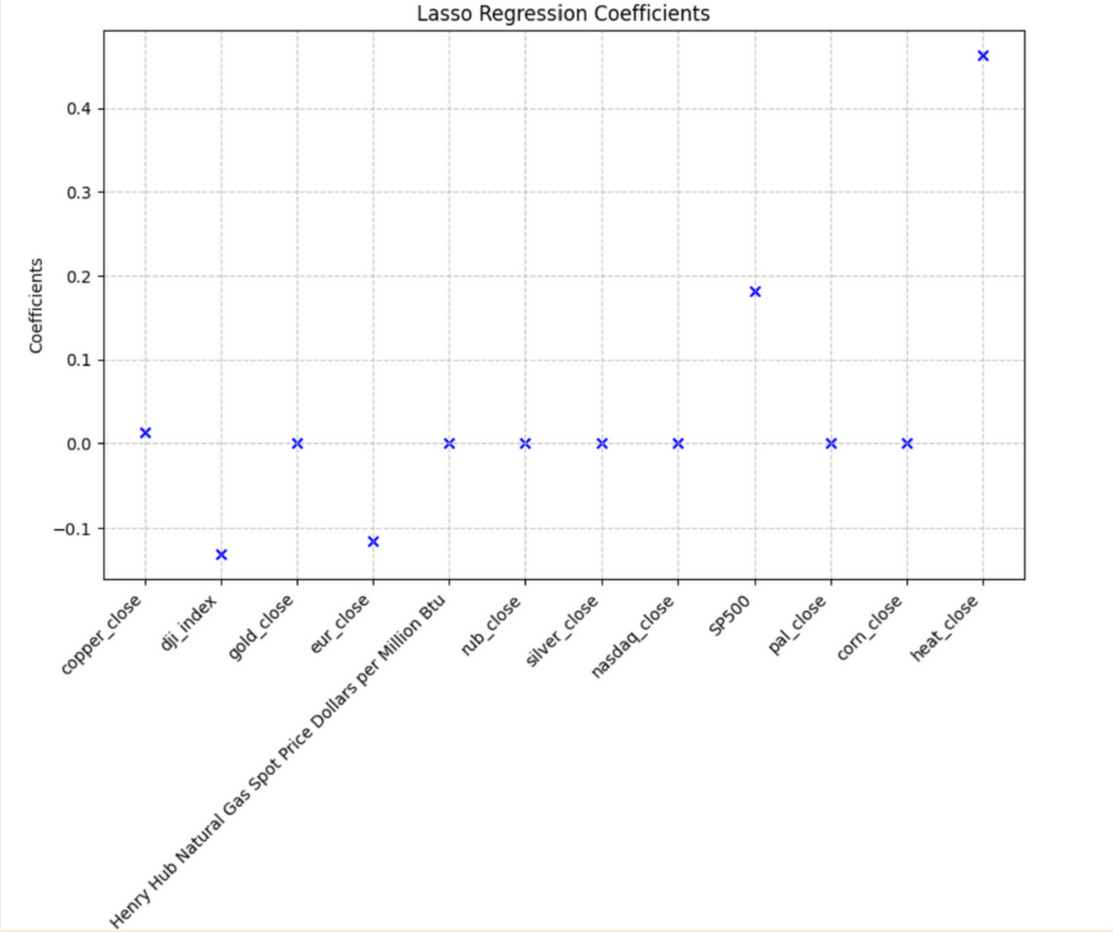
Linear Regression Basis **Regularization Term (L1 Penalty):**



Feature Selection **Feature Coefficient Shrinkage**

Selected Features

copper_close, dji_index, Henry hub, eur_close, heat_close



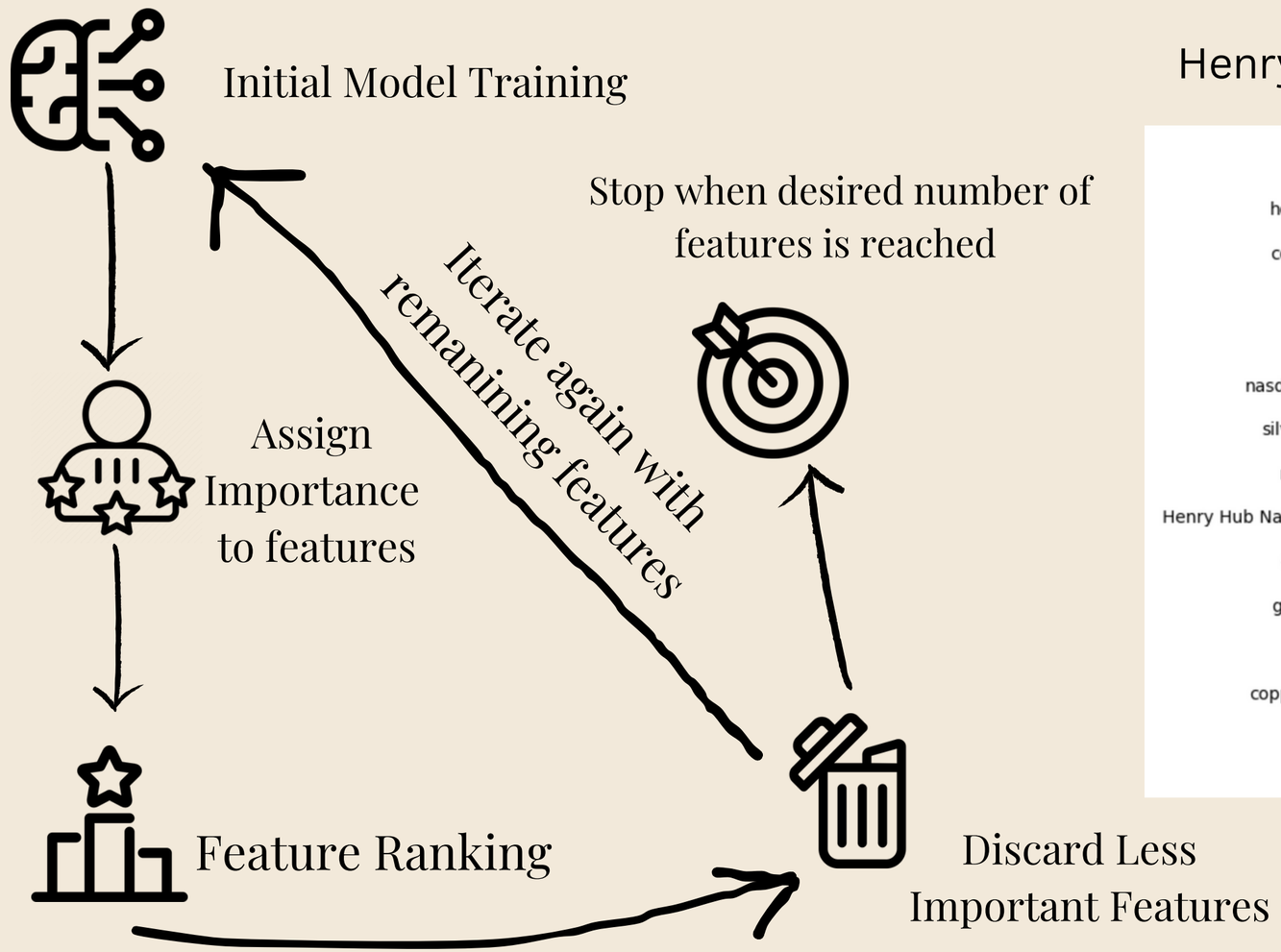
3.2 FEATURE ELIMINATION

- Definition: Process of systematically removing less relevant or redundant features from a dataset.
- Improve model simplicity, reduce overfitting, and enhance computational efficiency.
- We are using Recursive Feature Elimination for feature selection*
- Once features are eliminated, we apply our models*

3.2.1 Recursive Feature Elimination(RFE)

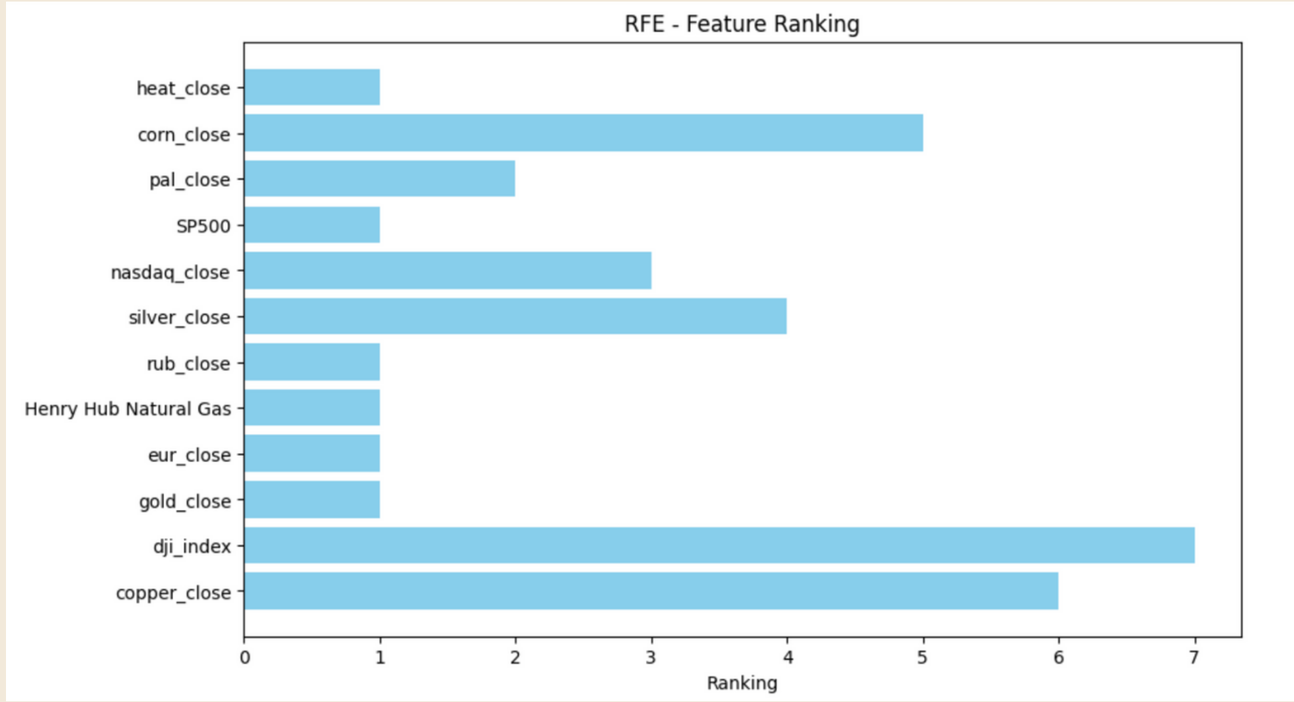
Overview: RFE iteratively eliminates less important features using importance-ranking metric to identify the most relevant subset of features.

Methodology

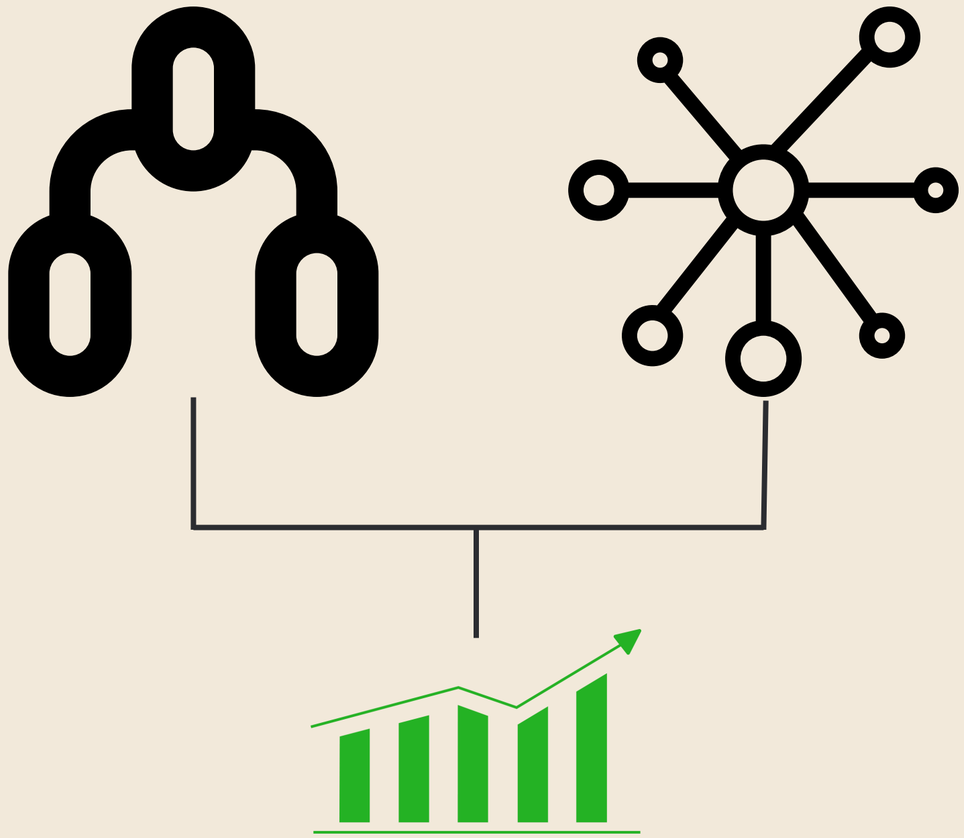


Selected Features

Gold_close, Eur_close, Henry_hub_natural_gas_, Rub_close, SP500, Heat_close



ENSEMBLE MODEL



- **Objective:** Predict current-day oil prices based on different feature combinations.
- Utilized an ensemble approach combining KNN and Random Forest models.
- KNN Configuration: Used 10 neighbors for the KNN model.
- Random Forest Configuration: Employed 128 estimators for the Random Forest model.

- Model Training and Prediction:
  - Data Fitting: Fitted the dataset to both the KNN and Random Forest models.
  - Prediction Process: Applied the trained models to testing data.
  - Ensemble Prediction: Final prediction derived as the average of predictions from both models.

METHOD	FEATURES SELECTED	MEAN SQUARE ERROR
HEAT MAP	Heat, Copper, Euro	**37.075**
XGBOOST	Henry Hub, Copper, Heat	**16.321**
Genetic	Copper, DJI, NASDAQ, PAL, Heat	**30.434**
Lasso	copper_close, dji_index, Henry hub, eur_close, heat_close	**69.09**
RFE	Gold_close, Eur_close, Henry_hub_natural_gas_spot, Rub_close, SP500, Heat_close	**3.33**

CONCLUSION

- After experimenting with different feature selection and feature elimination techniques we came to a conclusion that Recursive Feature Selection (RFE) gives the best features for the input and we get an mean square error of **3.33** using the Random Forest and KNN ensemble model.

- Using ***Gold_close, Eur_close, Henry_hub_natural_gas_spot, Rub_close, SP500*** and ***Heat_close*** as input features, we were able to predict the WTI oil prices with minimal error.

- Our research can help predict future oil prices and can also be used to study how different features affect oil prices.

Conclusion and Citations

1. Zhang, Xuan. “Dynamic Correlations between Crude Oil Futures Prices.” Energy RESEARCH LETTERS, vol. 3, no. 1, 17 Jan. 2022, https://doi.org/10.46557/001c.30057.
2. Cen, Zhongpei, and Jun Wang. “Crude Oil Price Prediction Model with Long Short Term Memory Deep Learning Based on Prior Knowledge Data Transfer.” Energy, vol. 169, Feb. 2019, pp. 160–171, https://doi.org/10.1016/j.energy.2018.12.016
3. Chen, Yanhui, et al. “Forecasting Crude Oil Prices: A Deep Learning Based Model.” Procedia Computer Science, vol. 122, 2017, pp. 300–307, https://doi.org/10.1016/j.procs.2017.11.373.
4. Guo, Junhui. “Oil Price Forecast Using Deep Learning and ARIMA.” 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Nov. 2019, https://doi.org/10.1109/mlbdbi48998.2019.00054.
5. Zhou, Yingrui, et al. “A CEEMDAN and XGBOOST-Based Approach to Forecast Crude Oil Prices.” Complexity, vol. 2019, 3 Feb. 2019, pp. 1–15, https://doi.org/10.1155/2019/4392785.