PREDICTING ARABICA COFFEE PRICES USING WEATHER PATTERNS IN BRAZIL

MASEEH FAIZAN

Abstract. This paper tries to determine global Arabica Future prices given weather patterns in Brazil. Brazil accounts for 40% [5] of world's Coffee production and main producer of Arabica Future prices, so variation in its production capacities has an impact on Coffee prices around the world. Coffee beans, like most of the commodities, are sold in US Dollars, so I also took into account the exchange rates between Brazil Real and US Dollars (BRL/USD). Using time series data and Machine Learning models, I was able to predict a steady rise in coffee prices. Machine learning models were effective in predicting prices with weather patterns. Ordinary Linear model, even in Machine Learning context, didn't yield effective results

Key words. Coffee, Climate Change, Weather, Brazil, Arabica Futures

1. Introduction.

Coffee is one of the most popular and most consumed hot beverages in the world. The world's coffee production is expected to reach 10.62 [6] Billion kg in 2023 with Brazil accounting for 40% of that production. Other countries include, Columbia, Vietnam and Ethiopia. Coffee is also one of the world's most traded commodities, with the European Union (EU) and Switzerland as top importer. The EU countries, trades 90% of unroasted green coffee beans, which are mostly Arabica Coffee beans, 77% of the trade happening in Switzerland.[6] As coffee beans are agricultural goods, produced mainly in Latin America, variation in weather patterns in the producing countries, due to Climate Change, will have an impact on its prices. Since Coffee is also a highly traded commodity, we can also assume that it is well priced with no arbitrage opportunity. Finally, unlike other commodities like petroleum, geopolitics has very little impact on its prices. I am applying various machine learning models to forecasting future prices of Arabica Coffee beans.

2. Description of the research question and the relevant literature.

In the realm of financial markets, commodity prices fundamentally hinge on the principles of supply and demand inherent to an efficient market. Conversely, stock prices are influenced by a multitude of factors including changes in management, corporate scandals, geopolitical events, and shifts in investor sentiment toward a specific company, rendering stocks considerably more volatile and challenging to predict. Among various commodities, petroleum stands out due to its significant geopolitical implications, largely influenced by The Organization of the Petroleum Exporting Countries (OPEC), which operates akin to a cartel and plays a pivotal role in controlling oil prices. In contrast, prices for agricultural commodities such as coffee, sugar, soy, and corn are predominantly governed by basic supply and demand dynamics and tend to exhibit greater stability.

The focus of this analysis is on coffee, a commodity whose price dynamics are particularly interesting due to the stable and progressively increasing global demand. The ubiquity of coffee consumption, whether directly or through caffeine-containing beverages, supports a relatively stable demand side for coffee beans. Thus, significant price fluctuations are primarily attributable to variables affecting supply.

Latin America, and Brazil in particular, are critical in the global coffee market as major producers. Disruptions in this region's coffee production can, therefore, have substantial repercussions on global coffee prices. A noteworthy study by the

United Nations identified weather patterns as the principal driver of Arabica coffee price volatility in Latin America. The research utilized time series data to explore how climate change, by introducing more frequent and severe weather anomalies, adversely impacts production capacity and, consequently, elevates Arabica coffee bean prices.

Moreover, the study underscored the importance of considering exchange rate fluctuations in the analysis. The primary consumers of coffee, located in the U.S. and Europe, conduct transactions in U.S. dollars, while producers, particularly in Brazil, receive payment in their local currency, the Brazilian real. Accounting for these currency dynamics is essential in constructing a robust model to analyze coffee prices, enhancing the understanding of how external economic factors interplay with agricultural production to influence commodity markets.

3. The methodology applied to address the research question. In contrast to previous research that predominantly employed linear and non-linear regression models with multiple control variables to analyze the impact of climate change on Arabica coffee prices using time series data, this study initially employs Ordinary Least Square (OLS) models with HC3 covariance type and Huber T. models. Although the results from these models prove to be statistically significant, they lack economic significance and are not well-suited for forecasting purposes with R^2 close to 0.

Building upon these findings, the present study aims to develop advanced machine learning models to forecast future Arabica coffee prices. This approach leverages historical price data and weather patterns, acknowledging the significant influence of climate change on coffee production. Previous findings, including a 2015 article by Michon Scott [4], indicate that Arabica coffee beans thrive in temperatures between 18-21°C. However, recent data from Brazilian coffee farming regions show temperature fluctuations ranging from 15°C to 24°C, which complicate coffee production and harvesting.

This study will collect and analyze average daily temperatures and precipitation data from cities in Brazil where Arabica coffee is produced. The data will be processed through four distinct machine learning models, each undergoing hyperparameter optimization. A rigorous cross-validation process will determine the model with the highest predictive accuracy, establishing its superiority for forecasting purposes.

4. A description of the data set. Obtaining precise and comprehensive historical weather data poses significant challenges. Initially, the Open Weather API [1] was considered for acquiring rich historical weather data; however, the costs associated with downloading extensive historical data were prohibitive. A more fruitful approach was subsequently employed using MeteoStat [3], an open-source platform supported by community contributions and Patreon. MeteoStat provides a Python API that locates the nearest weather station based on provided longitude and latitude coordinates, offering detailed historical data, including daily weather conditions and precipitation levels.

To effectively utilize the MeteoStat API, it was essential to first acquire geographic coordinates for cities and farms within Brazil, where coffee is predominantly cultivated. This data was sourced from Simplemaps [2], which organizes longitude and latitude information by city and administrative regions within Brazil. Subsequent research identified Minas Gerais and São Paulo as the primary coffee-producing regions. Cities outside these areas were excluded from the analysis, focusing the data collection on relevant locations through the MeteoStat platform. Focusing on cities where most of the farming occurs helps to have cleaner data. The resulting dataset includes daily time series data on average temperatures and precipitation from 2019 to 2024. Data preceding 2019 was excluded due to its inconsistency and frequent absence of daily records.

In parallel with weather data collection, daily Arabica coffee bean prices were retrieved from the Yahoo Finance API, focusing on futures prices which reflect both current and anticipated future market conditions. Arabica coffee futures are traded daily, indicating a highly efficient market environment without price stickiness. This price data covers the same period as the weather data, from 2019 to 2024.

Additionally, daily exchange rates between the Brazilian Real and the US Dollar were downloaded from Yahoo Finance for the same time frame. This financial data, combined with the collected weather data, was compiled into a single panel data structure using the pandas DataFrame. Each row of the DataFrame corresponds to daily observations, integrating average weather conditions, Arabica coffee prices, and exchange rate information.

Analytical methods were then applied to this consolidated dataset. Specifically, the returns and cumulative returns on Arabica coffee prices were calculated to facilitate a more nuanced analysis of price trends. Cumulative returns, in particular, provide a clearer visual and analytical representation of price fluctuations over time, allowing for a comprehensive assessment of how weather patterns and economic factors influence coffee market dynamics.

This systematic approach to data collection and analysis underscores the complexities of forecasting commodity prices, highlighting the importance of integrating diverse data sources to enhance the robustness of predictive models in agricultural economics. At the end of the data collection and cleaning process, we had a time series panel data

5. A discussion of the implementation. Prior to any analytical procedures, the data is normalized by converting all prices into Brazilian Reals, thereby eliminating variations attributable to exchange rate fluctuations. Additionally, the study acknowledges a temporal lag between shifts in weather patterns and their impact on coffee futures prices. Typically, adverse weather conditions do not instantaneously affect coffee yields within the same year; rather, it may take several weeks for such conditions to significantly influence annual production and subsequently alter market prices. Consequently, the adjusted prices are shifted forward by 40 days in the dataset. This temporal adjustment ensures that the analysis accounts for the delayed effect of weather changes on coffee prices.

Initially, an examination of coffee bean price trends over several years was conducted, revealing a sharp increase from 2021 to 2022, followed by stabilization in 2022, and a subsequent decline approaching 2023. Subsequent analysis focused on the average daily temperatures across various cities within the production region. These temperatures exhibited a distinct cyclical pattern with a pronounced V-shape, peaking between September and February and declining between June and September. Typically, the temperatures ranged from lows of around 12°C to highs of approximately 25°C, occasionally reaching up to 30°C, an atypical occurrence for the region. Despite these variations, initial visual assessments did not suggest a correlation between average temperatures and price fluctuations.

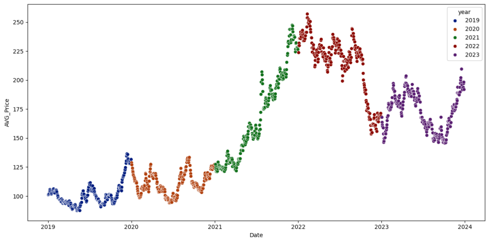

Fig. 5.1. *Coffee Prices*

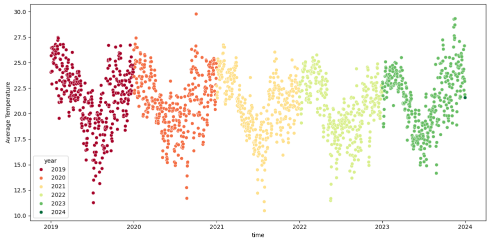

Fig. 5.2. *Average Tempratures*

A more structured approach involved conducting a visual linear regression analysis using Seaborn's regression plot, where the dependent variable was the coffee prices adjusted for exchange rate effects and lagged by 40 days, against the independent variable of non-lagged average temperatures. This analysis indicated a weak linear relationship, suggesting that higher temperatures might inversely affect coffee prices, although the presence of numerous outliers complicates definitive conclusions.

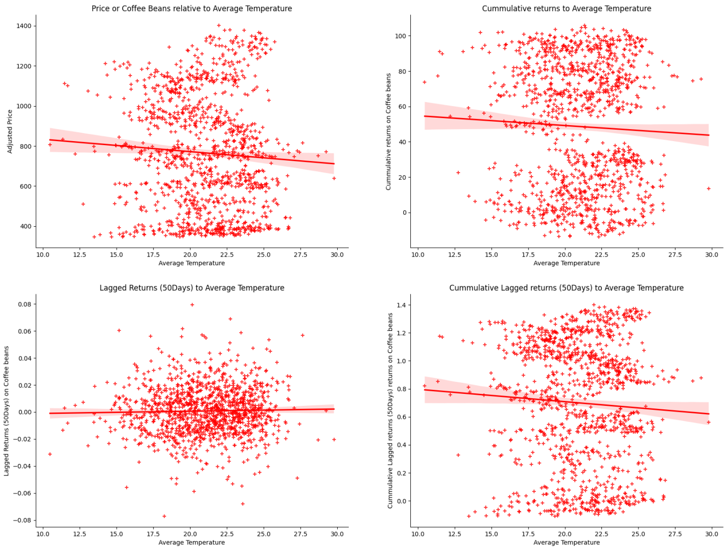

Fig. 5.3. *Prices relative to Average Tempratures*

Further analysis extended to examining precipitation patterns, which similarly showed cyclical behavior with minimal amounts for most of the year but significant increases in December and January. Despite these patterns, initial visual regression analysis on precipitation data failed to reveal any clear relationships with coffee prices. This lack of evident correlation underscores the complexity of these variables' interactions, necessitating further detailed analysis to ascertain the definitive impact of weather conditions on coffee pricing trends.

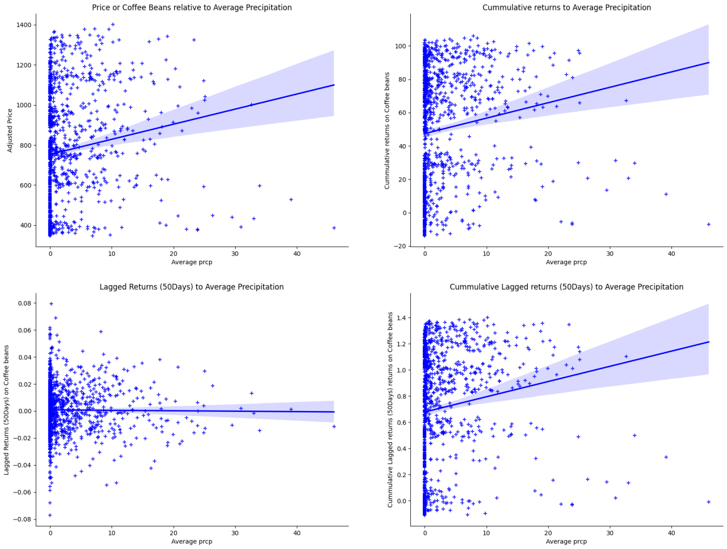

Fig. 5.4. *Prices relative to Average Precipitation*

Given the presence of outliers identified through visual analysis, I opted to employ robust regression techniques to enhance the reliability of the results. Specifically, I utilized the HC3 model, which is advantageous when dealing with heteroscedastic data—where variance varies across different values of the independent variable, X. The HC3 model is particularly effective, providing robustness even in smaller samples by correcting for variations in sample size. Additionally, I implemented the Huber T model, a robust regression method tailored to mitigate the influence of outliers. The Huber T model adjusts the impact of extreme values, ensuring they do not disproportionately affect the overall model accuracy.

Armed with initial findings, I shifted focus to leveraging machine learning techniques to enhance the prediction of coffee prices and associated weather pattern behaviors. Utilizing the SciKit Learn library, I set out to develop and implement a machine learning model tailored for this purpose.

The dataset for the model includes features such as average temperature, average precipitation, year, and month, with the target variable being the cumulative lagged returns. From lessons learned in academic settings, I adhered to best practices by splitting the data into training and testing subsets. Specifically, 80

Given the nature of the data, I categorized the variables into two types: categorical and numerical. The 'Year' and 'Month' were treated as categorical data. Although 'Month' is numerically represented from 1 to 12, it does not possess ordinal properties that justify numerical treatment, and 'Year' is categorized similarly for consistency in data handling. This categorization is crucial as it addresses the non-linear relationship these variables may have with the target variable.

Before proceeding with model training, it was essential to perform preprocessing on the dataset. For the numerical data—average temperatures and precipitation—I employed the 'StandardScaler' from SciKit Learn. This scaler standardizes the features by removing the mean and scaling each feature to unit variance. This normalization process transforms the features to have a zero mean and a standard deviation of one, which is particularly beneficial for algorithms that presume a normal distribution of input data. Standardizing the data in this way ensures that no single feature disproportionately influences the model due to scale differences.

Conversely, for the categorical data, I utilized the 'OneHotEncoder', also from SciKit Learn. This technique converts categorical variables into a binary numerical format. It achieves this by creating separate columns for each category within a feature, where the presence of a category is marked by '1' and its absence by '0'. This method is employed to effectively manage categorical input within machine learning models, ensuring that each unique category is properly represented without implying any ordinal relationship.

By preprocessing the data using 'StandardScaler' for numerical features and 'OneHotEncoder' for categorical features, the model is well-prepared to interpret the features accurately and perform robustly in predicting the dependent variables. This methodological approach not only aligns with academic teachings but also enhances the predictive capability of the machine learning model, providing a solid foundation for insightful data analysis in the context of coffee market dynamics.

To optimize the prediction of coffee prices and weather patterns, I employed various machine learning models, each with distinct characteristics and strengths, suitable for the complexities of the dataset at hand. Utilizing ScikitLearn's robust functionalities, I approached the model selection and hyperparameter tuning systematically.

The methodology began with defining a nested dictionary to organize the different models and their respective hyperparameters, an approach that aligns with ScikitLearn's requirements for implementing Pipeline and GridSearchCV frameworks. The models selected for this study were Random Forest Regression, Gradient Boosting Regression, Linear Regression, and K-Nearest Neighbors Regression, each chosen for its unique capabilities in handling specific aspects of the dataset.

6. Model Descriptions and Rationales. 1. **Random Forest Regression**: This model constructs numerous decision trees during training and outputs the mean prediction of these trees. It is particularly effective for datasets with high dimensionality and a mixture of numerical and categorical features, providing robustness against overfitting through its averaging mechanism.

2. **Gradient Boosting Regression**: By building an ensemble of weak prediction models sequentially, each correcting its predecessor's errors, Gradient Boosting optimizes both bias and variance, making it adept at handling complex, noisy datasets.

3. **Linear Regression**: As a fundamental regression approach, it assumes linearity between the dependent and independent variables. It includes options to manage multicollinearity, such as ridge regression or Lasso, enhancing model generalization through regularization.

4. **K-Nearest Neighbors Regression (KNN)**: This non-parametric method predicts outcomes based on the averages of the 'k' closest training examples, offering

high interpretability and flexibility, especially in scenarios where the data distribution does not meet conventional statistical assumptions.

To validate these models effectively, a 10-fold cross-validation was set up, splitting the data into ten parts to ensure that each fold serves as a robust test set. Additionally, an empty list was prepared to collect the results of each model's performance.

The process continued with implementing GridSearchCV, a powerful tool in ScikitLearn designed to conduct an exhaustive search over specified parameter grids. This method not only automates the optimization of model parameters but also ensures that the best model performance is achieved by testing all possible combinations of parameters across the folds of cross-validation.

During the GridSearchCV execution, each model was evaluated, and the results were documented, including the model name, the best score (typically the R^2), and the hyperparameters that led to the optimal performance. This phase was critical for identifying the most effective model configurations.

Following the identification of the best models from the grid search, another round of analysis was conducted. In this phase, a more refined set of hyperparameters, closely aligned with the best-performing models, was tested to fine-tune the models further. The performance of these configurations was assessed using both R^2 and Mean Squared Error (MSE) as metrics, providing a comprehensive view of each model's predictive accuracy.

The final step involved deploying the selected models with the identified optimal parameters to predict the outcomes. The predictions were then plotted alongside the actual coffee prices to visually assess each model's performance. This visualization was crucial in understanding the practical effectiveness of each predictive model in mirroring the real-world data and making informed decisions based on the model outputs.

Through this meticulous approach, leveraging ScikitLearn's advanced capabilities in machine learning model selection and hyperparameter tuning, the study aimed to not only predict coffee prices accurately but also to provide insights into the influential factors affecting these prices, such as weather patterns and their interaction with market dynamics.

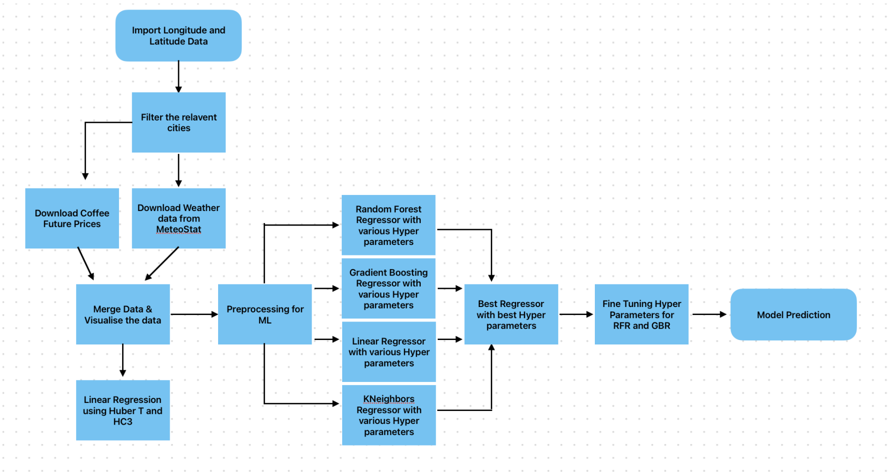
Import Longitude and Latitude Data
Filter the relavent cities
Download Coffee Future Prices
Download Weather data from MeteoStat
Merge Data & Visualise the data
Linear Regression using Huber T and HC3
Preprocessing for ML
Random Forest Regressor with various Hyper parameters
Gradient Boosting Regressor with various Hyper parameters
Linear Regressor with various Hyper parameters
KNeighbors Regressor with various Hyper parameters
Best Regressor with best Hyper parameters
Fine Tuning Hyper Parameters for RFR and GBR
Model Prediction

FIG. 6.1. *Summary of Implementation chart*

7. Results. In examining our linear regression models, which are robust and account for heterogeneity, we observe that the results are statistically significant. The coefficient for Average Temperature is approximately 0.9 at a 1% significance level, suggesting that at a theoretical temperature of zero, the baseline cumulative return of coffee beans is around 0.9. Furthermore, an increase in Average Temperatures correlates with a decrease in cumulative returns by approximately 0.009, a result statistically significant at the 5% level. However, these results lack economic significance, primarily because the highest average temperatures still fall within the optimal range for coffee production. This implies that any rise in temperatures could enhance production capacity, thereby reducing prices, as indicated by the negative coefficient.

Additionally, the analysis of Average Precipitation reveals that the constant across both models is approximately 0.67 at a 1% significance level, indicating that on days without rain, the cumulative returns stand at 0.67. However, each millimeter of rain increases cumulative returns by about 0.0116, also significant at the 1% level. This suggests that rainy days tend to increase cumulative returns and hence prices, which is counterintuitive since increased rainfall should theoretically boost production capacity and lower prices.

Given the mixed results and the potential non-linear distribution of data, the economic significance of these models remains questionable. This underscores the need for further investigation, possibly with non-linear modeling approaches, to better understand the dynamics at play.

Table 7.1

	[HC3] I	[HUBER] I	[HC3] II	[HUBER] II
const	0.8856***	0.9214***	0.6789***	0.6753***
	(0.0942)	(0.1028)	(0.0159)	(0.0155)
Average Temperature	-0.0089**	-0.0103**		
	(0.0044)	(0.0048)		
Average prcp			0.0116***	0.0142***
			(0.0033)	(0.0026)
R-squared	0.0032		0.0187	
R-squared Adj.	0.0023		0.0178	
NO. observations	1217	1217	1171	1171
R-squared	0.00		0.02	

Standard errors in parentheses.
* p¡.1, ** p¡.05, ***p¡.01

Machine Learning Models: The two models, Gradient Boosting Regression and Random Forest Regression, demonstrated superior performance. The efficacy of these models can be attributed to their inherent capabilities to navigate the complexities often present in datasets characterized by non-linear relationships and intricate interactions among variables.

Random Forest Regression proved adept due to its robustness in managing high-dimensional data. This model employs an ensemble of decision trees, each constructed from a randomly sampled subset of the data. At each node of a tree, a random subset of features is selected, contributing to diversity in the model’s predictions. This technique, known as ”bagging” or bootstrap aggregating, effectively mitigates variance without substantial increase in bias, thereby enhancing the overall prediction accuracy. Moreover, Random Forest’s capability to assimilate diverse data types and its proficiency in capturing non-linear variable interactions make it particularly suited for modeling the dynamic interplay between climatic conditions and coffee price movements.

On the other hand, **Gradient Boosting Regression** excels by constructing a sequence of weak prediction models, typically decision trees, where each tree incrementally corrects the errors of its predecessors. This iterative correction process enables the model to improve continuously, optimizing both bias and variance through successive refinements. Gradient Boosting is adept at handling complex and noisy data, a common characteristic of datasets where weather conditions directly influence economic outcomes. Unlike Random Forest, which generates trees independently, Gradient Boosting strategically focuses on areas poorly explained by previous models, thereby enhancing model precision over iterations. This methodological approach is particularly effective in detecting subtle patterns within the data, crucial for forecasting the effects of nuanced meteorological variations on coffee prices.

Both models leverage distinct methodologies to address the challenges inherent in predicting economic indicators influenced by environmental factors. **Random Forest’s approach of reducing variance through averaging multiple independent trees and Gradient Boosting’s technique of focusing on sequential error reduction** coalesce to form robust predictive frameworks. These models not only demonstrate high accuracy in forecasting, but also provide insights into the complex relationships between weather patterns and coffee market dynamics, underscoring their utility in economic and meteorological studies.

Figure 7.1 & Figure 7.2 Figure 7.1 illustrates that while both Random Forest Regression, in green, and Gradient Boosting Regression, in blue, provide reasonable price predictions, in red, a noticeable lag exists between the predicted and observed prices, attributable to the predictive nature of the models and a deliberate shift in cumulative returns by several weeks. In contrast, Figure 7.2, which corrects for this lag, by shifting back the returns and displays significantly improved alignment between the model predictions and actual prices, demonstrating enhanced accuracy in the forecasting model.

Figure 7.1 & Figure 7.2 These are the hyperparameters used for both Gradient Boosting Regressors and Random Forest Regressor to have best R-sq results using Scikit Learn without overfitting.

```

# This is the Pipline for Gradient Boosting Regressor
pipeline = Pipeline([
    (’preprocessor’, preprocessor),
    (’regressor’, GradientBoostingRegressor(
    # These are the hyperparameters that result in the best
    R-Sq results
        learning_rate=0.1,
        max_depth=6,
        min_samples_leaf=1,
```

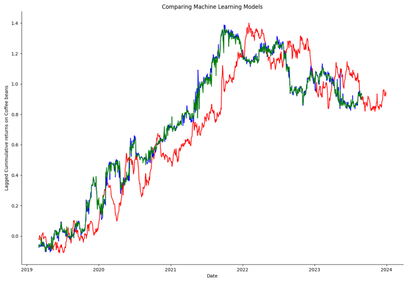

Fig. 7.1. *Models Optimizing for R-Sq lagged*

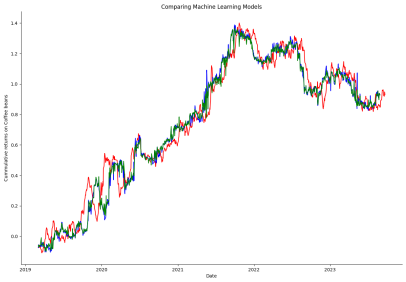

Fig. 7.2. *Models Optimizing for R-Sq*

```
        min_samples_split=2,
        n_estimators=150,
    ))
])
pipe =pipeline.fit(X_train, y_train)

###########################################################

# This is the Pipline for Random Forest Regressor
pipeline = Pipeline([
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(
    # These are the hyperparameters that result in the best
    # R-Sq results
        max_depth=25,
        max_features='sqrt',
        min_samples_leaf=1,
        min_samples_split=4,
        n_estimators=150,
    ))
])
pipe = pipeline.fit(X_train, y_train)
```

Listing 1
Processing data

Figure 7.3 & Figure 7.4

In these figures, the actual Arabica coffee prices are depicted by the red line, predictions from the Gradient Boosting Regressor by the blue line, and those from the Random Forest Regressor by the green line. The Gradient Boosting Regressor closely mirrors the actual prices, demonstrating enhanced accuracy in Figure 7.4 after correcting for the lag, thus significantly improving the predictive performance. Conversely, the Random Forest Regressor displays considerable inconsistency, especially when optimized for Mean Squared Error (MSE). Characterized by a relatively flat trend interspersed with abrupt fluctuations, this model's limited efficacy is attributable to a restrictive max_depth parameter set to 1, in stark contrast to the deeper 25 layers used previously. These observations suggest that in this instance, the Random Forest Regressor is less suitable for predicting coffee prices based on weather patterns.

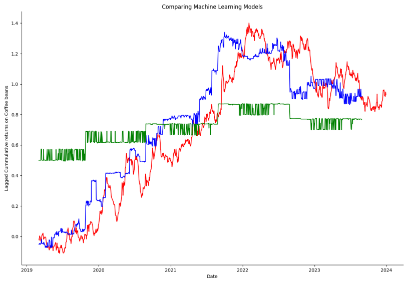

Fig. 7.3. *Models Optimizing for MSE lagged*

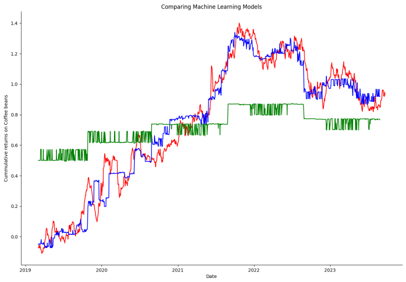

Fig. 7.4. *Models Optimizing for MSE*

These are the hyperparameters used for both Gradient Boosting Regressors and Random Forest Regressor to have best MSE results using Scikit Learn without overfitting.

```

# This is the Pipline for Gradient Boosting Regressor
pipeline = Pipeline([
    (’preprocessor’, preprocessor),
    (’regressor’, GradientBoostingRegressor
    # These are the parameters used to Minimise
    # Mean Squared Error
        learning_rate=0.1,
        max_depth=4,
        min_samples_leaf=2,
        min_samples_split=2,
        n_estimators=50,
    ))
])
pipe =pipeline.fit(X_train, y_train)

###########################################################

# This is the Pipline for Random Forest Regressor
# The preprocessors are standard scaler for numerical data and
# OneHotEncoder for Categorical data
pipeline = Pipeline([
    (’preprocessor’, preprocessor),
    (’regressor’, RandomForestRegressor(
    # These are the parameters used to Minimise
    # Mean Squared Errors
        max_depth=1,
        max_features=’sqrt’,
        min_samples_leaf=1,
        min_samples_split=2,
        n_estimators=150,
    ))
])
pipe = pipeline.fit(X_train, y_train)
```

Listing 2
Processing data

8. Conclusion. In this study, we found that linear regression models, particularly those adjusting for outliers like HC3 and Huber T, are effective in predicting coffee prices using weather patterns, yielding statistically significant results. Further exploration with machine learning models, especially the Gradient Boosting Regressor, after tuning hyperparameters, also proved successful in accurately forecasting coffee prices.

9. Appendix. OpenAI's ChatGPT was used for coding assistance and correcting writing issues

REFERENCES

[1] O. W. Map, *Open weather map*. https://openweathermap.org/, 2024. Accessed on 2024-04-01.

[2] S. Maps, *Simple maps brazil*. https://simplemaps.com/data/br-cities, 2024. Accessed on 2024-04-15.

[3] MeteoStat, *Meteostat*. https://meteostat.net/en/, 2024. Accessed on 2024-04-15.

[4] M. Scott, *Climate coffee*. https://www.climate.gov/news-features/climate-and/climate-coffee, 2015. Accessed on 2024-05-01.

[5] USDA, *World coffee markets and trade*. https://fas.usda.gov/data/production/commodity/0711100, 2023. Accessed on 2024-05-01.

[6] USDA, *World coffee markets and trade*. https://apps.fas.usda.gov/psdonline/circulars/coffee.pdf, 2023. Accessed on 2024-05-01.