Assessing Avocado Pricing Dynamics Utilizing Climate, Transportation Cost, and Macroeconomic Metrics in California

Ken K. Hong

khong18@gatech.edu

Abstract

The rising global demand for avocados, fueled by increasing consumer awareness and social media influence, has led to a need for deeper insights into the factors influencing avocado prices. This study focuses on California, a major avocado-producing and consuming region, to analyze key drivers of avocado price fluctuations. Utilizing a comprehensive dataset encompassing climate conditions, transportation costs, and macroeconomic indicators, this research develops predictive models to assess avocado pricing dynamics. Data sources include climate metrics from key production regions (Fallbrook, California, and Uruapan, Mexico), transportation cost data from the U.S. Energy Information Administration, and economic indicators from the Federal Reserve Economic Data repository. The study applies advanced statistical and machine learning techniques, including regression trees and random forests, to evaluate feature importance and forecast price trends.

Results indicate that organic classification, personal consumption expenditures, and energy costs significantly impact avocado prices, while unemployment rates show minimal influence. The random forest model outperforms the regression tree model, achieving an R^2 of 0.947 for training data and 0.798 for testing data, demonstrating strong predictive accuracy and generalizability. These findings provide valuable insights for stakeholders in the avocado supply chain, offering data-driven guidance for pricing strategies and market forecasting.

Contents

1. Introduction

Avocados have seen a remarkable surge in popularity across social media platforms in recent years, driving an exponential increase in demand and establishing themselves as one of the most popular fruits globally. Research conducted by Mordor Intelligence [MordorIntelligence] forecasts substantial growth in the avocado market, with projected market size expected to rise from USD 22.69 billion in 2024 to USD 35.55 billion by 2029. Additionally, the USDA reports an annual per capita consumption of avocados exceeding 8 pounds in the North American market, with local production and imports serving as the primary sources of supply.

Given the perishable nature of avocados, maintaining fruit quality heavily depends on a robust supply chain that efficiently manages harvesting, transportation, and distribution processes. As avocado demand continues to escalate, understanding the driving factors behind market expansion becomes imperative.

This study concentrates on California, a leading region in both avocado cultivation and consumption, with the specific aim of identifying primary drivers influencing avocado price fluctuations within the state. By analyzing critical components of the avocado supply chain—production, transportation, and consumer behavior—the objective is to develop a robust regression model that not only forecasts future avocado prices but also highlights the importance of each factor influencing these prices. The goal is to achieve a comprehensive understanding of avocado pricing dynamics by incorporating features representing various stages of the supply chain.

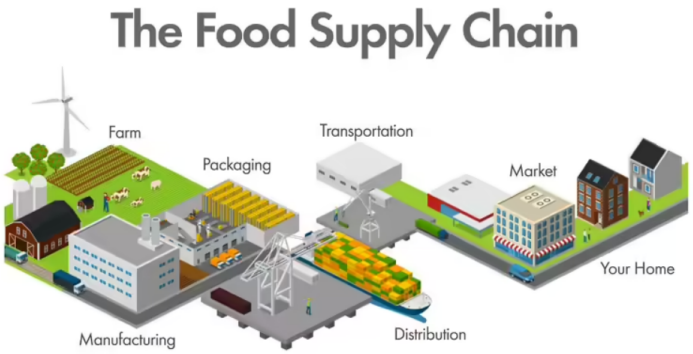

Figure 1 | Food Supply Chain [Kiger]

2. Data Source

2.1. Production Data

Avocado pricing is significantly influenced by climatic conditions in key production regions such as Uruapan, Michoacan (Mexico), and Fallbrook, California (USA), both of which are known as the "Avocado Capital of the World." Climate data sourced from *Weather of the World* (available at `https://rp5.ru/Weather_in_the_world`) covers the period from January 2014 to February 2024. This dataset includes daily information on temperature, precipitation, humidity, wind speed, atmospheric pressure, and other factors critical for avocado cultivation. This data is leveraged to assess the impact of climatic conditions on avocado production.

2.2. Transportation Cost Data

Transportation costs, a major component of avocado pricing, primarily include energy consumption expenses. Unit prices of electricity (in cents per kilowatt-hour), petroleum (gasoline and diesel, in dollars per gallon), and natural gas (dollars per cubic feet) in California are sourced from *U.S. Energy Information Administration (EIA)* (available at https://www.eia.gov/). Incorporating these data allows for an understanding of the influence of transportation costs on avocado pricing dynamics.

2.3. Consumer and Economic Data

Obtained from *the Federal Reserve Economic Data (FRED)* (available at https://fred.stlouisfed.org/), the economic variables include indicators such as the Consumer Price Index, Personal Consumption Expenditures, Producer Price Index, Inflation Rate, and Unemployment Rate. These economic metrics serve as valuable indicators to assess how changes in the economic environment correlate with fluctuations in avocado prices.

2.4. Target Variable

The target variable in the regression model is derived from avocado price data obtained from Kaggle [Kiggins]. This dataset comprises weekly avocado prices and sales volume data spanning from 2015 to 2023 across major U.S. areas. For the purpose of this report, the focus is on weekly California price data, utilizing it for model development and interpretation.

Components	Attribute	Source
Avocado Price Data	Year	Kaggle
	Month	Kaggle
	Is_Organic	Kaggle
	Avocado_Volume	Kaggle
	Avocado_Price	Kaggle
Plantation Data	Temperature_Uruapan	Weather in the World
	Sea_Level_Pressure_Uruapan	Weather in the World
	Loc_Pressure_Uruapan	Weather in the World
	Humidity_Uruapan	Weather in the World
	Wind_Speed_Uruapan	Weather in the World
	Visibility_Uruapan	Weather in the World
	Dew_point_temperature_Uruapan	Weather in the World
	Temperature_Fallbrook	Weather in the World
	Sea_Level_Pressure_Fallbrook	Weather in the World
	Loc_Pressure_Fallbrook	Weather in the World
	Humidity_Fallbrook	Weather in the World
	Wind_Speed_Fallbrook	Weather in the World
	Visibility_Fallbrook	Weather in the World
	Dew_point_temperature_Fallbrook	Weather in the World
California Energy Data	California_Gas_Price	U.S. Energy Information Administration
	California_NG_Price (Natural Gas)	U.S. Energy Information Administration
	California_Electricity_Price	U.S. Energy Information Administration
Economic Data	Unemployment_Level	Federal Reserve Economic Data (FRED)
	Unemployment_Rate	Federal Reserve Economic Data (FRED)
	Median_Consumer_Price_Index	Federal Reserve Economic Data (FRED)
	Personal_Consumption_Expenditures	Federal Reserve Economic Data (FRED)
	Number_Unemployed_for_27_Weeks_and_over	Federal Reserve Economic Data (FRED)
	Average_Hourly_Earnings_of_All_Employees	Federal Reserve Economic Data (FRED)
	Federal_Funds_Effective_Rate	Federal Reserve Economic Data (FRED)
	Employed_Persons_in_California	Federal Reserve Economic Data (FRED)
	Labor_Force_Participation_Rate_for_California	Federal Reserve Economic Data (FRED)

Table 1 | Dataset Components and Sources

3. Data Examinations

As outlined in the Data Source section, a thorough analysis will be conducted to assess the impact of various features on avocado price fluctuations. This examination begins with identifying missing values or outliers within the dataset. Given the variability in data frequency and granularity, each feature will be aggregated into monthly data points, and monthly averages will be computed for model development purposes.

During this examination, each attribute from different components of the supply chain (plantation, energy consumption, economic factors) will be scrutinized. The primary objectives include:

- Reviewing outliers and identifying data anomalies.
- Checking correlations between different attributes to understand interdependencies.

The results of this examination will provide insights critical for developing a comprehensive understanding of the data sources and refining feature selection for model development.

3.1. Avocado Price Data

Data Description

	count	mean	std	min	25%	50%	75%	max
Year	216.000000	2019.000000	2.587987	2015.000000	2017.000000	2019.000000	2021.000000	2023.000000
Month	216.000000	6.500000	3.460071	1.000000	3.750000	6.500000	9.250000	12.000000
Is_Organic	216.000000	0.500000	0.501161	0.000000	0.000000	0.500000	1.000000	1.000000
Avocado_Volume	216.000000	3036990.358403	2852572.101494	82368.292500	222202.941875	2169131.171000	5841954.779500	7653057.877500
Avocado_Price	216.000000	1.509203	0.355675	0.825000	1.223255	1.481788	1.790987	2.325000

Figure 2 | Data Description

Avocado Price over Time

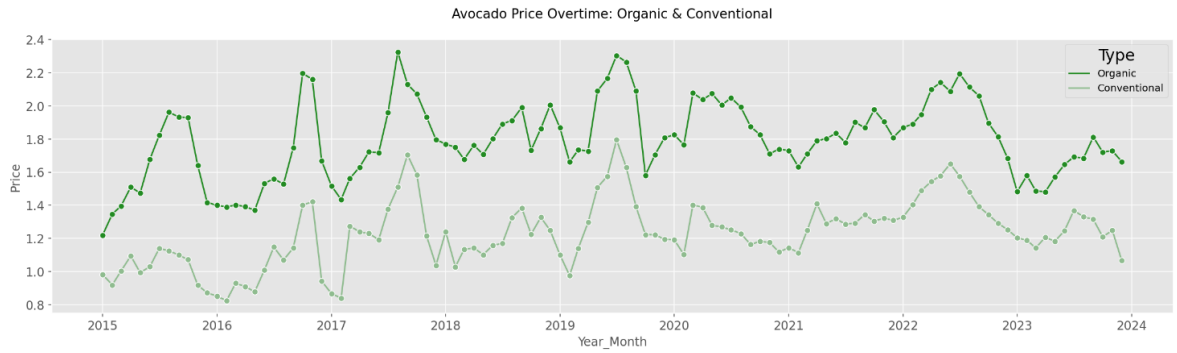

Figure 3 | Avocado Price over Time

3.2. Plantation Data - Fallbrook

Data Description

	count	mean	std	min	25%	50%	75%	max
Temperature_Fallbrook	216.000000	17.055303	3.596314	9.760880	14.216468	16.461810	20.562465	23.619718
Sea_Level_Pressure_Fallbrook	216.000000	742.658457	1.427656	740.314450	741.386022	742.240842	743.934006	745.576340
Humidity_Fallbrook	216.000000	66.793920	10.396502	31.916667	61.128533	69.056502	74.600565	85.405244
Wind_Speed_Fallbrook	216.000000	1.857397	0.312388	1.186298	1.617894	1.809621	2.091785	2.660128
Visibility_Fallbrook	216.000000	14.792203	0.627279	13.053333	14.414525	14.882023	15.236255	15.897059
Dew_point_temperature_Fallbrook	216.000000	9.192613	5.230041	-4.534498	4.933694	8.579012	13.898736	16.815232

Figure 4 | Data Description

KDE of Each Attribute

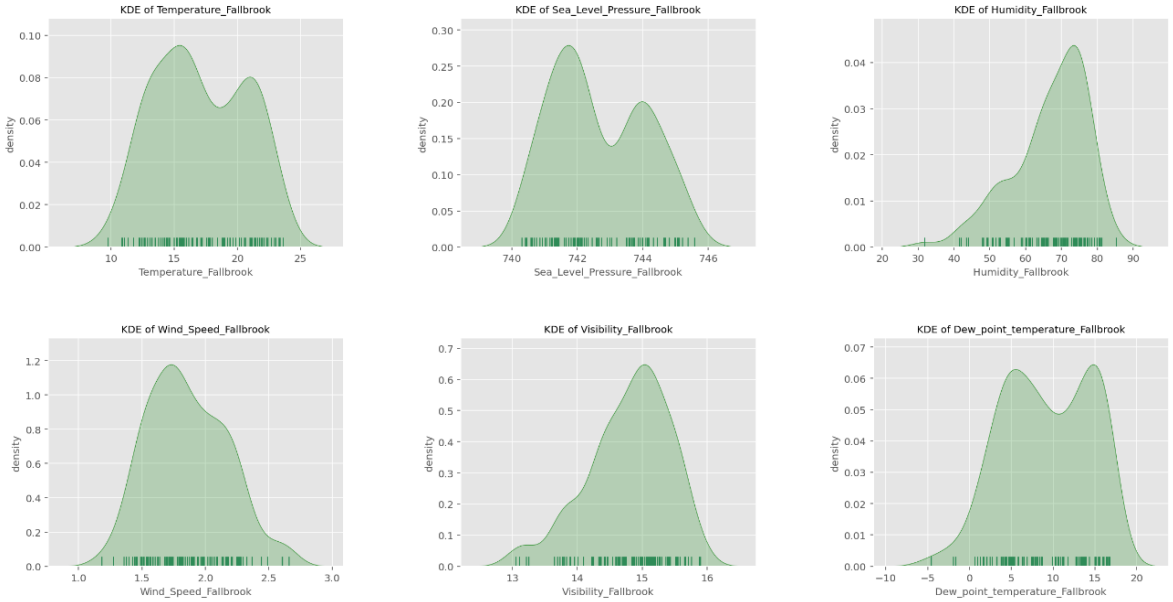

Figure 5 | KDE of Each Attribute

Correlation Between Attributes

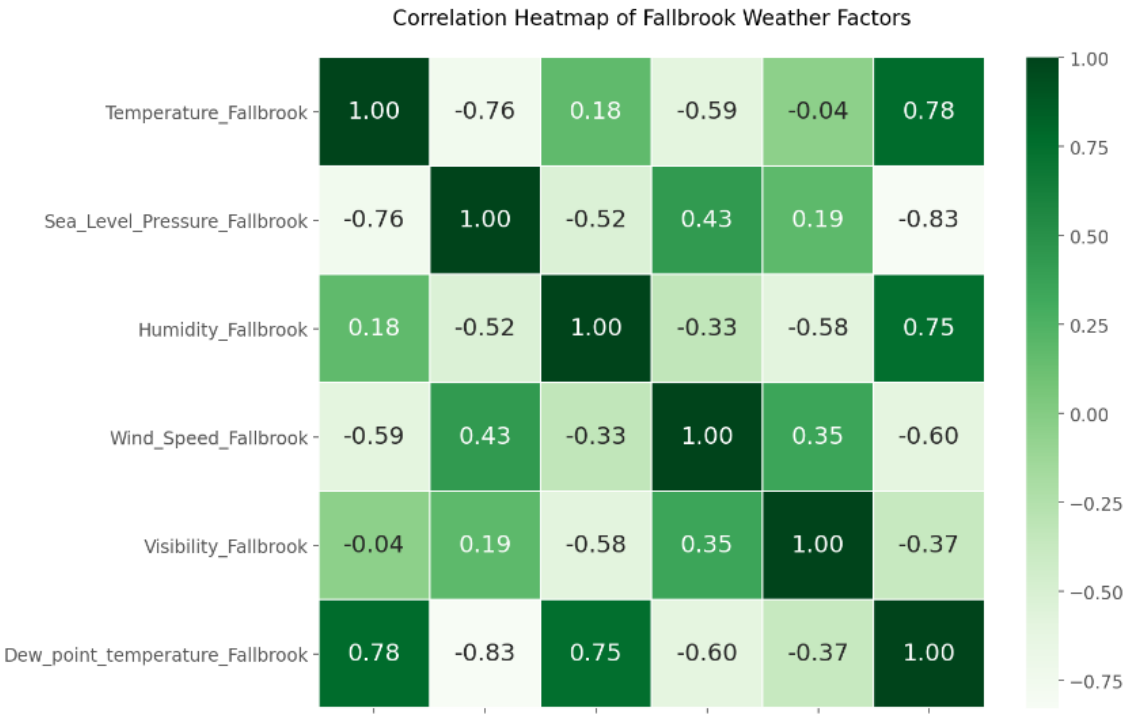

Figure 6 | Correlation Between Attributes

Pairwise Distribution of Attributes

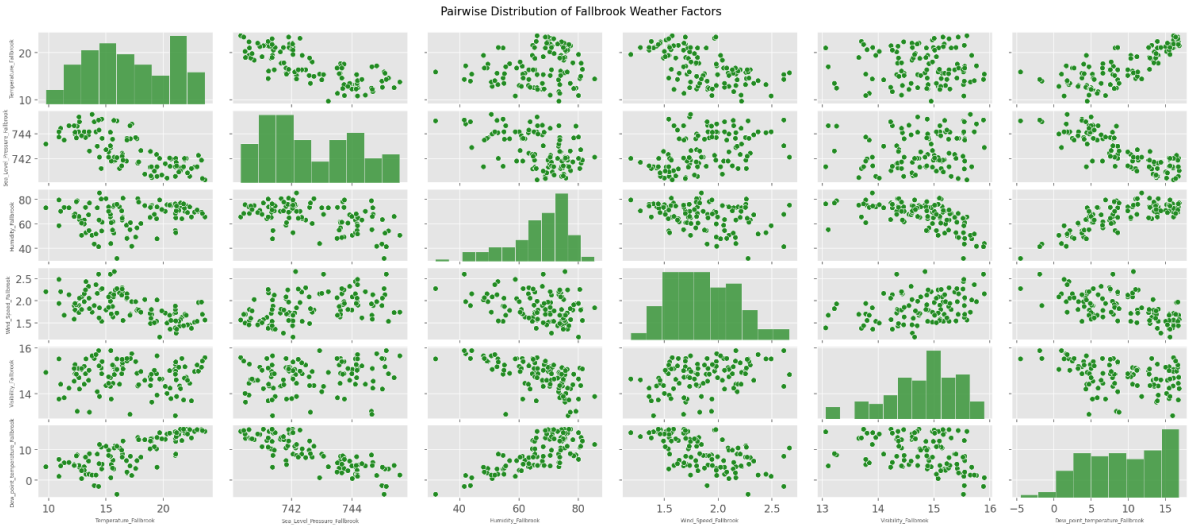

Figure 7 | Pairwise Distribution of Attributes of Fallbrook Weather Factors

3.3. Plantation Data - Uruapan

Data Description

	count	mean	std	min	25%	50%	75%	max
Temperature_Uruapan	216.000000	20.603093	1.415372	17.280303	19.741589	20.443585	21.600767	24.635443
Sea_Level_Pressure_Uruapan	216.000000	635.813458	0.572428	634.337121	635.451312	635.816259	636.230783	637.113317
Humidity_Uruapan	216.000000	67.251374	13.404664	36.744063	57.362114	68.695882	79.500787	85.725191
Wind_Speed_Uruapan	216.000000	1.799516	0.314965	0.967662	1.561724	1.754706	2.009551	2.506596
Visibility_Uruapan	216.000000	13.408285	2.519041	7.799501	11.546415	13.192788	14.922268	20.525341
Dew_point_temperature_Uruapan	216.000000	13.461364	3.381558	5.142480	10.256261	13.792843	16.698894	17.846154

Figure 8 | Data Description

KDE of Each Attribute

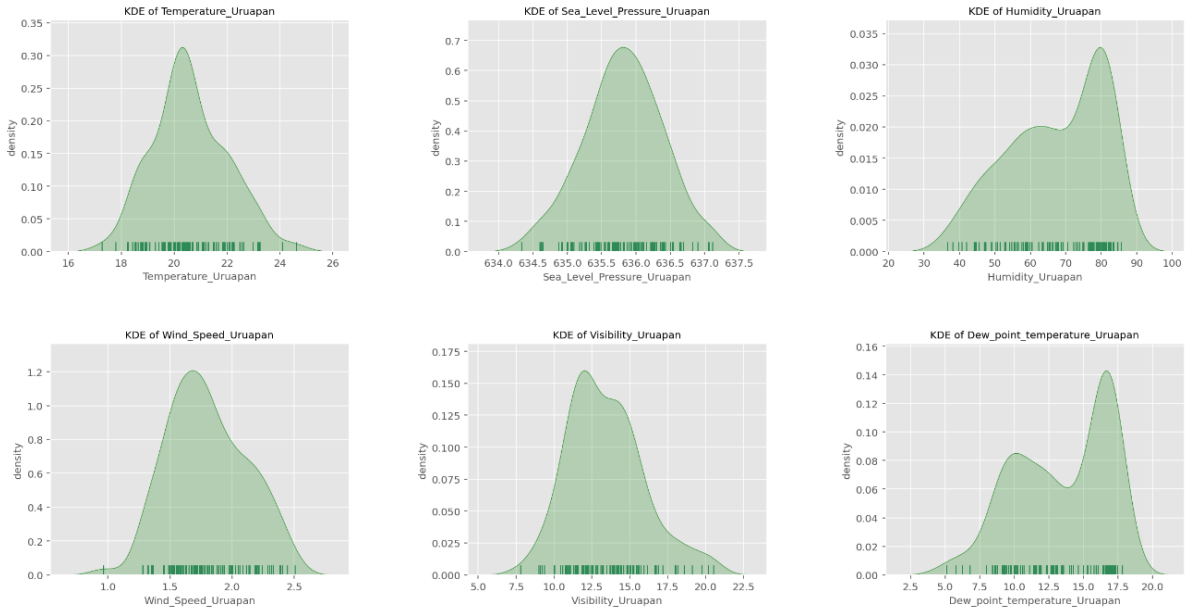

Figure 9 | KDE of Each Attribute

Correlation Between Attributes

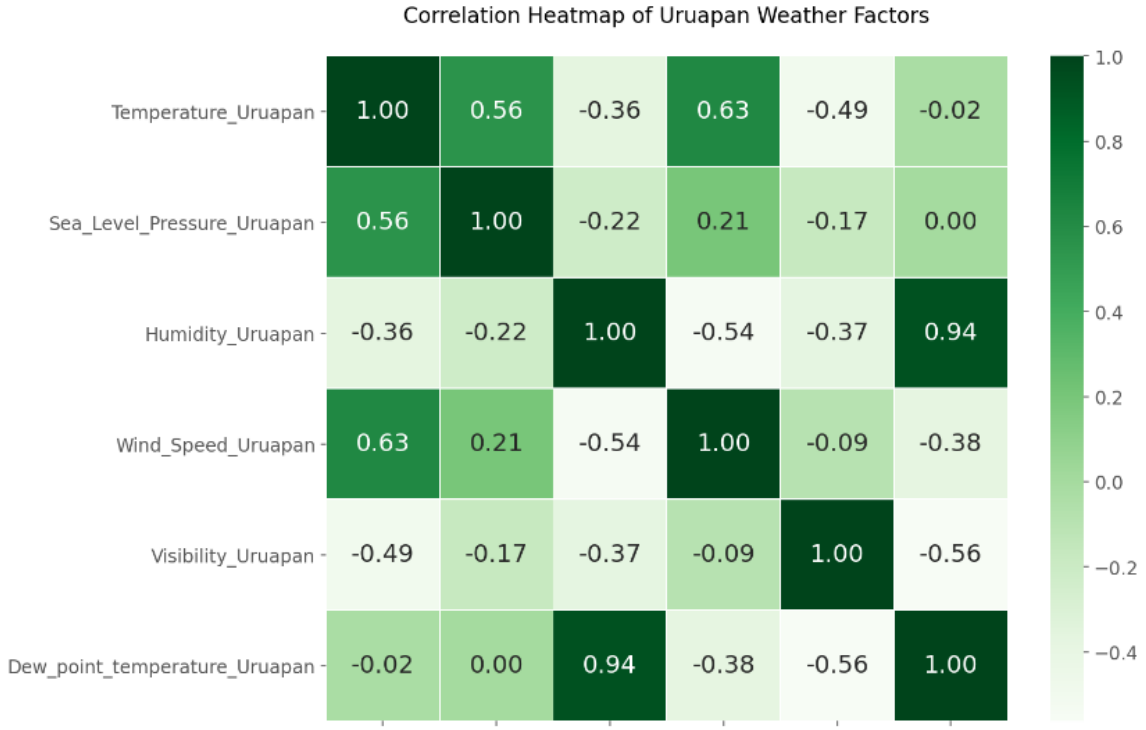

Figure 10 | Correlation Between Attributes

Pairwise Distribution of Attributes

Figure 11 | Pairwise Distribution of Attributes of Uruapan Weather Factors

3.4. California Energy Data

Data Description

	count	mean	std	min	25%	50%	75%	max
California_Gas_Price	216.000000	3.757833	0.894529	2.477000	3.095750	3.555500	4.331250	6.294000
California_NG_Price	216.000000	11.070895	3.149315	0.000000	9.085833	9.953333	12.345833	22.953333
California_Electricity_Price	216.000000	18.180000	3.393409	12.600000	15.497500	17.310000	20.007500	27.740000

Figure 12 | Data Description

KDE of Each Attribute

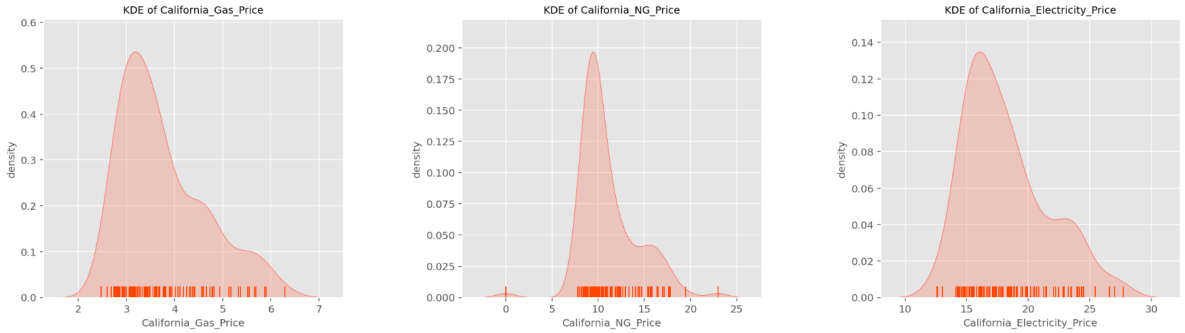

Figure 13 | KDE of Each Attribute

Correlation Between Attributes

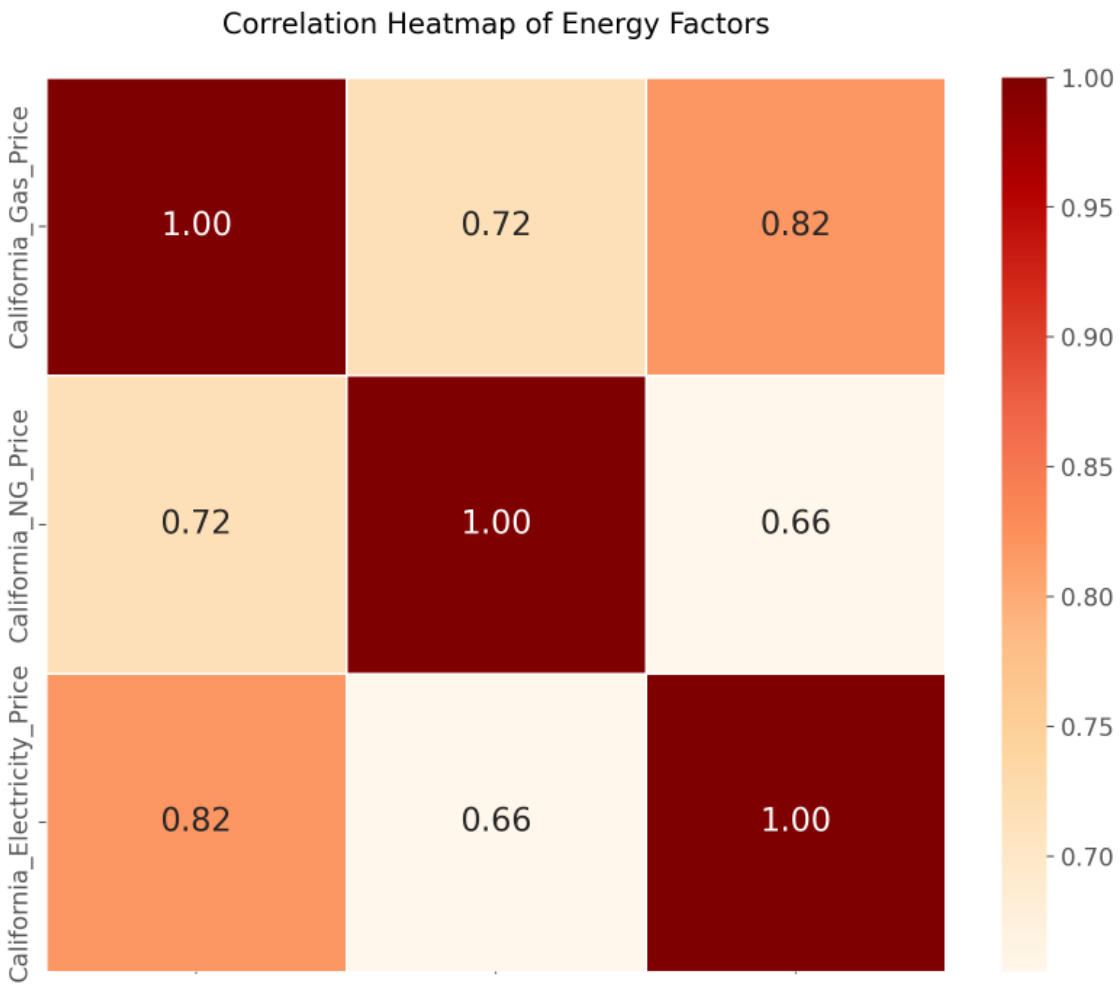

Figure 14 | Correlation Between Attributes

Pairwise Distribution of Attributes

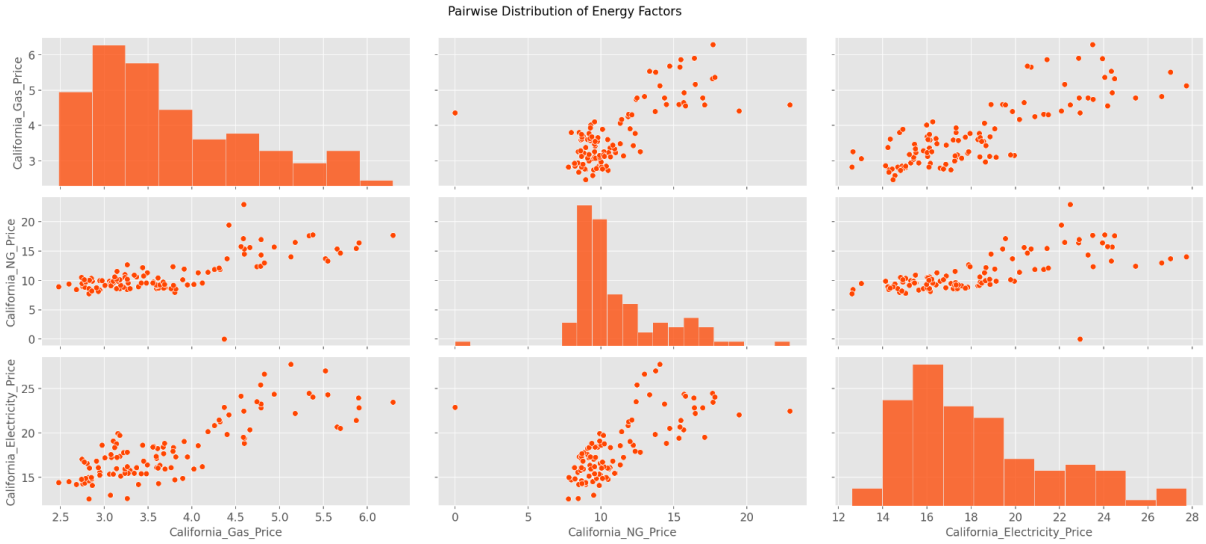

Figure 15 | Pairwise Distribution of Attributes of California Energy Factors

3.5. Economic Data

Data Description

	count	mean	std	min	25%	50%	75%	max
Unemployment_Level	216.000000	7666.101852	2812.828794	5698.000000	6070.000000	6753.500000	7968.750000	23090.000000
Unemployment_Rate	216.000000	4.752778	1.808068	3.400000	3.700000	4.200000	5.025000	14.800000
Median_Consumer_Price_Index	216.000000	3.436477	1.710645	1.092895	2.293325	2.897937	4.134864	8.048036
Personal_Consumption_Expenditures	216.000000	14777.619444	2070.833871	12066.700000	13127.025000	14201.500000	16452.625000	19013.700000
Number_Unemployed_for_27_Weeks_&_over	216.000000	1828.305556	808.172057	965.000000	1262.250000	1553.500000	2127.000000	4174.000000
Average_Hourly_Earnings_of_All_Employees	216.000000	28.668981	2.887263	24.750000	26.155000	27.985000	31.000000	34.340000
Federal_Funds_Effective_Rate	216.000000	1.409167	1.572856	0.050000	0.120000	0.845000	2.145000	5.330000
Employed_Persons_in_California	216.000000	18030973.333333	602621.626383	15643297.000000	17752987.000000	18286905.500000	18419367.500000	18732265.000000
Labor_Force_Participation_Rate_for_California	216.000000	61.912037	0.665945	59.600000	61.800000	62.100000	62.300000	63.000000

Figure 16 | Data Description

KDE of Each Attribute

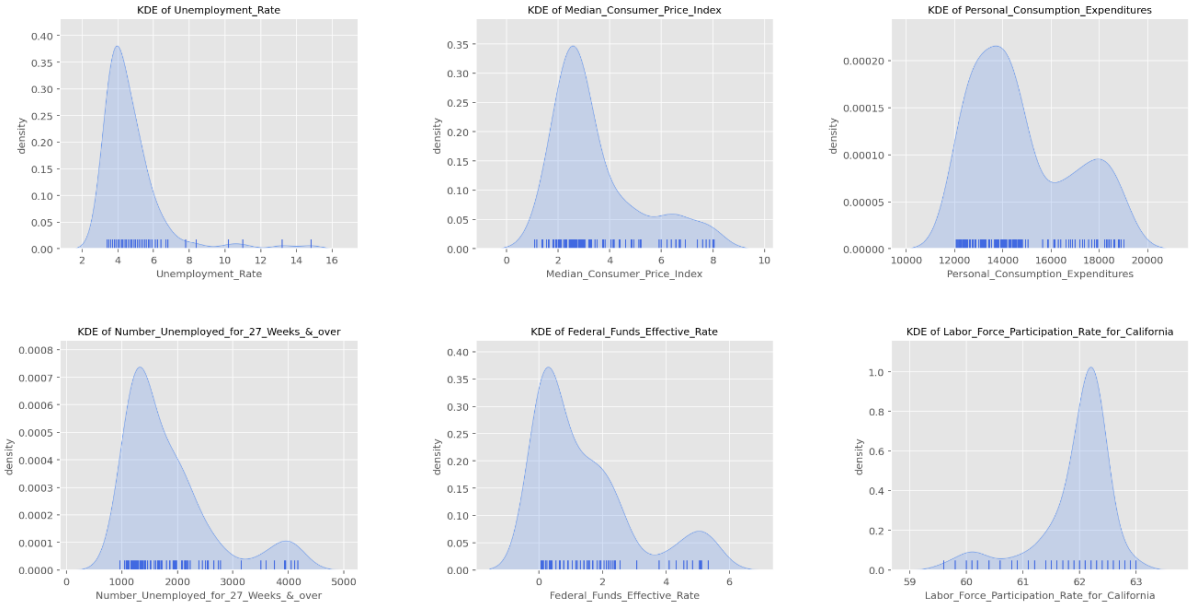

Figure 17 | KDE of Each Attribute (Highly Correlated Attributes Removed)

Correlation Between Attributes

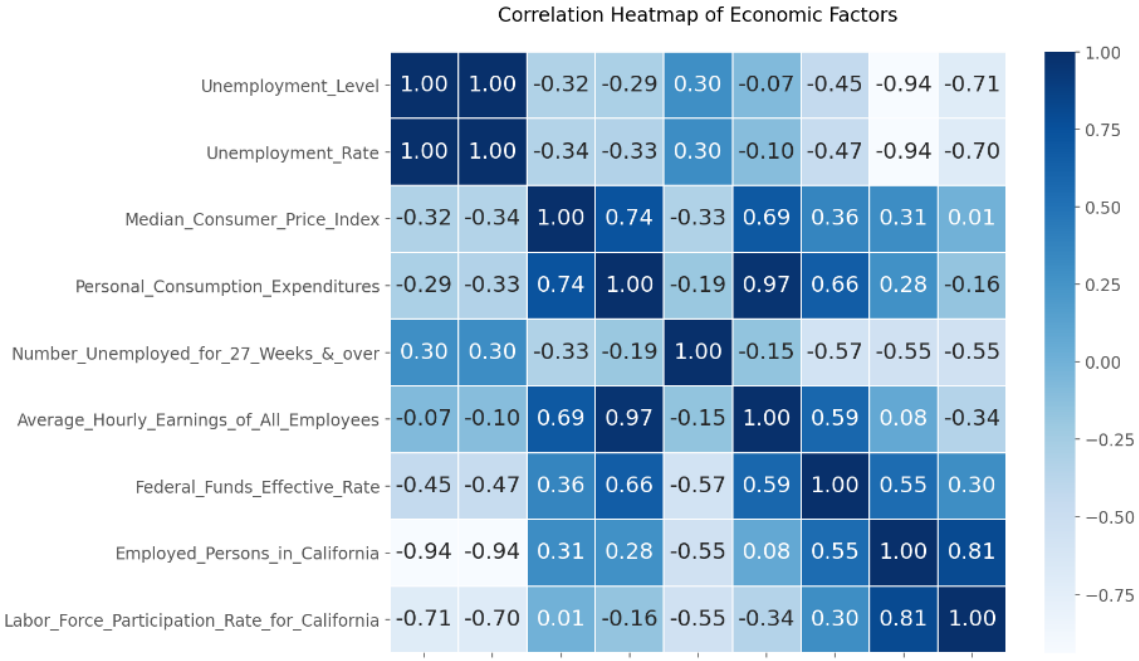

Figure 18 | Correlation Between Attributes

Pairwise Distribution of Attributes

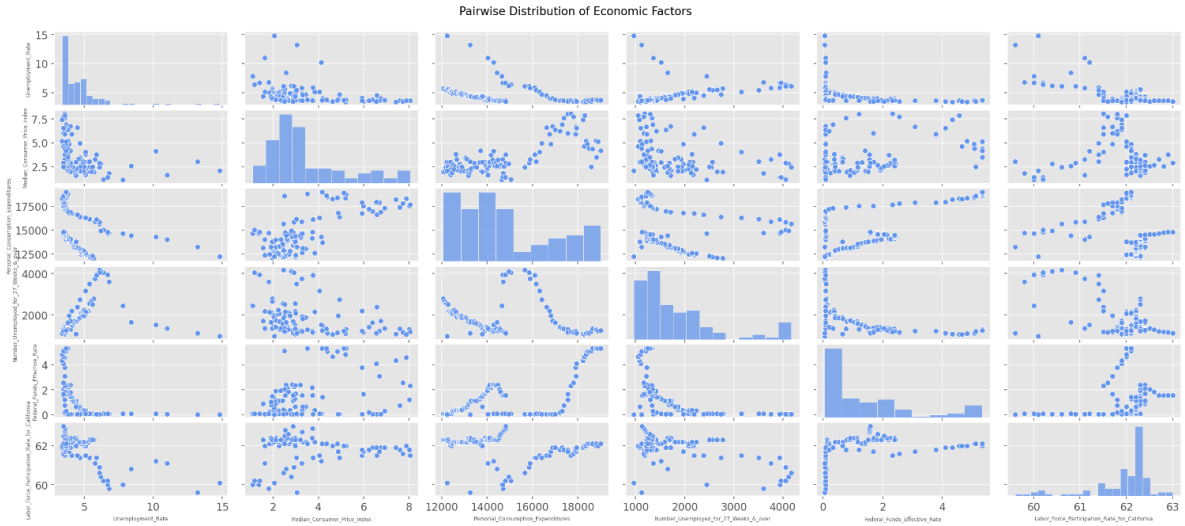

Figure 19 | Pairwise Distribution of Attributes of Economic Factors

3.6. Summary of Data Review

The analysis of 216 monthly avocado price points spanning from January 2015 to December 2023, encompassing both organic and conventional avocados, yielded the following insights:

1. **Data Completeness**: No null values were found across all fields within the dataset.
2. **Price Stationarity**: Visual inspection of price point distribution plots suggests stationarity. Augmented Dickey-Fuller (ADF) Test results with a p-value of approximately 0.028, below the chosen significance level of 0.05, reject the null hypothesis, indicating that the price data is indeed stationary.
3. **Plantation Data Evaluation**: Evaluation plots of plantation data from Fallbrook and Uruapan, assessed using Kernel Density Estimation (KDE) plots with automatic bandwidth selection, indicate normally distributed attributes without any outliers.
4. **Energy Data Evaluation**: Similar KDE evaluation of energy data in California reveals normally distributed attributes with no identified outliers.
5. **Economic Data Correlation**: Analysis of economic data using an attribute correlation heatmap reveals significant correlations between certain variables. Notably, `Unemployment_Level` correlates highly with `Unemployment_Rate`; `Average_Hourly_Earnings_of_All_Employees` correlates with `Personal_Consumption_Expenditures`; and `Employed_Persons_in_California` correlates with `Labor_Force_Participation_Rate_for_California`. Consequently, `Unemployment_Level`, `Average_Hourly_Earnings_of_All_Employees`, and `Employed_Persons_in_California` are excluded from the model development process.
6. **Economic Factors Manipulation**: Economic factors are adjusted into leading indicators, implying that changes in economic factors may lag behind changes in avocado prices. For example, a rise in unemployment may not immediately impact avocado prices within the same month but could affect the price changes two months later. Considering this lag effect, each avocado price point incorporates economic factors not only from the month of avocado pricing but also from two months prior.

4. Methodology and Model Development

4.1. Multiple Linear Regression

The most commonly employed technique for estimating avocado prices based on input features is the linear regression model. This model represents a linear combination of features and their respective coefficients. Advanced techniques in linear regression may involve mapping input attributes to higher-dimensional feature spaces and incorporating interaction terms to capture synergy effects. Additionally, local data fitting methods, such as locally weighted regression, can enhance model accuracy.

In the development of a linear regression model, selecting the most important features is crucial. Three commonly used feature selection techniques in linear regression are Forward Selection, Backward Elimination, and Recursive Feature Elimination. Below are pseudocode descriptions of each technique:

- **Forward Selection:**
  - Begins with an empty set of selected features.
  - Iteratively selects the best feature to add to the selected set based on improvements in model performance.
  - Continues this process until the desired number of features (k) is selected.
- **Backward Elimination:**
  - Starts with all features included in the selected_features set.
  - Iteratively removes the worst-performing feature based on degradation in model performance.
  - Continues until the number of selected features reaches the minimum desired number.
- **Recursive Feature Elimination:**
  - Iterates through each remaining feature.
  - Adds the feature to the selected features.
  - Trains the model using the selected features and X_train.
  - Evaluates the model performance using a chosen metric.
  - If the model performance improves, updates the best score and best feature.
  - Backtracks by removing the feature to explore other features.

In this study, linear regression with feature elimination is not considered due to the simplicity of the linear regression model. While linear regression offers good interpretability compared to other regression models, its simplicity may not effectively model the complex nonlinear relationships inherent in avocado prices. Instead, regression trees and random forest models will be developed using 80% of the input data as training points. The remaining 20% of the data will be used to evaluate the performance of the regression tree and random forest models.

4.2. Regression Tree

A *regression tree* involves recursively partitioning the feature space into smaller regions based on the values of input features. At each node of the tree, the algorithm selects the feature and split point that best separates the data into groups in terms of the target variable. This splitting process continues until a stopping criterion is met, such as reaching a maximum tree depth or a minimum number of samples per node. While regression trees are intuitive and capable of

capturing complex nonlinear relationships between input features and the target variable, they are prone to overfitting and may not perform well with unseen data.

In this study, the stopping criterion for the regression tree was determined using five-fold cross-validation. The goal was to identify the optimal minimum number of samples per node. The results of the cross-validation, evaluated in terms of R^2 (coefficient of determination), are presented below:

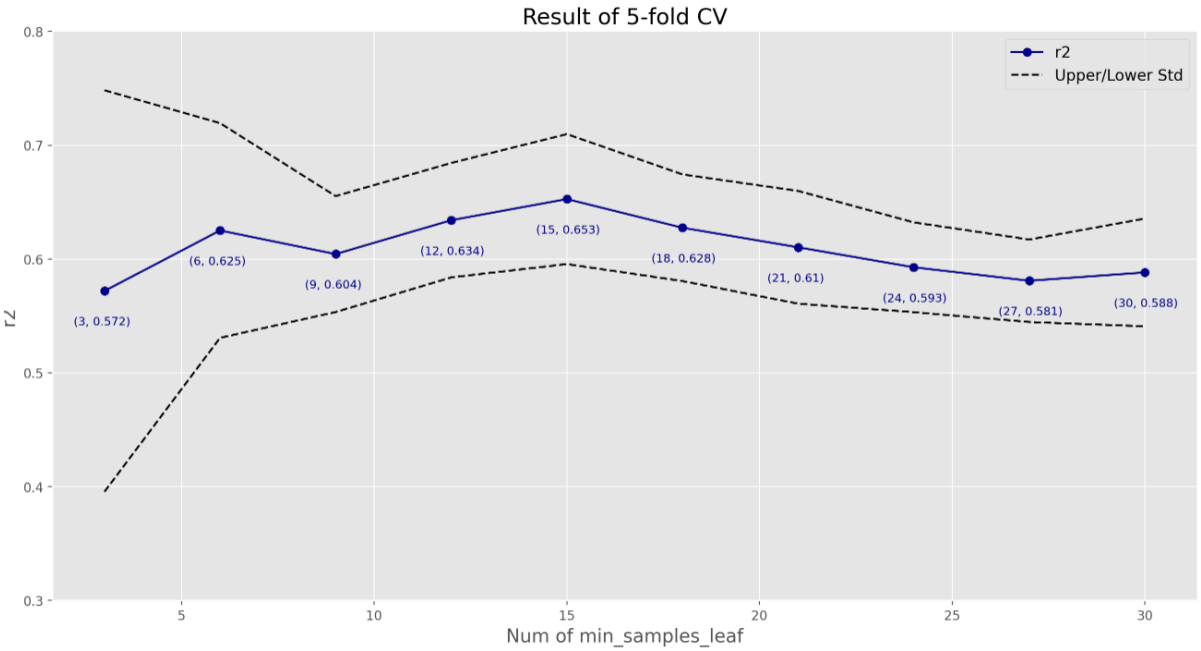

Figure 20 | Regression Tree: Results of 5-fold Cross-Validation

Based on the cross-validation results, a minimum number of 15 samples per node is recommended. Using this parameter, the regression tree model is implemented and will be evaluated for performance in the final result evaluation section.

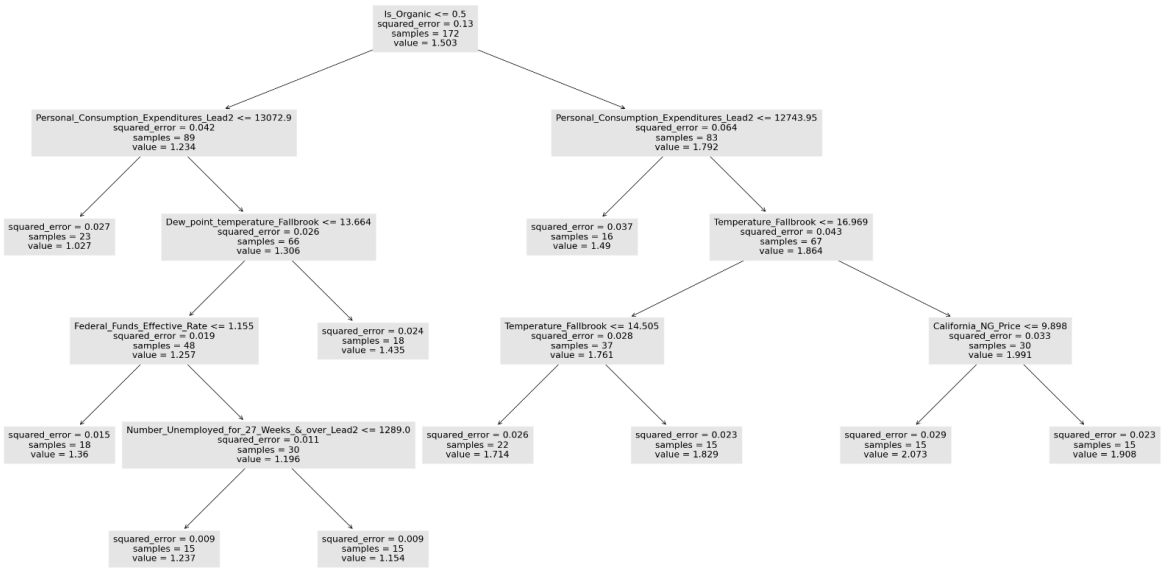

Figure 21 | Classification And Regression Tree (CART)

4.3. Random Forest

To address the risk of overfitting associated with a single CART tree, an ensemble method known as *Random Forest* is employed. This technique constructs multiple decision trees during training by utilizing bootstrapped samples with replacement. Each tree is trained on a random subset of the bootstrap sample and a random subset of features at each split. The objective is to decorrelate the trees and mitigate overfitting. Random Forests exhibit robustness against outliers, are less sensitive to noise in the data, and provide a measure of feature importance.

In this study, the performance of the Random Forest model is evaluated using the grid search cross-validation method. Two key parameters, the number of trees (`n_estimators`) and the minimum number of samples per leaf node (`min_samples_leaf`), are tuned to optimize the model's performance. The results of the grid search cross-validation are presented below, illustrating the impact of parameter tuning on the model's effectiveness.

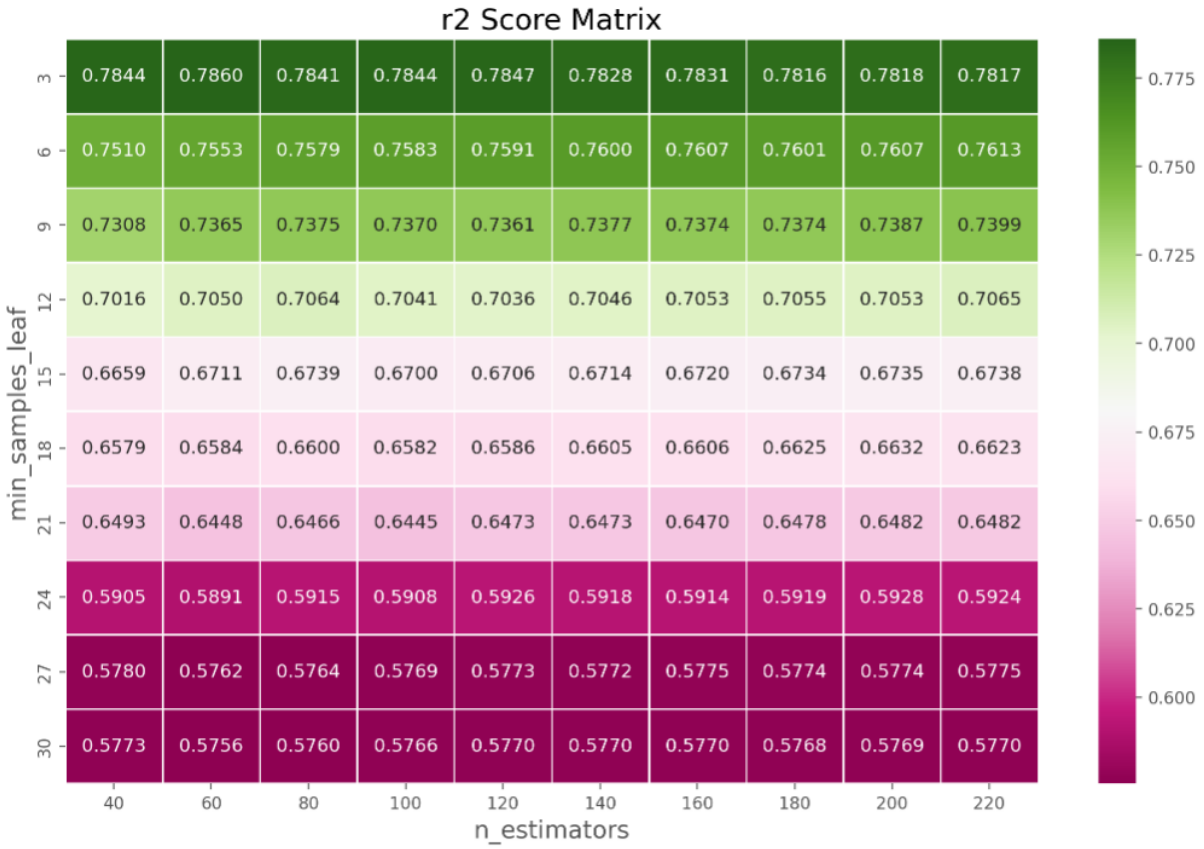

Figure 22 | Random Forest: R^2 Score Matrix

Based on the result matrix from grid search cross-validation, the combination that yielded the best performance consists of `min_samples_leaf = 3` and `n_estimators = 60`. Using this optimal parameter combination, the Random Forest regression model is developed to predict avocado prices. The performance of this model will be thoroughly evaluated and discussed in the final results section.

5. Results Evaluation

The performance of the developed models will be evaluated using the 20% evaluation dataset obtained from the previous step. The primary metric for assessing model performance will be the R^2 score, also known as the coefficient of determination.

Definition of R^2: R^2 is a statistical measure that quantifies the proportion of variance in the dependent variable explained by the independent variables in a regression model. It ranges from 0 to 1, where:

- $R^2 = 1$ indicates a perfect fit, meaning the model explains all the variability in the target variable around its mean.
- $R^2 = 0$ suggests that the model does not explain any variability and merely predicts the mean of the target variable for all observations.

A higher R^2 score indicates a better fit of the model to the data, capturing more variance in the target variable. R^2 can be calculated using the equation below:

$$R^2 = 1 - \frac{\text{Sum of Squared Residuals}}{\text{Total Sum of Squares}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

Regression Tree Evaluation: To determine the predictions of the regression tree ($\hat{y}$) for the 20% testing dataset, each testing point undergoes evaluation through the rules of the regression tree, starting from the root node down to a leaf node. The predicted value is derived from the mean of the target variable in the leaf node's training samples.

Upon evaluating the predicted prices, the training R^2 is calculated using the formula described earlier and is computed as 0.827, indicating a high level of explanation of variance within the training data. However, when applying this regression tree model to the testing dataset, the resulting R^2 is calculated as 0.589, which is 0.238 lower than the training R^2. This discrepancy suggests that the regression tree model's generalizability to unseen data is limited, potentially due to the smaller amount of data available for training.

Random Forest Model Evaluation: For the random forest model, it is constructed based on the results of the grid search cross-validation, utilizing the parameters: 3 for `min_samples_leaf` and 60 for `n_estimators`. Once the 60 random forest trees are built using the bagging technique, new predictions for the data are estimated based on the mean of these 60 trees.

The training R^2 for the random forest model is calculated as 0.947, and the testing R^2 is 0.798, which is 0.149 lower than the training R^2. This performance suggests that the random forest model not only excels in the training data but also exhibits better generalizability to unseen data. This advantage may be attributed to the characteristics of random forests, which aim to reduce overfitting by averaging the performance of multiple trees developed with the goal of decorrelating properties.

Feature Importance Analysis: Now that the R^2 values for both the regression tree and random forest models have been determined, further investigation into the key features driving these R^2 values is conducted through feature importance analysis. This analysis is performed to discover the most critical features for both the regression tree and random forest models. The calculation of feature importance is relatively straightforward in this context, focusing on the change in R^2 when specific features are included, excluded, set to zero, or shuffled to maintain the overall probability of each feature.

The results are shown below:

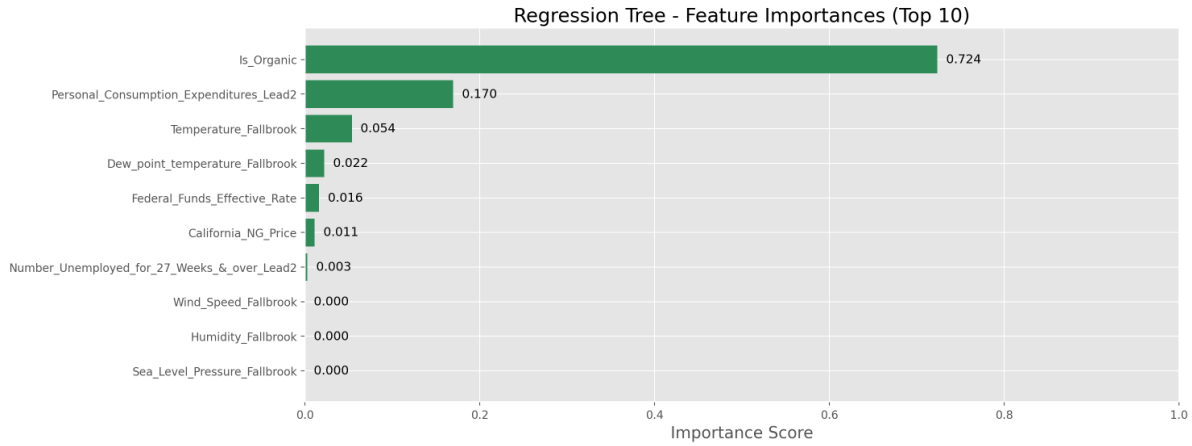

Figure 23 | Regression Tree - Feature Importance (Top 10)

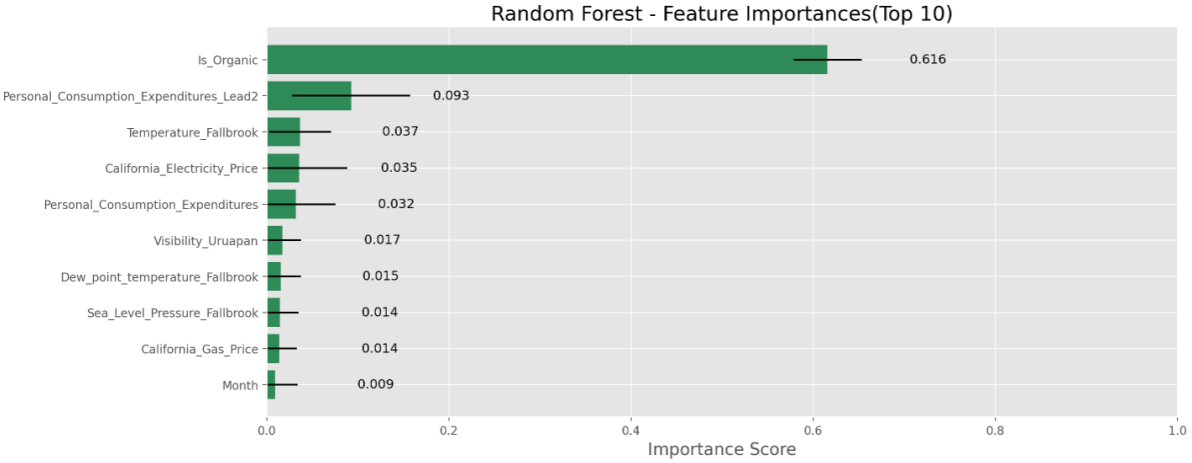

Figure 24 | Random Forest - Feature Importance (Top 10)

Based on the plots, it is evident that knowing whether the avocado is organic or conventional plays a crucial role in estimating its price. This aligns with the expectation that organic food typically has a higher price compared to conventional options. The second important feature is prior personal consumption expenditure, indicating that changes in personal spending may not immediately impact avocado prices but could influence them two months later. Additionally, weather conditions in Fallbrook, California, particularly temperature changes, have a minor effect on avocado prices due to the fruit's perishable nature, which depends on freshness. Weather conditions in Fallbrook may also correlate with broader California weather patterns, potentially affecting avocado prices. The price of electricity in California is another significant feature, as higher electricity prices likely contribute to higher avocado prices. Personal consumption expenditures for the month rank as the fifth important feature, although less significant compared to earlier conditions, as discussed.

6. Conclusion

In conclusion, this report evaluated the performance of two regression models: Regression Tree and Random Forest. The Random Forest model achieved acceptable accuracy with an R^2 of 0.947 in training and 0.798 in testing, demonstrating higher generalizability and less overfitting to the training data. While Random Forest outperformed the Regression Tree, it's important to note that Random Forest lacks the interpretability to explain how the result is determined in a simple form. In scenarios where interpretability is a key criterion for model selection, Regression Tree or other regression models should be considered.

Furthermore, the feature importance analysis of the Random Forest model revealed critical factors contributing to R^2. Notably:

- Organic avocados have a higher price.
- Changes in personal consumption expenditure impact avocado prices two months later.
- Electricity prices also influence avocado pricing.

Understanding these important features and leveraging the Random Forest model allows for the estimation of future avocado prices. This methodology can be extended to comprehensively analyze food prices in California, providing insights into the drivers of price changes.

Surprisingly, unemployment rates do not significantly impact avocado price predictions, possibly indicating that avocados remain affordable even for unemployed populations. Overall, leveraging these important features and the Random Forest model enables the estimation of future avocado prices, with potential for broader analysis of food price dynamics across California.

References

P. J. Kiger. Supply chain 101: What happens when our food supply is disrupted by a pandemic?, May 11 2020. URL https://money.howstuffworks.com/food-supply-chain-pandemic.htm.

J. Kiggins. Avocado prices and sales volume 2015-2023, January 12 2024. URL https://www.kaggle.com/datasets/vakhariapujan/avocado-prices-and-sales-volume-2015-2023.

MordorIntelligence. Avocado market insights, 2024. URL https://www.mordorintelligence.com/industry-reports/avocado-market.