**Strategic Analysis of Fuel Prices in Brazil Using Machine Learning: Trends, Inequities, and Predictive Modelling**

**Table of contents**

* + 1. Abstract ……………………………………………………………......……….3
    2. Introduction …………………………………………………………………….3
    3. Background and context of the Irish Housing Market …………………………3
    4. Major challenges faced by the Irish Housing Market and possible solutions…..3
    5. Scope………………………………………………....…………………………4
    6. Working with dataset……………………………………………………………5
    7. Exploratory Data Analysis………………………………………………………5
    8. Data Preparation…………………………………………………………...……6
    9. Machine Learning Model………………………………………………………10
    10. Cross-validation and Hyperparameter tuning………………………………….15
    11. Feature Importances……………………………………………………..……..17
    12. Conclusion …………………….……………………...……………………….19
    13. References and sources…………………………………………..……………20

**Table of figures and tables**

Figure 1. Boxplot of variable “propertySize”……………………………..……………….7

Figure 2. Pie chart of properties sales by county…………………………………………..8

Figure 3. Properties sold in each county…………………………………………..……….9

Figure 4. The Corelation Heatmap………………………………………………….……..10

Figure 5. Distribution of house bedrooms……………………………………..………….12

Figure 6. Table of metrics calculated by each machine learning model…………………..14

Figure 7. Plotting actual vs predicted values for different models……………………..…15

Figure 8. Table of cross-validation scores using different amount of k-folds…………….16

Figure 9. Table of metrics results calculated by Random Forest before and after hyperparameter tuning……………………………………………….……………………………17

Figure 10. Barchart of 10 most important features………………………………………….18

Figure 11. Barchart of 10 least important features……………………………...…………..19

1. **Abstract**

Fuel price volatility and geographical unevenness in Brazil represent significant threats to economic stability and strategic decision-making. This project examines past fuel price data from the National Agency of Petroleum (ANP) to identify trends, disparities, and possibilities for forecasting. Through data science and machine learning techniques, we aim to create models capable of forecasting average fuel resale prices based on temporal, geographic, and product features. The results provide actionable insight to government and industry leaders, informing policy that enhances market integrity, operational efficiency, and data transparency.

1. **Introduction**

Fuel prices have been a central point of interest in Brazil’s economy and politics. Differences in prices between states and states, and sporadic increases due to inflation, tax adjustments, and supply disruptions, have a direct impact on customers, businesses, and government policy.

As fuel is a central input in logistics, transport, and daily life, an understanding of its price behaviour is important to strategic decision-making. Price prediction is also useful for government agencies to better regulate and plan subsidies, and for private companies to optimize supply chain and pricing strategies.

The objective of this project is to analyse the determinants and trends of fuel prices across Brazil. With the combination of statistical modelling and machine learning techniques, we aim to develop models that predict average resale price and derive strategic insights from both temporal and regional perspectives.

1. **Problem Definition**

Brazilian fuel prices are influenced by a complex mix of domestic taxes, local infrastructure, distribution, and international economic forces. The mix translates into significant price fluctuations not only over time but also geographically and state-to-state.

The project addresses the following key questions:

* What are trends in regional and temporal fuel prices in Brazil?
* Are there seasonality or repeat pattern which would be amenable to be used as predictor?
* How accurately can machine learning models predict mean resale fuel prices based on available features?
* Does the data mis-represent any regions or fuel types, and how does this affect model fairness?

By addressing these questions, we seek to contribute to a more transparent, data-driven understanding of fuel price behaviour in Brazil, supporting strategic planning and policy development.

1. **Challenges in Fuel Price Regulation and Forecasting in Brazil: Causes, Impacts, and Data-Driven Solutions**

Fuel prices in Brazil are influenced by an array of variables that range from international volatility of oil prices to domestic infrastructure and taxation. The interlocking variables render the regulation and prediction of fuel prices an extremely challenging undertaking. Understanding and appreciating the structural and economic constraints involved is essential for putting the need for predictive models and data-based planning into perspective.

One of the principal problems is regional imbalance in distribution network and transportation costs. The continental nature of Brazil means that fuel must be transported over vast distances, especially to serve northern and interior states. These transportation costs are not evenly distributed across the country and often include additional risks or delays because of poor infrastructure, leading to consistently higher prices in certain regions.

Another major issue is taxation, namely the difference in state-level taxation. Both fuel states have their own rates(e.g., ICMS), which contribute to considerably to final consumer prices. This is great price disparities between neighbouring states and makes nationwide pricing schemes or subsidies challenging.

In addition, the dependence on global oil prices and exchange rates makes fuel prices in Brazil extremely vulnerable to general economic conditions worldwide. Because some of Brazil’s refined fuels are imported and are subject to dollar rate volatility, it makes the domestic market vulnerable to price shock worldwide that is, far too frequently leading to spur-of-moment pump adjustments.

Another operation problem is the variability in data reporting. States report different numbers of surveyed gas station, from a few to hundreds, and sparse data, resulting in biased data quality and quantity. This not only decreases confidence in the national average but can also lead to model and policy bias.

Policy uncertainty makes the situation even worse. Fuel subsidies, price controls, and export and import bans have the tendency to change rapidly with the whims of political convenience. These changes introduce idiosyncrasies into the data that are data that are difficult to model or predict by traditional economics.

With such challenges, data-driven modelling and machine learning is a possible rout ahead. Looking across time, geography, and fuel type, predictive models are able to pick up underlying patterns and make more accurate projections. These findings can then be used to inform smarter regulatory planning, subsidies targeted where they are needed most, and fairer pricing policy that not only reduces inequality but also makes the economy more efficient.

Conquering these obstacles requires not only technical tools but also a commitment to greater transparency, better data quality, and coordinated policymaking at the federal and state levels.

1. **Scope of the Study**

The study is targeted at fuel price forecast in Brazil using data-driven techniques. It’s overall purpose is to identify previously existing patterns, spatial variations, and seasonality in resale prices by type of fuel, as well as by states. Its focus is on using machine learning models in forecasting fuel prices, as well as creating strategic insights that can guide policy making, regulation-level planning, and operational effectiveness in both public and private sectors.

The research in limited to working with publicly available data from Brazil’s National Agency of Petroleum, Natural Gas, and Biofuel (ANP), which provides monthly data from Brazilian fuelling stations in all states. The data include an array of features such as fuel type, measure unit, average resale price, prices by distribution, price margins, state, region, and number of stations includes. The data feature (start date, month, year) supports trend analysis as well as temporal modelling.

The focus is on resale prices, as opposed to distribution or wholesale prices, because they most immediately apply to final consumers, as well as to aggregate economic impact of fuel price uncertainty. Though data include variables that permit geographic variation and product-level variation to be analysed, they lack macroeconomic variables (world prices for crude oil, inflation rates, or exchange rates). The analysis does not attempt to control for external shocks or broad-based economic crises explicitly, thus, but their impact is dealt with in result interpretation. This capstone is not attempting to create a Brazilian fuel price model but is behaviour from provided from provided data. This project has two deliverables: one set of trained predictive models with estimated accuracy, and some data-driven recommendations to support subsequent open, efficient and fair management of fuel price policy in Brazil.

**5. Working with dataset**

Data for our study from Brazil’s Nacional Agency for Petroleum, Natural Gas , and Biofuels(ANP). The dataset is built on monthly observations for gas stations in all 26 Brazilian states and in the Federal District. The data provides fine-grained data on all fuel types, including Gasoline, Diesel, Ethanol, and Natural Gas (GNV), for multiple years. The data is highly granular, with potential for geographical as well as temporal analysis.

Every row in the data is na individual observation for na individual type of fuel in na individual location for na individual time period. Place (Whether by state or by region), fuel type, measurement unit (e.g. R$/liter, R$/kg, R4/m3), and the varying statical measures on the resale price and the distribution price are most essential available features.

The most important columns to examine are:

* start\_date: The reference date for the pricing survey
* region and state: Geographic classification
* product: Fuel type (e.g., gasoline, diesel)
* unit\_of\_measurement: Unit used to report prices
* average\_resale\_price: Main target variable for prediction
* standard\_deviation\_resale, minimum\_resale\_price, maximum\_resale\_price: Variation measures
* average\_resale\_margin: Difference between resale and distribution prices
* number\_of\_stations\_surveyed: Number of stations reporting that fuel in that period
* average\_price\_distribution, standard\_deviation\_distribution: Distribution-level pricing stats

The dataset also contains derived temporal fields:

* month, year, and month-year: Calculated from the start\_date column to allow for time series grouping.

Initial data exploration revealed that prices vary by location as well as by fuel type. Resale prices range on average from R$ 4 to R$ 8 with higher peaks associated with economic shocks such as inflation, tax policy shocks, and world price shocks in petroleum. The number\_of\_stations\_surveyed field is also highly relevant to establishing the representation and validity of individual observations. Small states with few reported stations, for example, will have less reliable means, and this was accounted for in the process filtering and operating on the data.

No valid missing values were present in the data, but less frequent densities in some areas, as well as very low variability in some columns, were present. Outliers were primarily in maximum resale prices and margins being local spikes or data reporting mistakes. These were retained but monitored closely in modeling.

This initial familiarization step with the data was also useful in determining features to include in the machine learning models and features on which to approach with care. The richness an depth in ANP data provided an excellent foundation for proceeding with predictive modeling, exploratory data analysis, and data preprocessing.

1. **Exploratory Data Analysis**

The exploratory data analysis step was created to expose key patterns in fuel prices over time, regionally, and by type of fuel. By way of a series of visualizations and descriptive analysis, we probed the most pertinent behaviours in the data to guide the feature selection process for modelling.

The first task was to understand in what direction fuel prices were trending between regions of Brazil. Plotting the line with “year\_month” and “region”, I could notice that the Southeast and South regions consistently exhibited higher average resale prices over years, whereas the North and Northeast experienced lower prices. This suggests strong regionalization in fuel pricing, which might be influenced by transportation and taxation.

Second, we examined differences by fuel type. I made a bloxplot to observe how the distribution of “average`\_resale\_price” changes by ‘product’. Gasoline and Ehtanol had greater medians and more spread, implying volatility. Diesel and Natural Gas (GNV) were less volatile and less variable. This again supports the necessity of including fuel type as a categorical feature in the predictive model.

Gráfico, Gráfico de caixa estreita

O conteúdo gerado por IA pode estar incorreto.

*Figure 1. Boxplot of average resale prices by fuel type.*

Then examined correlations between numerical attributes with a heatmap. There were strong positive correlations between resale price and maximum price, and distribution price metrics. Standard deviation and resale margin variables also had moderate correlations, and the coefficient of variation was weakly correlated to the target variable.

To investigate seasonal trends, a time series decomposition was conducted on the Southeast region. The decomposition showed a high positive trend and strong annual seasonality. This is in agreement with the inclusion of time-based features such as month and year into the model.

We also verified the representativeness of data by plotting number of surveyed stations for each state. Sparsely populated states likes São Paulo and Minas Gerais yielded the most data. On a normalized basis, though plotting number of surveyed stations per 100,000 population the sparsely populated states of Roraima and the Federal District had comparatively improved coverage.

Briefly, the EDA revealed:

* Cumbersome regional variations in fuel prices;
* Variations in volatility between various types of fuel;
* Prices seasonality;
* Data imbalance for representation in regions;

All these were very essential observations needed for feature engineering and had a direct influence on model selection and model testing elaborated in further sections.

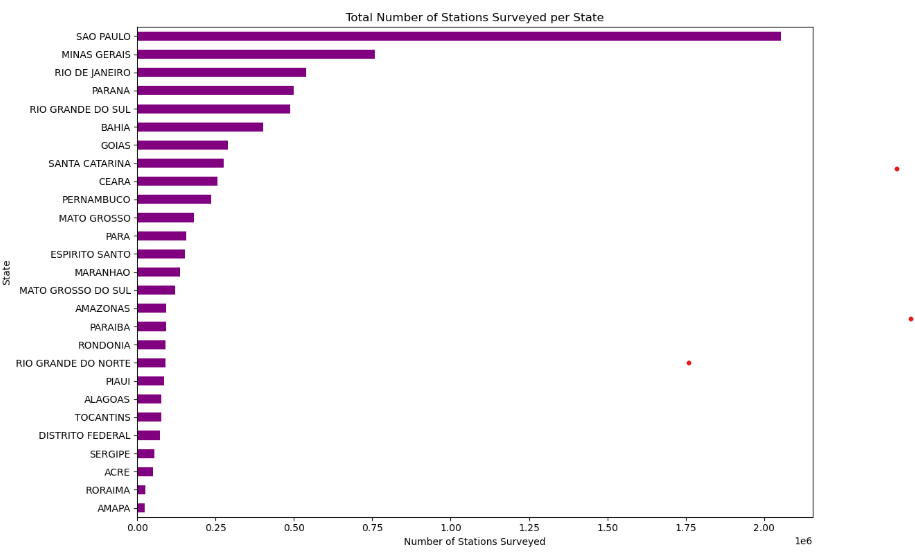
**7 . Data Preparation**

Data preparation was a critical step in getting the dataset clean, consistent, and machine learning model-ready. The process involved several tasks, from column renaming to data type conversion, creating time-based features, encoding categorical variables, to feature selection.

The data set had a very inconsistent set of column names with some of them continuing spaces and varied case. First, all column names were made standard by converting them to lower case and replacing spaces with underscores. This it easier to refer to them during the data manipulation and modelling process.

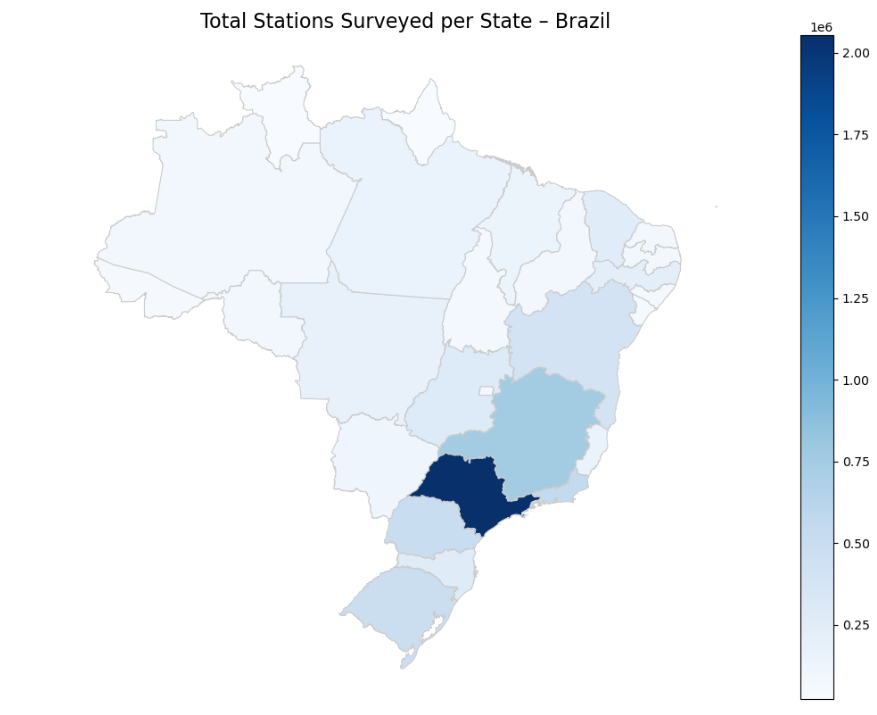
The ‘start\_date’ column, that is, the initial date of each fuel price survey, was converted to datetime type. Its transformation gave birth to three additional temporal columns: ‘year’, ‘month’, and ‘month\_year’. They were required in order to label seasonality, and trends across a longer timeframe in fuel prices.

One of the notable characteristics of the data set is ‘number\_of\_stations\_surveyed’ column, which is the number of data being collected from different states. To better visualize this characteristic, the number of surveyed stations by state was plotted.

***Figure 2 – Bar plot: Total stations surveyed per state***

Next let’s have a look at the map shows that São Paulo, Minas Gerais and Rio de Janeiro have highest number of surveyed station, reflecting the economic cost and infrastructure density in these regions.

For geographical representation, a choropleth map was produced, showing the geographical distribution of surveyed fuel stations in Brazil



*Figure 3 – Choropleth map: Total stations surveyed by state*

This spatial views shows the large regional variations in survey density of data collection. While some states are heavily covered, others ( most notably in the North and Northeast) have much lower numbers of surveyed stations.

But absolute numbers can be deceiving if one does not take into account population variations. So, the ratio of stations surveyed per 100,000 population was also calculated and mapped to evaluate survey fairness by state.

Mapa

O conteúdo gerado por IA pode estar incorreto.

*Figure 4. Stations per 100,000 Inhabitants.*

This second choropleth shows that denser states such as Roraima and the Federal District have comparatively more coverage relative to their population size, while smaller coverage is offered for more populous states such as Bahia and Pará.

Following these analyses, data preparation continued with one-hot encoding categorical features such as ‘region’ and ‘product’. This allowed the machine learning algorithms to handle these important features without bias or artificial ordering.

Numerical features were also inspected, and highly variable or redundant features were dropped to improve the efficiency of modelling.

* 1. **Machine Learning Model**

The modelling process attempted to predict the mean resale price of fuel across Brazilian states from a set of temporal, geographical, and fuel-specific features. Given that the target variable is continuous, regression algorithms were the most appropriate selection.

Two regression algorithms were employed within this study:

- Linear Regression

- Random Forest Regressor

These two models were slected in order to provide both a straigh forward, interpretable baseline (Linear Regression) and more advanced, high-power predictive model able to model non-linear relationships(Radom Forest).

**8.1 Feature Preparation**

Before training the models, preprocessing tasks were performed on the data. Temporal information was derived from the ‘start\_date’ column and new columns such as ‘year’, ‘month’ and ‘month\_year’ were created. Categorical features such as ‘region’ and ‘product’ were one-hot encoded in order to facilitate machine learning algorithms to process them without introducing ordinal biases.

Additionally, numerical features such as ‘number\_of\_stations\_surveyed’, ‘average\_resale\_margin’, and price distribution statistics were incorporated to allow for statistical variation within products and locations.

Significantly, there was no normalization of features because Linear Regression and Random Forest Regressor are both scale-insensitive.

**8.2 Model Training and Testing**

The data was divided into two sets: training and test, 80% and 20% respectively. The division was random but with a fixed random seed to allow for reproducibility of results. This is critical for the purpose of ensuring models are tested against unseen data, providing an accurate estimation of how well the models generalize.

Cross-validation can also be used but is not applied here due to time constraints and computational cost. Future work could extend this and utilize k-fold cross-validation to make it more robust.

**8.3 Linear Regression**

Linear Regression is a basic statistical method that assumes the relationship between input features and target variable to be linear. It is basic, but an excellent baseline model.

The model was trained on the training set and then tested on the test set. Performance metrics such as Mean Squared Error (MSE) and R-squared (R²) were computed. Linear Regression provided a fast, interpretable model but showed poor capability in detecting complex relationships involved in fuel pricing behaviour.

**8.4 Random Florest Regression**

Random Forest Regressor is an ensemble method that trains multiple decision trees and gives the average prediction of the trees. It performs extremely well with non-linear interactions and relationships between the variables, and hence performs extremely well in this case.

Hyperparameters such as the number of trees (‘n\_estimators=100’) and random state were set constant to ensure model stability and performance. No hyperparameter tuning was performed here rigorously but is recommended for future work to realize maximum performance.

The Random Forest model outperformed Linear Regression on all assessment metrics, indicating that fuel price behaviour is governed by non-linear interactions between features.

**8.5 Model Comparison**

Both models were validated with Mean squared Error (MSE) and R-squared (R²) metrics. Preliminary results indicated that the Random forest Regressor had significantly reduced prediction errors relative to Linear Regression and explained a higher percentage of variance.

Gráfico, Gráfico de dispersão

O conteúdo gerado por IA pode estar incorreto.

***Figure 5*** *→ Scatter plot of predicted vs actual resale prices using Random Forest*

Figure 5 shows the predicted vs. actual average resale prices scatter plot of the Random Forest Regressor. For an ideal situation, points should lie near the red diagonal line (ideal prediction). As seen, most predictions are grouped together  closely, highlighting the strength of prediction performance of the model.

**8.6 Conclusion of Modelling Conclusion**  
  
The aim of this phase was to compare the performance of two regression models (Linear Regression and Random Forest Regressor) in the estimation of mean fuel resale prices for Brazilian states.

Table summarizes the performance of both models as regards Mean Squared Error (MSE) and R-squared (R2).

Interface gráfica do usuário, Aplicativo

O conteúdo gerado por IA pode estar incorreto.

*Figure 6. Comparison of MSE and R-squared values between Linear Regression and Random Forest.*

As Figure 6 shows, the Random Forest model did better on both evaluation measures compared to Linear Regression. It had a much lower MSE, i.e., fewer means errors, and effective in describing nearly all variance in the target variable.

The better performance of Random Forest is because it can handle complex non-linear interactions between temporal, spatial, and fuel features, which Linear Regression can not fully encapsulate because of its linear nature.

To better illustrate these varying performance, Figure 7 presents the scatter plot of actual resale price vs predicted resale price for both models.

Gráfico, Gráfico de linhas

O conteúdo gerado por IA pode estar incorreto.

*Figure 7. Comparison of Actual vs Predicted Resale Prices.*

Figure 7 visually confirms that the Random Forest predictions are much nearer to the perfect prediction line than Liner Regression. This visual congruence with the numerical conclusions further attests to the power and predictive advantage of ensemble learning methods for this type of economic forecasting task.

Based on these results, Random Forest Regressor was selected as the ultimate forecasting model for this study. Its increased precision, immunity to overfitting, and capacity for estimating intricate relationships make it an extremely suitable choice for strategic applications in fuel price prediction and examination.

Hyperparameter tuning through cross-validation and incorporation of external macroeconomic indicators can be future studies directed towards further increasing predictive performance.

* 1. **Strategic Recommendations**

The section provides strategic recommendations grounded in data analysis, exploratory findings, and predictive modelling conducted as part of the project. It is an attempt to propose viable concepts that could lead regulatory authorities, fuel distribution, supply chain managers, and policymakers to make data-driven decisions to decrease fuel price volatility, increase market transparency, and streamline operational efficiency.

**9.1 Improve Regional Monitoring and Data Collection**

Exploratory data analysis showed that data coverage is distributed unevenly across Brazilian states with some states, especially those located in the North and Northeast regions, being underrepresented distorts predictive models and limits the accuracy of national price forecasts.

To address this issue, the regulatory agencies should implement a unifor

* **References:**

Agrawal, R., 2023. *https://www.analyticsvidhya.com.* [Online]   
Available at: https://www.analyticsvidhya.com/blog/2021/05/know-the-best-evaluation-metrics-for-your-regression-model/

Anon., 2023. [Online]   
Available at: https://www.simplilearn.com/tutorials/data-science-tutorial/bayesian-linear-regression#:~:text=In%20contrast%20to%20conventional%20regression,variance%20and%20mean%20are%20normalized).

Anon., 2023. *statista.com.* [Online]   
Available at: https://www.statista.com/statistics/1416471/average-two-bedroom-mortgage-repayment-dublin/#:~:text=The%20repayment%20amount%20in%20Dublin,at%20943%20euros%20per%20month.  
[Accessed 10 05 2024].

Anon., n.d. s.l.:s.n.

Brennan, C., 2023. *https://www.irishexaminer.com.* [Online]   
Available at: https://www.irishexaminer.com/news/arid-41140702.html

Koehrsen, W., 2018. *https://towardsdatascience.com.* [Online]   
Available at: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74  
[Accessed 10 05 2024].

McNally, T., 2023. *https://www.irishexaminer.com.* [Online]   
Available at: https://www.irishexaminer.com/news/politics/arid-41200186.html#:~:text=Data%20from%20Inside%20Airbnb%20shows,total%20of%2027%2C439%20separate%20listings.  
[Accessed 2023].