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

**Table of contents**

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

The goal of this project is to analyse house prices going forward, to explore possible solutions regarding availability and affordability. Dataset used is publicly available of one of the main housing portals daft.ie - " daft\_ie\_v1.csv", containing 3967 records of house sales in 2021 - 2022. All these homes were advertised on daft.ie. Dataset contains 22 variables describing house size, number of bedrooms and bathrooms, marketing style, agencies etc.

1. **Exploratory Data Analysis**

To start dataset needs to be checked for duplicates and it shows there are none of them, all the records are unique.

When checking missing values, it was discovered that there are 355 of them, all in column “propertySize”.

Dataset contains 22 features and 3967 observations. Let’s have a look at each of these observations as some are more important than others.

Starting with unimportant features, they include variables like “AMV\_price”, “seller\_name”, “seller\_branch” and others. These features won’t have any influence on the analysis and will be dropped.

“id”. This variable shows unique ID number of each advertised house. It doesn’t contribute to machine learning model but is very useful to perform data manipulation actions as each value is unique.

“price”. Price is a most important feature and the target value of the analysis.

“title”. Title displays the full address of the property. As knowing street name or house number contributes nothing to the analysis, the important feature that needs to be extracted is County. This is discussed in the next section (Data Preparation).

“propertySize”. Another important feature that will be used for machine learning model. This variable has a lot of missing values (355), that will be filled at the later stage. Also “propertySize” contains a number of outliers, 241 in total. Analysis of statistical values for “propertySize” shows that they vary from 1 to 8600, with standard deviation of 255, suggesting data scaling before performing machine learning model. Here is the boxplot of “propertySize”.

A diagram of a box plot

Description automatically generated

*Figure 1. Boxplot of variable “propertySize”*

“numBedrooms” and “numBathrooms” are similar variables, displaying the number of bedrooms and bathrooms each property has and they corelate with property’s price.

“category”. This variable shows if the property is newly built or second-hand.

“ber\_rating” is another important feature, Building Energy Rating. Ratings vary from A1 to F.

1. **Data Preparation**

The first new column created is “County”. The values for this feature are extracted from column “title”, that includes a lot of additional information such as house numbers that are not needed for the analysis. “County” only contains names of the counties. 15 observations don’t mention counties name in the full address, and google maps were used to find out to which county the address belongs, and missing counties were filled.

Let’s have a look at properties sales by counties.

**A pie chart with numbers and text

Description automatically generated**

*Figure 2. Piechart of properties sales by county.*

Next let’s have a look at the variable “propertySize”. It is the variable that is crucial for running machine learning model, but it contains both outliers and missing values. Let’s deal with the outliers first.

The scatterplot shows that there are 4 outliers that are significantly higher than the rest of the values. Analysis of these 4 outliers shows that one of the properties includes ruined house with a lot of land around it <https://www.daft.ie/for-sale/detached-house-rathlikeen-mullinavat-co-kilkenny/3967177>, Property in co. Wexford is also a site. The remaining two properties are standard houses in co. Kildare https://www.myhome.ie/residential/brochure/2719-dara-park-newbridge-kildare/4557218 and co. Dublin https://www.myhome.ie/residential/brochure/13-donomore-crescent-tallaght-dublin-24/4553476, suggesting that imputed property size is inaccurate. The observations with the two sites are deleted and the “propertySize” of the two standard houses are replaced with median value of 3-bedroomed homes. To calculate median value, dataframe is filtered, and the result of median value is 100. After removing two rows, dataframe is reduced to 3965 observations.

“propertySize” still contains 355 missing values, represented as NaN and nan. Missing values are found in houses have with various number of bedrooms. These values are replaced accordingly, with the median values of the similar-sized houses.

A graph of a number of blue dots

Description automatically generated

*Figure 3. Properties sold in each county.*

After filling in the values for houses up to ten bedrooms, the last house left is a historic place in Dingle Ballintaggart house, with 23 bedrooms, and 6 acres of land <https://seeinsidedingle.com/ballintaggart-house/>. After extensive research it was impossible to find out the exact size of the property, and the decision was taken to drop this particular row.

With all the missing values filled, incorrect values replaced, and big outliers dropped it’s time to reduce the dataset and prepare it for machine learning model. Insignificant variables are dropped, variable names replaced with more user-friendly names.

Here is the Heatmap showing corelations between the most important numerical values.

A screenshot of a graph

Description automatically generated

*Figure 4. The Corelation Heatmap.*

Heatmap indicated strong corelation between number of bedrooms and overall properties size. There is a moderate corelation between Price and all other numerical parameters.

1. **Machine Learning Model**

The first model used is Linear Regression. For the first run of the model just current numerical variables are used (“Bedrooms”, “Bathrooms” and “Size”). The target variable is “Price”. After running Linear Regresion model this result was achieved:

A table of numbers with numbers

Description automatically generated

The are the errors calculated.

A black text on a white background

Description automatically generated

The evaluation of results. Calculated Mean Absolute Error shows that average predictions are off by almost 152000. Mean Squared Error is quite big figure, indicating that it is heavily influenced by the outliers. Root Mean Squared Error is another way of measuring to measure the difference between predicted and actual values, and this error is also quite high. Overall, the model didn’t perform particularly well and needs an improvement.

To improve the performance of the machine learning models few avenues can be explored. Dataset can be reduced eliminating large outliers, as they don’t represent typical, dominating housing. Dataset also needs to be scaled and encoded. More machine learning models can be performed, hyperparameters tuned.

Let’s start with filtering the dataset. Dataset contains variable “Bedrooms” showing that they range from 1 to 16.

A graph of a number of rooms

Description automatically generated

*Figure 5. Distribution of house bedrooms.*

Distribution shows heavily skewed data to the right. Houses with more than 5 bedrooms would be untypical for Ireland. However there still are 58 of 6-bedroomed houses in the dataset making 1.5% of the total dataset. It makes sense to keep theses houses, deleting all the houses with more than 6 bedrooms, as they are unique housing estates, distorting the shape of dataset. After this distribution of Bedrooms becomes close to normal.

Similarly, houses with more than 6 bathrooms are taken out of the dataset.

Taking biggest houses out of dataset normalised it and in addition to that one more very expensive house is taken out leaving the dataset to 3923 observations.

As previous results on Linear Regression results aren’t satisfactory this time data is encoded and scaled. Encoding is done using pandas dummies method. Scaling is done using MinMax scaler as variables “Price” and “Size” are skewed. “Year” and “Month” are also scaled to bring all the variables onto the same scale.

With encoding and scaling done, time to move on to machine learning algorithms. Target variable is “Price”.

4 machine learnings models are tried: Linear Regression, K-Nearest Neighbour, Bayesian Ridge Regression and Random Forest Regressor. Testing sample size is 20% of the data. With all the models 4 metrics are calculated:

* Mean Absolute Error (MAE). This metric simply calculates mathematical difference between predicted and actual values.
* Mean Squared Error (MSE). It’s similar to MAE but instead of calculating absolute difference it calculates squared difference.
* Root Mean Squared Error (RMSE). It’s related to MSE, as it is square root of MSE.
* R Squared (R2). This is an important metric as it evaluates the performance of the model. R2 give result in percentange, which can tell what percentage of the variance in the target variable can be explained by the model. As the result is closer to 1 as better.

So to build and deploy a generalized model is required to evaluate the model on different metrics which helps to better optimize the performance, fine-tune it, and obtain a better result. (Agrawal, 2023).

Starting with Linear Regression model the following results were calculated:

Mean Absolute Error: 0.038505747526977

Mean Squared Error: 0.0038924614539166654

Root Mean Squared Error: 0.06238959411565895

R-squared: 0.48133570916408885

To evaluate them model needs to be compared to the other models.

The next model tried is K-Nearest Neighbour. It is slightly different algorithm as it calculates the average or weighted average of the target variable “Price” if the nearest neighbours. Therefore it’s important to pick the best number of neighbours. To do that list of neighbours is created ranging from 1 to 16. Number of neighbours can’t be too high as it can lead to overfitting the model. Running the program the best number of neighbours returned is 12. KNN model is fitted with 12 neighbours and the results are:

Mean Absolute Error: 0.04082953721911113

Mean Squared Error: 0.004159506143888466

Root Mean Squared Error: 0.06449423341577498

R-squared: 0.44575242943594884

These results are worse than the ones calculated by Linear Regression.

Next model tried is Bayesian Regression. In Bayesian linear regression, the mean of one parameter is characterized by a weighted sum of other variables. (Anon., 2023)

These are the results:

Mean Absolute Error: 0.038019194797133525

Mean Squared Error: 0.003890211574539317

Root Mean Squared Error: 0.06237156062292587

R-squared: 0.4816355020086769

The last model tried is Random Forest Regressor. This model creates sets of decision trees to calculate best fits. Here are the results of Random Forest’s model:

Mean Absolute Error: 0.035471822327157225

Mean Squared Error: 0.0035250141184498067

Root Mean Squared Error: 0.0593718293338668

R-squared: 0.5302974815350642

With all the calculations done now it’s the time to bring all the results into one table for comparison and pick the best performing model.

A table of numbers and symbols

Description automatically generated

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

K-Nearest Neighbour performed worst out of all 4 models with all metrics being the worst. Bayesian Ridge and Linear Regression performed very similarly, but Random Forest comes on top. It’s R2 is the highest explaining 53% of variance in the target variable. Mean Absolute Error is also the lowest.

To further confirm all the best fit regression lines are plotter together.

A diagram of different values

Description automatically generated

*Figure 7. Plotting actual vs predicted values for different models.*

Plotting of the regression lines calculated by all 4 models also confirms that Random Forest performed best. The RED line indicates actual values, what would be the perfect prediction and the PURPLE line representing Random Forest Regressor is the closest one to the red line.

Therefore, Random Forest Regressor is picked for further analysis – cross-validation and hyperparameter tuning.

1. **Cross-validation and Hyperparameter tuning**

Cross-validation is a technique of resampling dataset into smaller samples, called k-folds. It gives the idea how accurate the original prediction is when the model was run just once. In this case model is run at 5, 10, 15 and 20 k-folds. These are the results:

A table of numbers and symbols

Description automatically generated

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

Results after cross-validation are in fact worse comparing to results after performing single train-test split. There could be several reasons for that, for example overfitting the model. Let’s try to tune hyperparameters to see if it makes a difference.

Hyperparameters include the number of decision trees in the forest and the number of features considered by each tree when splitting a node. (The parameters of a random forest are the variables and thresholds used to split each node learned during training). The best hyperparameters are usually impossible to determine ahead of time, and tuning a model is where machine learning turns from a science into trial-and-error based engineering. (Koehrsen, 2018).

To look for best parameters method know RandomisedSearchCV is used. It creates the grid of hyperparameter ranges and takes random samples from the grid. There are number of hyperparameters in random forest but most important are six:

* n\_estimators - number of trees in the foreset
* max\_features - max number of features considered for splitting a node
* max\_depth - max number of levels in each decision tree
* min\_samples\_leaf - min number of data points allowed in a leaf node
* min\_samples\_split - min number of data points placed in a node before the node is split
* bootstrap - method for sampling data points (with or without replacement)

After performing grid search best parameters are established:

{'n\_estimators': 400,

'min\_samples\_split': 10,

'min\_samples\_leaf': 4,

'max\_features': 'auto',

'max\_depth': 70,

'bootstrap': True}

With these best parameters in place, Random Forest Regressor model is fitted again and new results for 4 metrics are calculated.

Mean Squared Error: 0.003460181623377258

Mean Absolute Error: 0.03503225240840679

Root Mean Squared Error: 0.05882330850417424

R-squared: 0.5389363082718301

To compare the results a table is created.

A white rectangular box with black numbers

Description automatically generated

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

Tuning hyperparameters improved the results slightly. MAE, RMSE and MSE all decreased, and R2 increased by nearly 1%, showing that now almost 54% variance of the target value is explained.

1. **Feature Importances**

With the result of the machine learning model now finalised let’s have a look at what features had the biggest impact on the model performance. To obtain importance scores for each feature attribute feature\_importances\_ is used. These scores indicate relative contribution of every feature to the model’s predictions.

There are in total 58 features in df\_scaled dataframe, and many of them have very minimal impact to the model predictions. Although these unimportant features still can be helpful in terms of making strategic decisions. Both, 10 most important features and 10 least important features are visualised.

A graph with different colored bars

Description automatically generated

*Figure 10. Barchart of 10 most important features.*

Graph tells those houses sold in county Dublin contributes 20% of the total importance in the model. This is very significant contribution, and having in mind that attached to county Dublin, counties Kildare and Wicklow are also in top ten, it indicates that a lot of activity is done in a greater Dublin area.

Total size of the house, number of bedrooms and bathrooms are also amongst key predictors with contribution around 15%.

Category Second\_Hand could be misleading as this dataset is dominated by second hand dwelling, new homes only having 87 observations.

The graph shows what features dominate the dataset but it doesn’t necessary mean that government should focus on providing more similar houses like that. For example Ber\_C1, energy rating feature is heavily contributing to the machine learning model, but it is not because people are looking for houses with this energy rating but because a lot of houses like that are already built, and they dominate second-hand house sector.

Let’s have a look what features contributed least to the machine learning model. Plotting 10 least important features:

A graph of different colored bars

Description automatically generated

*Figure 11. Barchart of 10 least important features.*

Although this graph may seem insignificant, it still can be helpful for analysts and strategic planners. First of all it can help to fix inaccuracies in the dataset. Here variables Type\_House, Ber\_B1, Type\_Studio have zero impact to model predictions, indicating, that these variables can be replaced, modified or even deleted. The further least important features are standing for less populated counties of Ireland, again indicating that these features have very little impact on price predictions, and perhaps very little market activity is happening in these counties comparing to large counties like Dublin.

1. **Conclusion**

This report provides an overview of our investigation into Irish housing market. It highlights the complexities and diverse factors that contribute towards the housing problems and emphasize on the importance of data to understand these issues and how to address them.

Linear Regression machine model was run. The predictions and the errors were calculated. This was done using three independent variables. Errors calculated are big and the model can be improved. Data used is unscaled, categorical variables that could have impact on the analysis are unencoded, different test sizes and different amounts of k-folds can be used in the future.

To improve the prediction results dataset was filtered, taking out the biggest outliers, encoded and scaled. 4 machine learning models were tried – Linear Regression, K-Nearest Neighbour, Bayessian Ridge and Random Forest. Best results were achieved by Random Forest. This model then was further used for cross-validation and hyperparameter tuning. Hyperparameter tuning further slightly improved results.

Finally, the feature importance scores were calculated. They proved that houses sold in county Dublin have biggest impact on predicting house prices. Closely followed features were size of the house, number of bedrooms and bathrooms.

These insights suggest that there is huge pressure on greater Dublin area. Government should focus on building new housing here, and without government’s intervention this crisis will never be solved. Focus should be on building more apartments or smaller houses, as due to high building costs, these are cheaper and faster to build. The shift towards apartment building in Dublin has started already.

Analysis is finished but it only really scratched the surface. There can be more done, different train/test splits can be tried, more models tried and more of different hyperparameters tuned. To build completely reliable house prices prediction model is impossible however, as these price heavily depend on internal and external factors, that can’t be easily foreseen.

1. **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].