Predictive Analysis of Fuel Mileage per Liter for Motor Vehicles

Wilmer Alexander Panqueva Caballero, Julieth Alejandra Páez Camargo

Abstract – This project aims to develop a predictive model to estimate fuel mileage per liter for vehicles in Mexico, considering characteristics such as brand, model, fuel type, and performance in different conditions. The tool will use data analysis and machine learning techniques to help consumers make accurate estimates, optimize their budgets, and make sustainable decisions. Given the current lack of precise tools for this estimation, the model seeks to address this need and promote greater awareness of the environmental impact of fuel consumption.

Índice de Términos – Predictive Analysis, Fuel Mileage per Liter, Motor Vehicles, Dataset, Algorithm, Linear Regression, Environmental Impact

Resumen – Este proyecto tiene como objetivo desarrollar un modelo predictivo para estimar el kilometraje por litro de combustible en vehículos en México, tomando en cuenta características como marca, modelo, tipo de combustible y rendimiento en distintas condiciones. La herramienta utilizará técnicas de análisis de datos y aprendizaje automático para ayudar a los consumidores a hacer estimaciones precisas, optimizar sus presupuestos y tomar decisiones sostenibles. Dado que actualmente existen pocas herramientas precisas para esta estimación, el modelo busca llenar esta necesidad y fomentar una mayor conciencia sobre el impacto ambiental del consumo de combustible.

I. INTRODUCTION

The cost of gasoline consumption is a key factor in the economy of users in Mexico. This project aims to develop a predictive model for vehicle fuel mileage per liter based on specific vehicle characteristics, such as brand, model, fuel type, city and highway performance, emissions, among others. Using data analysis and machine learning techniques, the goal is to create a tool that allows consumers to accurately estimate fuel mileage per liter and make informed decisions, optimizing their budget and fostering greater awareness of environmental impact.

II. PROBLEM

In Mexico, estimating the fuel mileage per liter for motor vehicles is a challenge for vehicle owners due to the variability in factors such as vehicle type, engine, and driving conditions. The lack of precise tools for making this estimation leaves users with limited information, which can lead to costly and inefficient decisions, both economically and environmentally. This project seeks to develop a predictive model that, based on specific vehicle characteristics, allows for an accurate calculation of fuel mileage per liter, helping consumers optimize their budget and make more sustainable decisions.

III. EXPLANATION OF PROGRESS MADE

We imported our project into a notebook to work through the problem:

- Imported libraries for data manipulation, such as pandas for data modification, seaborn for visualization, and sklearn for data scaling, among others.
- Loaded the dataset and created a DataFrame for analysis.
- Verified the dataset information and described its columns, including the number of records and data types for each column.
- Renamed some columns to make them more descriptive and remove unnecessary characters, e.g., renaming "Potencia (HP)" to simply "Potencia" and "Trans." to "Transmisión".
- Checked for duplicate values using the duplicated function.
- Removed duplicates with the drop_duplicates function.
- Identified non-numeric, null, or missing data using the isnull and isna functions.
- Removed rows with null values, as only a few records had this characteristic.
- Checked unique values in each column to distinguish between categorical and numerical types.
- Examined outliers visually by calculating limits using standard deviation and creating boxplots with seaborn.
- Scaled the most relevant columns (Rendimiento_ciudad, Rendimiento_carretera, Rendimiento_mixto, Rendimiento_ajustado, Emisiones_CO2, Emisiones_NOx, Gas_efecto_invernadero) using the MinMaxScaler method.
- Created a correlation matrix to identify key features for addressing our problem.
- Applied a RandomForest model to highlight the most relevant features and compared them to the correlation matrix and heatmap, selecting features such as Rendimiento_mixto, Rendimiento_ajustado, Emisiones_CO2, and Rendimiento_carretera.
- Developed visualizations to analyze column behavior.
- Compared combined performance to mixed performance to understand the advantages for vehicles with this fuel efficiency type.
- Compared vehicle fuel type with city performance and engine size to determine the most efficient fuel type based on performance and engine size relevance.

Document presented on November 6, 2024, to Jorge Eduardo Hernández Rodriguez.
Submission of Final Project Progress.

- Compared city performance by vehicle brand to evaluate brand impact on fuel consumption.
- Analyzed the three performance types by vehicle brand to identify which vehicles have optimal fuel efficiency.
- Assessed city performance by vehicle model, finding that the vehicle model year directly influences performance.
- Identified vehicles with the lowest city performance, concluding that they tend to be sports vehicles.
- Found that vehicles with the highest mixed performance include Ford, KIA, and Honda.

IV. Next Steps

V. Development of the Linear Regression Model

- Implement a linear regression model to predict mileage based on the features selected during preprocessing. This involves:

- Model Training: Utilize the preprocessed dataset to train the model, ensuring that the training data includes relevant features (such as vehicle weight, horsepower, and fuel type).

- Model Validation: Split the dataset into training and testing data to evaluate the model's accuracy and prevent overfitting.

- Parameter Optimization: Adjust parameters and evaluate error metrics, such as Mean Squared Error (MSE), to refine the model.

VI. Challenges Encountered

1. Complexity in Deciding Whether to Remove Less Relevant Features: We find it challenging to determine whether to eliminate less relevant features from the dataset.

2. Difficulty in Selecting Relevant Features: Choosing the most relevant features is difficult, especially considering that many of them are interrelated.

3. Comparing Combined Performance with Mixed Performance: We aim to compare combined performance with mixed performance to understand the benefits that vehicles with this type of performance have over traditional fuel usage.

4. Identifying Anomalous Values: Another challenge is to identify anomalous values in the dataset.

VII. Explanation of the Algorithm to be Used for the ML Model

Description: Linear regression is a supervised learning method that estimates the relationship between a dependent variable (in this case, mileage) and one or more independent variables. The objective is to find a line (for a single independent variable) or a plane (for multiple variables) that best fits the data points in space, minimizing the difference between the model's predictions and the actual values.

Application: Linear regression is chosen for its ability to model simple, direct relationships between mileage and relevant variables. In this context, it is expected that factors such as vehicle weight and horsepower will directly influence mileage, making linear regression an appropriate method.

Evaluation Metrics: To assess model performance, metrics such as Mean Squared Error (MSE) and the Coefficient of Determination (R^2) will be used. MSE helps to understand the average prediction error, while R^2 indicates the proportion of variation in the data explained by the model.

VIII. Deploying the Solution on a Cloud Server.

Platform Choice: Use a cloud platform such as AWS, Google Cloud, or Microsoft Azure to deploy the model. These platforms provide scalable infrastructure and facilities for machine learning projects.

Environment Setup: Create a cloud instance with the necessary resources to run the linear regression model. Dependencies such as Python, machine learning libraries (scikit-learn, numpy, pandas), and a web server environment like Flask or FastAPI should be installed to expose the model as a service.Conclusions

Model Implementation as an API: Transform the model into a REST API that accepts input data requests and returns mileage predictions. This web service will allow other systems or applications to submit vehicle data and receive real-time predictions.

Testing and Scalability: Conduct tests to verify the model's response and prediction time. Monitor resource usage to determine if server configuration improvements are needed, ensuring that the service can handle greater loads in the future if required.

IX. Conclusions

Relevance of Features: The appropriate selection of features is crucial for the performance of the regression model. While some features may seem less relevant, their removal could affect the model's accuracy due to the interrelationship between variables.

Benefits of Performance Comparison: Comparing combined performance with mixed performance provides valuable insights into vehicle efficiency. This not only helps understand the impact on fuel consumption but also guides consumers in choosing more sustainable vehicles.

Importance of Anomaly Detection: Identifying anomalous values is essential to ensure data quality. Outliers can distort analysis results and model performance, making it crucial to detect and appropriately handle them.