Machine Learning

US Electric Vehicle Population

Universidade do Porto

Faculdade de Engenharia

FEUP

Authors :

Lehmoudi Kamel
up202301437

Philius Moraless
up202302762

***Abstract*:**

This machine learning project focuses on using a comprehensive dataset containing information on the US vehicle population, including electric range, city, state, brand, year, and model. The project addresses three distinct problems:

Electric Range Prediction:

Methodology: Linear Regression, Decision Trees, Hyperparameter Tuning, Random Forest, Gradient Boosting, and Cross-Validation techniques were employed.

Evaluation: Mean Squared Error (MSE) was computed to assess the predictive performance of each model during cross-validation.

Brand Classification:

Methodology: Naive Bayes, Logistic Regression, Neural Networks, and Support Vector Machines (SVM) were utilized.

Evaluation: Classification accuracy and other relevant metrics were employed to evaluate the effectiveness of the models.

State Prediction:

Methodology: Linear Regression and Ridge Regression were employed to predict the state of vehicles.

Evaluation: Model performance was assessed using appropriate regression evaluation metrics.

The project provides insights into electric range prediction, brand classification, and state prediction, addressing practical challenges in the automotive domain. The diverse set of machine learning algorithms used for each problem showcases the versatility of the models. The evaluation metrics help in comparing and selecting the most suitable models for each task. The findings contribute to the broader field of machine learning applications in the automotive industry.

Introduction :

In recent years, the automotive industry has undergone a significant transformation driven by technological advancements and a growing emphasis on sustainability. With an increasing number of electric vehicles entering the market, understanding and predicting various facets of their performance and characteristics have become pivotal. This machine learning project delves into the analysis of a comprehensive dataset encompassing crucial information

about the US vehicle population, including electric range, city, state, brand, year, and model.

The primary objective of this project is to address three key challenges within the automotive domain, each presenting distinct machine learning problems:

Electric Range Prediction/Brand Classification/State Prediction

I. Data Overview

A. Dataset features

The dataset includes 17 features but we will focus on the following key features:

Electric Range: The driving range of electric vehicles on a single charge, a crucial factor in assessing their usability and efficiency.

City: The city in which the vehicle is registered, providing geographical context and potential regional variations.

State: The state in which the vehicle is registered, offering insights into state-specific trends and patterns.

Brand: The brand or manufacturer of the vehicle, a categorical variable reflecting the diversity of vehicles in the dataset.

Year: The manufacturing year of the vehicle, allowing for temporal analysis of electric vehicle characteristics.

Model: The specific model or make of the vehicle, offering a detailed breakdown of the dataset.

Vehicle type: plug-in hybrid electric vehicles (PHEVs) and battery electric vehicles (BEVs).

B. Data Preprocessing Steps

Prior to model development, the dataset underwent a series of preprocessing steps to ensure its suitability for machine learning tasks:

Handling Missing Values: Any missing values in the dataset were addressed through methods such as imputation or removal, based on the nature and distribution of missing data.

Categorical Variable Encoding: Categorical variables, such as "City," "State," and "Make," were encoded using suitable techniques to facilitate their inclusion in machine learning models.

Feature Scaling: Numerical features were scaled to ensure consistent ranges and prevent certain features from dominating the model training process.

Data Splitting: The dataset was divided into training , testing and validation sets to facilitate model training and evaluation.

The careful preprocessing of the dataset is essential for ensuring the robustness and reliability of the machine learning models applied to the subsequent problems.

II. Problem statements

Here are the 3 problems we covered in this project.

Problem 1: Predictive model for electric range

Objective: Build a model to predict the electric range of vehicles based on features like model year, brand, and electric vehicle type.

Problem 2: Customer Segmentation for Electric Vehicles

Objective: Segment electric vehicle customers based on demographic features to tailor marketing strategies and services.

Problem 3: Charging Station Placement Optimization

Objective: Predict the State for electric vehicle charging stations in different locations based on factors such as city, Model Year, Range, Brand, Electric Vehicle Type.

III. Methodology

In this section, we detail the methodologies employed to address the three distinct challenges within the automotive domain. Leveraging the rich dataset described in the previous section, our approach involves a strategic combination of traditional and advanced machine learning techniques.

The following sub-sections provide detailed insights into each problem's unique challenges and the tailored methodologies adopted to unravel patterns, optimize models, and derive meaningful predictions.

A. Electric range prediction

In addressing the first problem of predicting the electric range of vehicles, a range of regression models were considered to capture the intricate relationships between key features. The following methodologies were employed:

a. *Linear regression*

A fundamental regression technique was initially applied to establish a baseline prediction model. This model assumes a linear relationship between the predictor variables (year,brand,vehicle type) and the target variable (electric range), allowing for straightforward interpretation of coefficients.

b. *Decision Trees*

Decision Trees were employed to capture non-linear relationships and hierarchical decision-making processes within the dataset. These models partition the feature space into regions, aiding in capturing complex patterns.

c. *Hyperparameter Tuning*

To optimize the performance of the models, hyperparameter tuning was conducted using techniques such as grid search. This process involved systematically searching the hyperparameter space to identify the configurations that yielded the best predictive results.

d. *Random Forest*

A Random Forest ensemble model was constructed to leverage the collective wisdom of multiple decision trees. This approach aims to enhance predictive accuracy and generalizability by reducing overfitting.

e. *Gradient Boosting*

A boosting ensemble technique was employed to sequentially build an ensemble of weak learners. This iterative process focuses on the mistakes made by previous models, improving overall predictive performance.

f. *Cross-Validation (5 fold)*

To assess the generalization performance of each model, cross-validation techniques, such as k-fold cross-validation, were applied. This involved partitioning the dataset into training and validation sets multiple times, providing a robust estimate of model performance.

B. Customer segmentation

The concept of customer segmentation aligns closely with the task of brand classification. By effectively classifying vehicles into their respective brands, we inherently create a segmentation that groups vehicles based on the manufacturers or brands they represent. This segmentation is pivotal not only for understanding the distribution and prevalence of different brands within the dataset but also for extracting valuable insights into brand-specific characteristics and trends. The process of brand classification serves as a practical approach to segmentation, allowing us to discern patterns, preferences, and variations that can be instrumental in shaping marketing strategies.

Addressing the challenge of brand classification involves the application of classification algorithms to accurately categorize vehicles into their respective brands.The predictor variables considered for this task include the manufacturing year and electric range, while the target variable is the brand of the vehicle. In order to streamline the segmentation process and facilitate a more in-depth analysis, we opted to condense the classes to the five most prevalent vehicle brands across the United States. This reduction in the number of classes aims to enhance the clarity of segmentation. The methodologies employed for this task are as follows:

a. *Naive Bayes*

A probabilistic classification algorithm was employed for its simplicity and efficiency. The assumption of independence among features makes it particularly suitable for brand classification tasks.

b. *Logistic regression*

Logistic Regression, a widely used classification algorithm, was chosen to model the probability of a vehicle belonging to a specific brand. This method provides insights into the importance of individual features in the classification process.

c. *Neural Network*

Neural Networks, known for their ability to capture complex patterns, were implemented for brand classification. A multi-layer perceptron architecture was used to model intricate relationships among features.

d. *Support Vector Machines (SVM)*

SVM, a powerful classification technique, was employed to find the optimal hyperplane that separates different brands in the feature space. The kernel trick was utilized to handle non-linear decision boundaries.

e. *Evaluation Metrics*

The performance of each classification model was evaluated using metrics such as accuracy, precision. These metrics provide a comprehensive assessment of the models' ability to correctly classify vehicles into their respective brands.

C. Charging Station Placement Optimization

Predicting the state of registration allows us to identify areas with a higher demand for charging facilities, enabling stakeholders in the electric vehicle ecosystem to plan and allocate resources effectively. Consequently, this predictive modeling not only aids in addressing the challenge of state prediction but also serves as a foundational element for the broader goal of optimizing the placement of charging stations to enhance the overall electric vehicle charging infrastructure.

The third problem involves predicting the state in which vehicles are registered, with the predictor variables being the year, model, city, vehicle type, electric range, and brand. The target variable, in this case, is the state of

registration. The methodologies employed for addressing this task are as follows:

a. Linear regression

Linear Regression was chosen as the initial modeling approach to establish a baseline for predicting the state of registration. This linear model considers the relationships between the predictor variables (6 variables) and the target state variable.

b. Ridge regression

To address potential multicollinearity and enhance the robustness of the linear model, Ridge Regression was applied. This technique introduces regularization to the linear regression, preventing overfitting and improving the model's generalization performance.

c. Evaluation Metric

The performance of each regression model was evaluated using standard regression metrics such as Mean Squared Error (MSE),. This metric provides insight into the accuracy and precision of the models in predicting the state of registration.

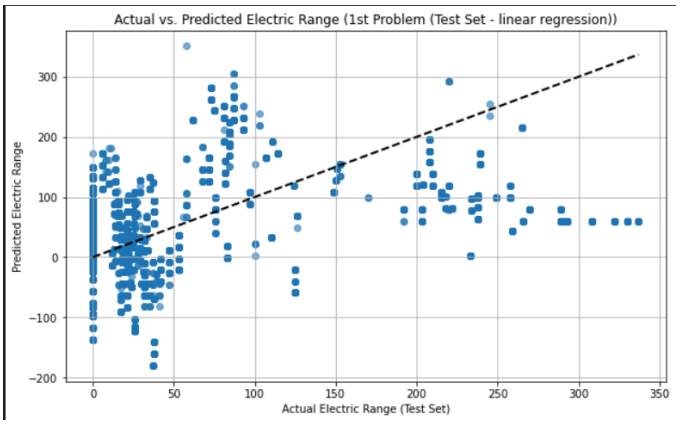
IV. RESULTS

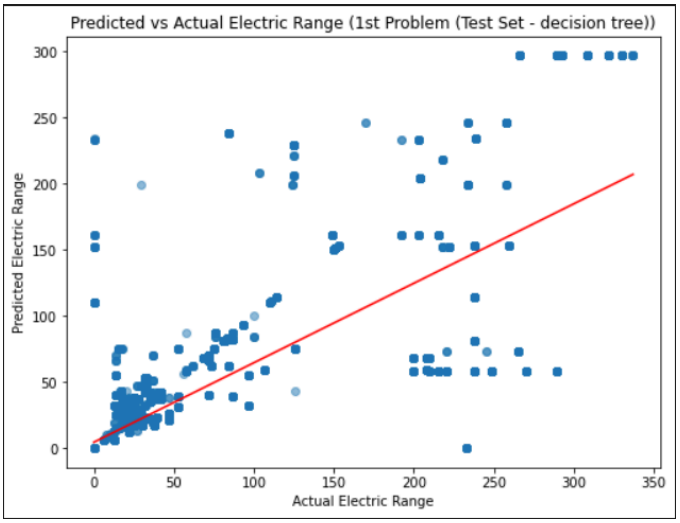
Here, we share what we found in our machine learning work. First, for predicting electric range, we check the Mean Squared Error. We compared different methods to see which one worked best. Then, for figuring out vehicle brands, we looked at how often our models got it right. Lastly, for guessing where vehicles are registered, we check the MSE. We'll talk more about what these results mean in the next parts.

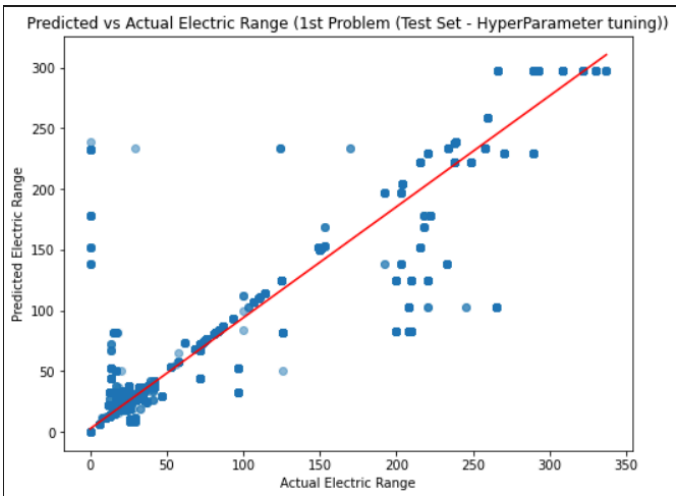
A. Problem 1 : Electric Range Prediction

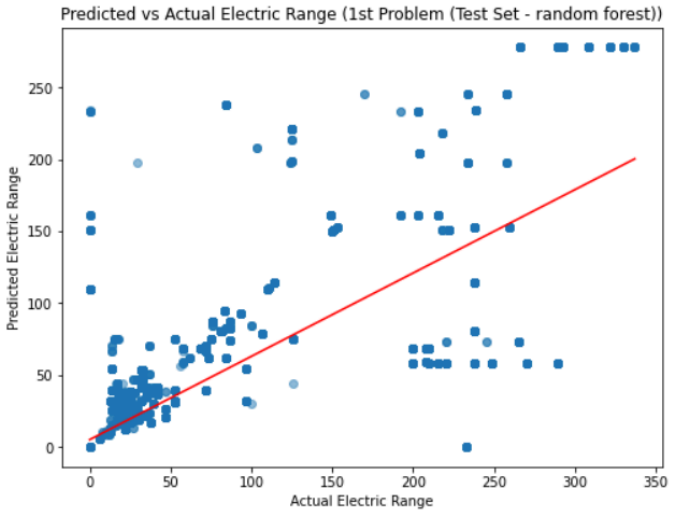
Model	MSE	
	Validation	***Test***
Linear regression	/	8982.49
Decision trees	/	3930.16
Hyperparameter tuning	/	804.83
Random forest	5235.12	3927.37
Gradient Boosting	1937.57	1673.17
Cross Validation 5 fold	114.96	116.47

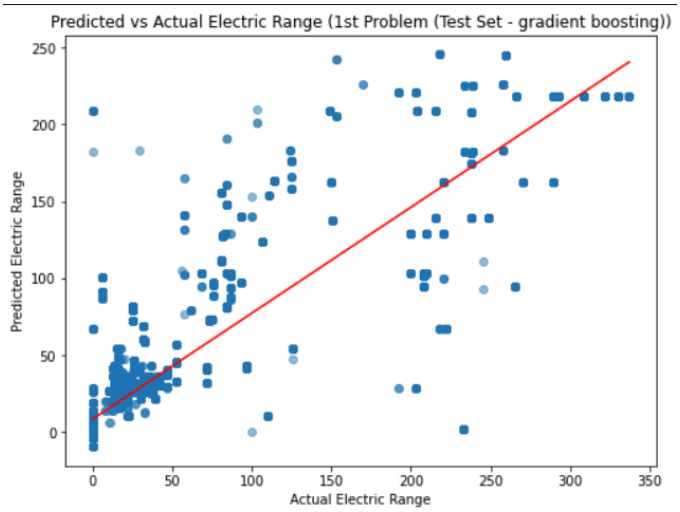
Fig. 1. Performance of different models in the problem 1

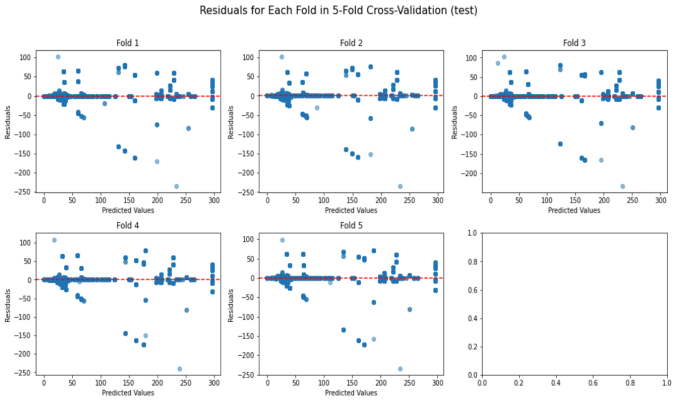












B. *Problem 2 : Brand classification*

Model	Test	
	accuracy	***misclassification***
Naive Bayes	0.701	0.299
Logistic regression	0.711	0.289
Neural Network	0.895	0.105
SVM	0.861	0.139

Fig. 2. Performance of different models in the problem 2

C. *Problem 3: State prediction*

Model	***MSE test***
Linear regression	65.053
Ridge regression	65.053

Fig. 3. Performance of different models in the problem 3

V. INTERPRETATION

In interpreting the results, the performance metrics from the test dataset tables offer crucial insights into the effectiveness of our machine learning models. For the electric range prediction task, lower Mean Squared Error (MSE) values indicate that our models were better at estimating electric range. Comparing MSE values across models helps identify which approach yielded the most accurate predictions. In the realm of brand classification, higher accuracy values signify a better match between predicted and actual brand labels. Looking at accuracy values for each model allows us to pinpoint which classification approach performed most effectively. For state prediction, regression metrics such as Mean Squared Error (MSE) offer a measure of how close our predictions were to the actual values. Understanding these metrics aids in evaluating the precision of our models.

While our machine learning models have provided valuable insights, it is essential to acknowledge certain limitations and assumptions inherent in our approach. One limitation pertains

to the assumption of linearity in the relationships modeled, especially in the case of regression tasks.

Non-linear patterns in the data may not be fully captured by our chosen models. Additionally, the effectiveness of our models relies on the quality and representativeness of the training data. Any biases or gaps in the dataset could impact the models' generalization to new, unseen data. Furthermore, assumptions about the stability of underlying patterns in electric vehicle characteristics and consumer behavior are implicit in our predictions. Changes in these patterns over time or external factors not accounted for in the data may affect the models' accuracy. Recognizing these limitations is crucial for a nuanced understanding of the applicability and reliability of our machine learning results.

VI. Conclusion

In summarizing our findings, this machine learning project has yielded valuable insights into the prediction of electric vehicle characteristics, brand classification, and state registration. Our models, though subject to certain limitations, have demonstrated promising predictive capabilities.

For electric range prediction, the comparative analysis of models highlighted those with superior accuracy, offering potential avenues for more precise estimations. Brand classification models showcased varying degrees of success, providing a foundation for understanding and improving the categorization of vehicles into brands. State prediction models,evaluated through regression metrics, provided a means to gauge their accuracy in forecasting vehicle registration locations.

As we reflect on our achievements, we recognize the need for continued refinement, addressing limitations, and exploring avenues for future research. Our contributions pave the way for more informed decision-making in the dynamic landscape of electric vehicles and charging infrastructure.

VII. References

[1] Data origine : https://catalog.data.gov/dataset/electric-vehicle-population-data

[2] https://www.commerce.wa.gov/growing-the-economy/energy/electric-vehicles/

[3] https://komonews.com/news/local/washington-state-continues-pushing-toward-frontlines-of-electric-vehicle-movement-in-us

[4] https://theicct.org/publication/update-on-electric-vehicle-adoption-across-u-s-cities/