Prediction of C02 Emissions

DATA SCIENCE TECHNIQUES PROJECT

Hung Nguyen (1671819)

Table of Contents

Preface

This research investigates the world of predictive modeling for CO2 emissions as the automobile sector struggles with the requirement to reduce carbon emissions and embrace sustainable practices. The study introduces an effective journey through data analysis, preparation, and modeling by focusing on a large dataset containing multiple vehicle parameters. Following that, the use of linear regression, random forest, and k-nearest neighbor models demonstrated the significant impact of careful feature selection on predictive accuracy. The research emphasizes the greater significance of estimating CO2 emissions from vehicle data, which aligns with the automotive industry's drive for sustainable practices. The main goal is to highlight the critical importance of feature selection in improving forecast accuracy and to provide valuable insights into the complex relationship of factors controlling CO2 emissions.

Introduction

We are living in a time of environmental awareness and sustainable practices; it is critical to comprehend and reduce carbon dioxide (CO2) emissions from every daily activity. Our societies seek more environmentally friendly choices, and understanding the factors influencing emissions becomes increasingly important. As a student at HAN University of Applied Sciences, I will conduct a data-driven investigation into the complex factors impacting CO2 emissions in the automotive landscape. By applying data science approaches, the purpose of this investigation is to uncover patterns, correlations, and insights from a large dataset. I hope to add significant knowledge to the discussion of cleaner, more sustainable transportation alternatives by diving into the complicated interplay of vehicle attributes, fuel usage, and environmental effects. This investigation is consistent with the larger commitment to promoting informed decision-making in the automobile sector and advocating for a greener, more sustainable future.

Data understanding

The dataset used in this research is from the Canadian Government's official open data website, and it is a thorough compilation of CO2 emissions by cars over a seven-year period. The richness of data includes a wide range of features, providing a detailed understanding of the complex

relationship between vehicle parameters and CO2 output. Using this dataset coincides with a commitment to transparency and reliance on authoritative sources, both of which are critical for conducting rigorous analyses in the field of environmental impact assessments. The use of government-approved data boosts the legitimacy of the findings, encouraging a data-driven investigation into the dynamics of CO2 emissions in the automotive industry.

A careful review of the information is necessary for understanding the complexities of automotive CO2 emissions. The dataset covers a wide range of attributes, each of which provides a unique perspective on the automotive landscape. The "Make" indicates the vehicle's manufacturer, while the "Model" reveals the specific variant. The term "Vehicle Class" divides autos into groups such as COMPACT, MID-SIZE, and SUV-SMALL. Engine characteristics such as "Engine Size(L)" and "Cylinders" define the vehicle's power and efficiency. The "Transmission" section contains information about the transmission type, such as automatic (A), automated manual (AM), or manual (M).

Fuel-related attributes are crucial in analyzing the environmental impact. "Fuel Type" indicates the type of energy, with codes such as Z for premium gasoline, D for diesel, and E for ethanol (E85). Fuel consumption is divided through "Fuel Consumption City (L/100 km)," "Fuel Consumption Hwy (L/100 km)," and "Fuel Consumption Comb (L/100 km)," along with the combined rating in miles per imperial gallon (mpg). These measurements provide on the efficiency of vehicles in both city and highway driving scenarios.

The CO2 emissions, which are measured in grams per kilometer, represent the environmental footprint of each vehicle, encompassing both city and highway driving. This comprehensive dataset, with its diverse features and variety of information, is the foundation for our investigation into the factors impacting CO2 emissions from a wide range of automobiles.

Data preparation

Several steps were taken in the early phases of data preparation to ensure the quality and reliability of the dataset for future regression analysis. To enhance the readability and consistency of the data set, column names are standardized and renamed in the same format. The subsequent exploration involves a detailed examination of the dataset’s information and

summary statistics, which would offer valuable insights into the structure of the data. Additionally, any missing values were carefully analyzed and addressed before it was suitable for analysis. In terms of addressing data integrity, there were a total of 1103 rows of duplicate entries within the dataset. In the context of vehicular CO2 emissions, each data entry should distinctly represent a unique vehicle observation. Duplicate entries could arise from various sources, such as recording errors, data integration processes, or multiple entries for identical vehicles. However, given the nature of the dataset, where each row represents a unique car, it was decided to keep only distinct observations and remove redundant rows. By executing the removal of duplicated rows, the dataset underwent a transformative refinement, ensuring that each vehicle's characteristics and associated CO2 emissions were accurately represented.

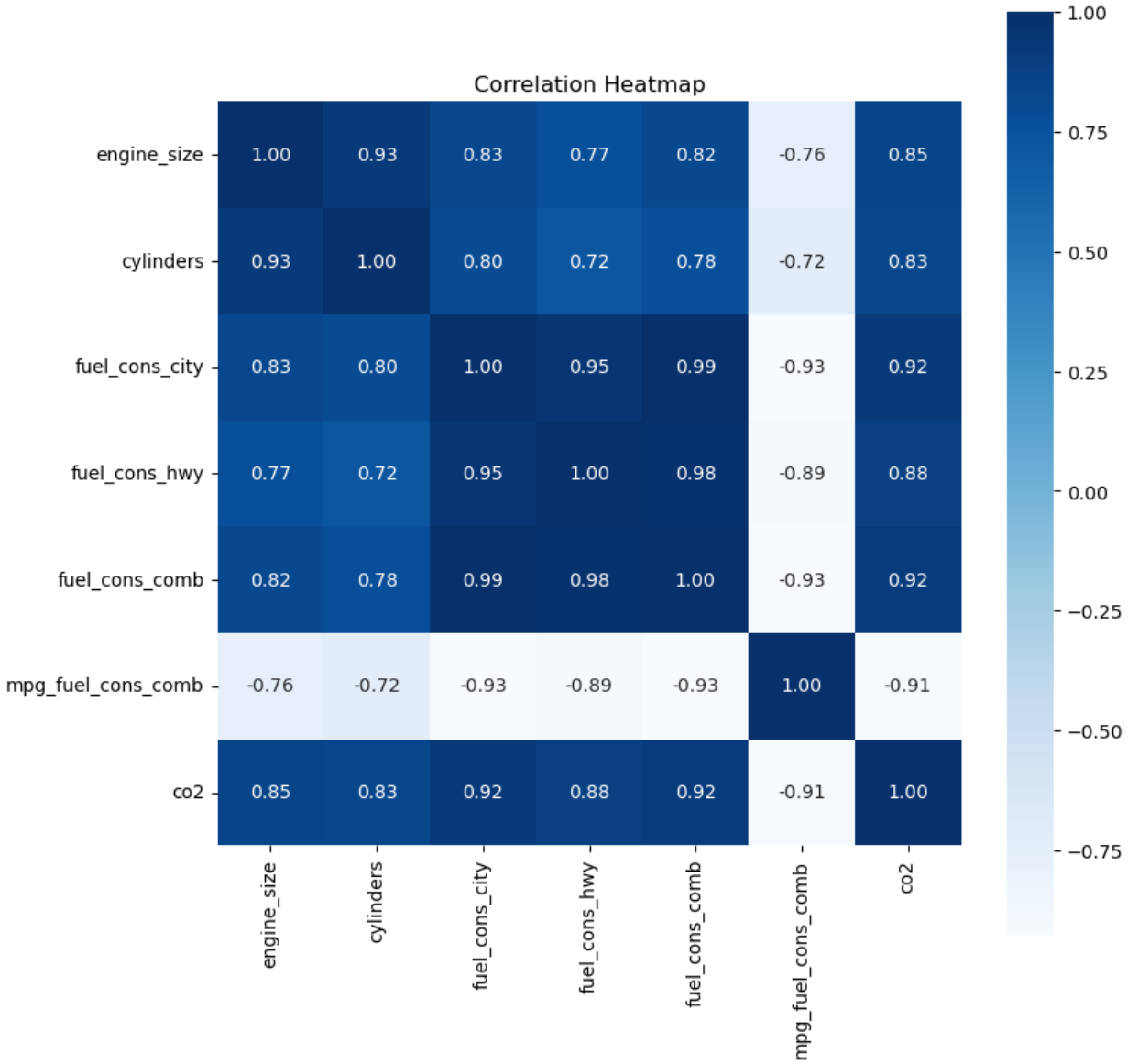
The next step in the data preparation phase included transforming categorical variables into a format suitable for model training. This process, known as one-hot encoding, was applied to categorical columns, including 'make,' 'model,' 'vehicle class,' 'transmission,' and 'fuel type'. Because of this transformation, categorical variables could be quantitatively represented, allowing them to be included in the study without enforcing ordinal relationships. These dummy variables were easily merged back into the original dataset, prefixed suitably for clarity.

Subsequently, a basic model, Linear Regression was implemented, where all independent variables were considered simultaneously. This is an effective way to analyze the effect of various factors influencing CO3 emissions. The dataset was divided into features (X) and the target variable (y), where 'X' encompasses all relevant features except for the CO2 emissions ('co2'), and 'y' represents the CO2 emissions. To ensure numerical consistency for model compatibility, the data types of 'X' and 'y' were converted to numeric formats, accommodating any potential discrepancies. This foundational model serves as a benchmark, providing insights into the baseline predictive capabilities of the chosen features on CO2 emissions.

Data visualization

To explore the relationships between numeric features in the dataset, a regression approach was employed to identify potential quadratic patterns in the relationship between independent

variables and the dependent variable – CO2 emissions. This was accomplished by developing a correlation matrix heatmap, which is a visual tool that displays the strength and direction of correlations between variables. The matrix was created by considering just numeric columns, ensuring a full picture of CO2 emissions correlations.



Based on the heatmap, most features exhibited positive correlations with CO2 emissions, indicating that as these features increase, CO2 emissions tend to increase as well. This aligns with expectations, as factors such as engine size, cylinders, and fuel consumption are commonly associated with higher emissions. However, the feature 'mpg_fuel_cons_comb' demonstrated a negative correlation with CO2 emissions. This implies that higher fuel efficiency, as measured in miles per gallon (mpg), is associated with lower CO2 emissions. But in fact, ‘mpg_fuel_cons_comb' and 'fuel_consumption_Comb_(L/100_km)' essentially represent the same information, with the latter having a higher correlation. This investigation set the framework for understanding the impact of individual characteristics on CO2 emissions. After that, the decision was made to keep 'Fuel_Consumption_Comb_(L/100_km)' for further research due to its greater correlation and practical significance in the context of the dataset.
In addition to examining correlations between numeric features, an analysis of relationships between categorical variables was critical for obtaining an overall understanding of the dataset.

While the correlation matrix focuses primarily on numerical correlations, investigating categorical correlations requires specific studies adapted to the nature of these features. This analysis focused on 'make,' 'model,' 'vehicle class,' 'transmission,' ‘engine size,’ ‘cylinder,’ and 'fuel type,' and involves computing mean CO2 emissions for separate categories within each characteristic. The changes in average emissions among different groups were then visually represented using bar graphs. By analyzing bar graphs, fuel type exhibited a relatively modest impact on CO2 emissions. In contrast, engine size and cylinder count emerged as pivotal factors influencing CO2 emissions, with larger engine sizes and cylinder counts correlating with heightened emission levels. These categorical insights underscore the multifaceted dynamics influencing CO2 emissions within the dataset, offering valuable cues for subsequent modeling and decision-making processes.

Modeling and evaluation

Linear Regression is a traditional statistical method that assumes a linear relationship between the selected features and the target variable, making it suitable for capturing simple, appropriate

correlations. Random Forest, on the other hand, employs an ensemble of decision trees, introducing nonlinearity and capturing complicated interactions between attributes. Its strength is in dealing with high-dimensional data and making accurate predictions. Meanwhile, K-Nearest Neighbours, a distance-based method, predicts data points in the feature space based on their proximity. KNN is sensitive to local patterns and excels in capturing complex correlations within specific data regions.

Based on the key understandings achieved from data visualization activities, I decided to use key findings from correlation matrices and bar charts, a selection of features—specifically, 'engine_size,' 'cylinders,' 'fuel_cons_comb,' 'fuel_cons_city,' and 'fuel_cons_hwy'—were deliberately selected for the improved model. The purpose of this feature selection was to focus on the most significant variables influencing CO2 emissions. The following linear regression modeling, which was adapted to these carefully chosen parameters, led to improved prediction performance. In comparison to the initial model with a Mean Squared Error of 2.3321450315167564e+20 and an R-squared of -6.462523858107161e+16, the new model demonstrated remarkable enhancements, boasting a significantly lower Mean Squared Error of 439.65 and an impressive R-squared of 0.8781. In the context of linear regression models, an R-squared of 0.8781 for the new model suggests that the selected features account for approximately 87.81% of the variability in CO2 emissions. This indicates a strong connection between the independent variables and the target variable, suggesting the model's ability to capture and explain CO2 emission patterns. In general, a higher R-squared value implies a better match, and in this case, the model shows the ability to estimate CO2 emissions based on the identified important factors.

Following with the research into predictive modeling, two further algorithms—Random Forest and K-Nearest Neighbours (KNN)—were used to improve perceptions of CO2 emissions based on the selected features. The Random Forest model had a valuable Mean Squared Error of 90.08 and a strong R-squared value of 0.9750, suggesting its high predictive accuracy and ability to capture complicated correlations in the data. Similarly, the KNN model had a Mean Squared Error

of 107.83 and a significant R-squared value of 0.9701, emphasizing the model's precision in predicting CO2 emissions. These results highlight the reliability of the chosen features and validate the suitability of machine learning algorithms for this predictive task. The high R-squared values across models indicate a constant and significant association between the selected features and CO2 emissions, validating the predictive models' ability to capture the underlying patterns in the dataset.

Conclusion

This project emphasizes the crucial significance of feature selection in improving model accuracy and accessibility in the field of predictive modeling for CO2 emissions in the automobile sector. The selection of essential variables, such as engine size, cylinders, and fuel consumption measurements, emerged as critical after analyzing the dataset and utilizing insights from the correlation matrix and bar charts. This careful selection of variables sought not only to improve forecast performance but also to provide a better understanding of the basic mechanisms controlling CO2 emissions. In a period where environmental sustainability has become importance to everyone, the automotive industry is pushed to make data-driven decisions that align with worldwide efforts to minimize carbon footprints. Predicting CO2 emissions from specific features not only provides stakeholders with useful insights into the environmental impact of various vehicle attributes but also allows manufacturers and customers to make more informed decisions. This study suggests the strategic use of predictive modeling to negotiate the complicated landscape of emissions in a technologically advanced and environmentally responsible.

Appendices

Link to Github repositories:

https://github.com/hungnguyen0409/Individual-assignment.git