MBA 442B: MACHINE LEARNING ALGORITHMS – I

CIA-1

DOMAIN-SPECIFIC MODEL BUILDING

SUBMITTED BY

JOJU FELIX

(2327465)

SUBMITTED

TO

PROF. HELEN JOSEPHINE V.L

CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

Fuel Consumption Analysis Report

1. Business Understanding

a. Problem Identification

The problem is to analyze the factors affecting CO2 emissions from vehicles and to build a predictive model to estimate CO2 emissions based on vehicle characteristics.

b. Variables

The dataset includes the following variables:
- Make
- Model
- Vehicle Class
- Engine Size
- Cylinders
- Transmission
- Fuel Type
- Fuel Consumption (City, Highway, Combined)
- CO2 Emissions

c. Objectives

The objectives are:
1. To understand the relationship between vehicle characteristics and CO2 emissions.
2. To build a predictive model for CO2 emissions using Multiple Linear Regression.
3. To interpret the model from a business perspective to identify significant predictors.

2. Data Understanding

a. Data Collection

The dataset is obtained from a secondary source (FuelConsumption.csv).

b. Data Exploration

Initial exploration includes summary statistics and visualization to understand data distribution and relationships between variables.

c. Assessing Data Quality

The data quality is assessed by checking for missing values, outliers, and inconsistencies.

3. Data Preparation

a. Data Integration

Not applicable as the data is from a single source.

b. Data Cleaning

- Missing Value Analysis: Identify and handle missing values.
- Data Imputation: Impute missing values if necessary.
- Variable Standardization: Standardize numerical variables.
- Feature Selection/Feature Engineering: Select relevant features and create new features if needed.
- Outlier Detection and Treatment: Identify and handle outliers.

4. Modeling (Multiple Linear Regression)

a. Model Selection and Assumptions

Multiple Linear Regression is selected. Assumptions include linearity, independence, homoscedasticity, and normality of residuals.

b. Model Output

Model Equation: CO2 Emissions = β0 + β1 * Engine Size + β2 * Cylinders + ... + βn * Fuel Consumption Combined
Explanation of Parameters: Each coefficient represents the impact of a unit change in the predictor on CO2 emissions.
Explanation of Coefficients: Significant coefficients are identified based on p-values.
Model Fit Indices: R-squared, Adjusted R-squared, and AIC are used to evaluate the model fit.
Any Other Method: Stepwise regression is used for model selection.

c. Model Interpretation

The model indicates that variables such as Engine Size, Cylinders, Transmission, and Fuel Consumption are significant predictors of CO2 emissions. Business strategies can focus on these factors to reduce emissions.

5. Model Evaluation and Diagnostics

The model's performance is evaluated using residual analysis, Mean Squared Error (MSE), and R-squared values. Diagnostics indicate good model fit with some outliers.

6. Democratizing the Solution

A note on making the solution accessible to stakeholders, including clear communication of findings, visualizations, and actionable insights.

Multiple Regression

The multiple regression analysis produced the following key findings:

Residuals:
- Min: -10.647
- 1Q (First Quartile): -0.002
- Median: 0.000
- 3Q (Third Quartile): 0.021
- Max: 10.647

Most residuals are close to zero, indicating good model fit for many data points, but with some large residuals indicating potential outliers or areas where the model does not fit as well.

Coefficients:
- Intercept:
  - Estimate: 280.5
  - Std. Error: 10.58
  - t value: 26.52
  - Pr(>|t|): < 2e-16 (highly significant)

Notable Coefficients:
- MAKEASTON MARTIN:
  - Estimate: 14.67
  - p-value: 0.012107 (significant)
- MAKEBENTLEY:
  - Estimate: 46.55
  - p-value: 4.62e-12 (highly significant)
- MAKELAMBORGHINI:
  - Estimate: 57.29
  - p-value: 4.50e-14 (highly significant)
- MODELA8 TDI CLEAN DIESEL:
  - Estimate: 39.25
  - p-value: 1.68e-08 (highly significant)
- MODEL1500 4X4 DIESEL:
  - Estimate: 38.34
  - p-value: 7.36e-12 (highly significant)

Interpretation:
- Significant Variables: Coefficients with low p-values (typically < 0.05) suggest a significant relationship with CO2 emissions. For instance, cars made by Bentley and Lamborghini significantly increase CO2 emissions.
- Non-Significant Variables: Variables with high p-values indicate no significant effect on CO2 emissions. For example, MAKEBMW has a high p-value (0.834729), suggesting no significant impact.

- Model Fit: The residuals indicate the model fits the data reasonably well for most observations but has some outliers. The significant coefficients and low p-values for several makes and models suggest these are strong predictors of CO2 emissions.
This output provides insight into how different car makes and models influence CO2 emissions, allowing for targeted strategies to reduce emissions by focusing on the significant contributors.

Lasso Regression

The Lasso regression model produced the following key findings:

Model Fitting: The Lasso regression model was fitted using cross-validation to find the optimal lambda value.

Evaluation Metrics:
- Mean Squared Error (MSE): 56.27
- R^2 (Coefficient of Determination): 0.9852

Interpretation:
- The low MSE indicates the model predicts well on average.
- The high R^2 indicates the model explains 98.52% of the variance in CO2 emissions.
- Lasso regression performs feature selection by shrinking some coefficients to zero, leading to a more sparse model.

Ridge Regression

The Ridge regression model produced the following key findings:

Model Fitting: The Ridge regression model was fitted using cross-validation to find the optimal lambda value.

Evaluation Metrics:
- Mean Squared Error (MSE): 1566.89
- R^2 (Coefficient of Determination): 0.5872

Interpretation:
- The MSE indicates how close the predicted values are to the actual values on average.
- The R^2 value suggests that the model explains about 58.72% of the variance in CO2 emissions.

Comparison of Multiple Regression, Lasso, and Ridge

The comparison of the three regression models indicates the following:

Residuals Analysis (Multiple Regression):
- Residuals close to zero indicate good model fit for many data points, with some large residuals indicating potential outliers or areas where the model does not fit as well.

Coefficients (Multiple Regression):
- Significant Coefficients: Coefficients with low p-values suggest a significant relationship with CO2 emissions, e.g., Bentley and Lamborghini significantly increase CO2 emissions.- Non-Significant Coefficients: High p-values indicate no significant effect on CO2 emissions, e.g., MAKEBMW.- Model Fit: Residuals indicate the model fits the data reasonably well for most observations but has some outliers.

Lasso Regression:
- The low MSE and high R^2 suggest that the Lasso model is highly effective in predicting CO2 emissions, performing feature selection by shrinking some coefficients to zero, leading to a more sparse model.

Ridge Regression:
- The higher MSE and lower R^2 compared to Lasso indicate that the Ridge model is less effective in predicting CO2 emissions. However, Ridge regression is more suitable when multicollinearity is a concern.

Conclusion

The analysis of CO2 emissions from vehicles using Multiple Linear Regression, Lasso Regression, and Ridge Regression models has provided valuable insights into the factors influencing CO2 emissions and the predictive capabilities of each model.

Multiple Linear Regression

The multiple regression model indicated that variables such as Engine Size, Cylinders, Transmission, and Fuel Consumption are significant predictors of CO2 emissions. The residuals analysis showed that most residuals were close to zero, indicating a good model fit for many data points, though some large residuals suggested potential outliers or areas where the model did not fit as well. Significant coefficients such as those for Bentley and Lamborghini demonstrated a strong relationship with increased CO2 emissions, while non-significant variables like MAKEBMW showed no significant impact.

Lasso Regression

The Lasso regression model demonstrated exceptional predictive performance with a low Mean Squared Error (MSE) of 56.27 and a high R^2 value of 0.9852, indicating that it explains 98.52% of the variance in CO2 emissions. Lasso's feature selection capability,

which shrinks some coefficients to zero, resulted in a more parsimonious model that still captured the significant predictors of CO2 emissions effectively.

Ridge Regression

The Ridge regression model had a higher MSE of 1566.89 and a lower R^2 value of 0.5872 compared to Lasso, indicating it explained only 58.72% of the variance in CO2 emissions. Ridge regression is useful for addressing multicollinearity, but in this case, it was less effective than Lasso in terms of predictive accuracy.

Comparison

Comparing the three models, the Lasso regression model emerged as the most effective in predicting CO2 emissions, with its ability to perform feature selection and achieve high predictive accuracy. The multiple regression model also provided valuable insights into the significance of various predictors, but with slightly less predictive accuracy. Ridge regression was the least effective in this scenario but remains valuable for addressing multicollinearity.

Recommendations

Based on the findings, it is recommended to use the Lasso regression model for predicting CO2 emissions due to its high accuracy and feature selection capability. The insights from the multiple regression model should be utilized to understand the impact of specific vehicle characteristics on CO2 emissions. This analysis can inform targeted strategies to reduce CO2 emissions by focusing on significant predictors such as Engine Size, Cylinders, Transmission, and Fuel Consumption.

Overall, this comprehensive analysis aids in understanding the factors affecting CO2 emissions and provides robust predictive models to estimate emissions based on vehicle characteristics.