

ALY 6020:

PREDCTIVE ANALYTICS

**Mid-Week 2: Linear Regression Analysis for Car Prices**

Submitted To:

Prof. Chinthaka Pathum Dinesh Herath Gedara, Faculty Lecturer

Submitted By: Abhilash Dikshit

Academic Term: Winter 2024

Northeastern University, Vancouver, BC, Canada

Master of Professional Studies in Analytics

January 20, 2024

**Title: Linear Regression Analysis for Car Prices**

1. **Abstract**

This report presents a linear regression analysis aimed at predicting the prices of cars using a dataset and data dictionary. The analysis involves feature selection, model fitting, and interpretation of results. The objective is to identify the most significant variables impacting car prices and evaluate the accuracy of the predictive model.

1. **Introduction**

The dataset under consideration contains information on various car attributes such as engine specifications, fuel type, and performance metrics. The analysis utilizes the linear regression model to establish relationships between these features and the target variable, which is the price of the cars.

1. **Methodology**

**Data Loading and Preprocessing**

The dataset was loaded and examined for any missing or inconsistent values. Non-numeric columns were dropped or encoded appropriately for compatibility with the linear regression model. The dataset was split into training and testing sets to assess the model's performance.

**Linear Regression Model**

A linear regression model was fitted to the training data using the Ordinary Least Squares (OLS) method. The model's summary provides insights into the coefficients, p-values, and R-squared values.

A screenshot of a computer

Description automatically generated

1. **Results and Discussion**

**Model Summary**

The linear regression model resulted in the following key statistics:

- R-squared: 0.865 (Adjusted R-squared: 0.851)

- F-statistic: 63.28

- P-value (Prob F-statistic): 1.71e-56

These values indicate that the model explains a significant proportion of the variance in car prices.

**Significant Variables**

Three notable methods—forward selection, backward elimination, and stepwise selection—were employed to identify the most significant variables.

A screenshot of a computer program

Description automatically generated

Stepwise:

A screenshot of a computer

Description automatically generated

**Interpretation of Coefficients**

The coefficients associated with each feature provide insights into their impact on car prices. For instance, the coefficient for 'enginesize' suggests that a one-unit increase in engine size results in a decrease of $12,723.7 in car prices.

1. **Accuracy Assessment**

The model's accuracy was evaluated based on the R-squared value, which indicates the proportion of variance in car prices explained by the model. With an R-squared of 0.865, the model demonstrates a high level of accuracy.

1. **Conclusion**

The linear regression analysis revealed that 'enginesize,' 'carwidth,' 'stroke,' 'car\_ID,' 'compressionratio,' 'peakrpm,' and 'citympg' are significant variables influencing car prices. Of these, 'enginesize' had the greatest positive influence, as indicated by its high coefficient.

The model, with an R-squared of 0.865, is deemed accurate in predicting car prices. However, it is essential to consider potential limitations, such as multicollinearity, as indicated by the large condition number.

1. **Recommendations**

The findings from this analysis can be valuable for car manufacturers and sellers to understand the factors contributing to car prices. Further refinements and validations may enhance the model's robustness.