**Car Price Prediction: A Linear Regression Analysis Report**

**Introduction**

The purpose of this study is to develop a linear regression model to predict car prices based on various features and characteristics of the vehicles. Predicting car prices can provide valuable insights for car buyers, sellers, and manufacturers, enabling them to make informed decisions in the automotive market. Car price forecasting has been a research area that has attracted much attention since it requires a lot of effort and knowledge from domain experts. (Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, Jasmin Kevrić, 2019) By utilizing a dataset containing information on car attributes and corresponding prices, my aim is to build a robust model that predicts car prices based on these features.

This report will focus on the research question: how the price of cars would be influenced by size of car, horsepower, milage in city and milage on highway. There are 728 related results on the University of Toronto Libraries website, illustrating the importance and popularity of this topic. Based on related research of car price (Ramgiri Siva, Adimoolam M, 2022), mileage and horsepower of car are two popular predictors. I split the milage into two parts: milage in city and milage on highway. Except for milage and horsepower, size of car is also an important factor that influences car’s price. Among all related papers, no research builds a linear regression model of car price that includes all of cars’ size, horsepower, milage in city and milage on highway as

predictors yet. The research will investigate a linear regression model that predicts car price, including cars’ size, horsepower, milage in city and milage on highway as predictors. In addition to the four variables, I choose several variables that might influence car price: car brand and car fuel type. They are also two popular predictors in research of predicting car price.

**Method**

The research aims to investigate the ‘best’ linear regression model that predicts car price, including cars’ size, horsepower, milage in city and milage on highway as predictors. The original data used in this research was scraped from Kaggle (Manish Kumar, 2019).

Before starting the research, I will perform a short exploratory data analysis on the car price dataset. The dataset, named "CarPrice\_Assignment.csv," contains information about various car features and their corresponding prices. I choose the numerical variables cars’ size, horsepower, milage in city and milage on highway as predictors. A histogram is plotted to visualize the distribution of car prices in the dataset. This provides an understanding of the range and spread of prices, and it shows that it is left-skewed. Scatter plots are generated to examine the relationships between selected numerical variables (e.g., horsepower) and car prices. This analysis aims to identify any linear or non-linear associations between these variables and the target variable. After assessing the dataset, I find that there exist no missing values.

After the EDA, I will split the data into two parts: a training dataset containing 50% collected data and a testing dataset containing the rest 50% collected data. Then I build the full model that consists of all possible predictors with the training dataset. I use the scatterplots between car price and predicted car price to check condition 1. If there is a functional pattern in the plots, then condition 1 is satisfied. I will use the scatterplots between numerical predictors in the models to check condition 2. If there are all linear or random patterns in the plots, then condition 2 is satisfied. If the model satisfies the two conditions, I will build residual plots for fitted response variable, numerical predictors, and normal QQ-plots to check assumptions.

To check the linearity assumption, if the residual plots have random patterns, then the assumption is satisfied. To check the independent assumption, if there are no apparent clusters in residual plots, then the assumption is satisfied. If points in residual plots distribute randomly, then homoscedasticity is satisfied. If the normal QQ-plot shows that the residuals' line doesn't deviate significantly from the standard line, then normality assumption is satisfied. If the model has no violations of assumptions, I will proceed to model selection. If it has some model assumption violations, then I will apply Boxcox transformation to handle the violations. After transformation, by comparing the original full model with the transformed model, I will choose the most appropriate one. For example, if all violations are fixed in transformed model, I will continue with the transformed model. If only some or no violations are fixed in transformed models, I will select the model with the slightest violation of assumptions to proceed as the full model.

After choosing a model from previous diagnostics, I will check multicollinearity for full model. If there are predictors with VIF >5, I will remove some predictors to ensure VIF < 5. Then I will use stepwise selection method based on AIC to build a new model, which will add and remove predictors in the current model until all further steps will increase AIC of current model. However, if the model has violations of assumptions, this method might be unreliable. I will also build a reduced model that consists of only the four predictors I am interested in: cars’ size, horsepower, milage in city and milage on highway. If the model assumptions of all candidate models are satisfied, I will choose the model with the highest adjusted and smallest AIC, AICc, BIC values. If all the models have violations of assumptions, I will check and choose the one that has the slightest violation of assumptions.

After deciding the preferred model, I will check leverage points, outliers, and influential points, which can be measured by three cutoffs: Cook’s Distance, DFFITS, DFBETAS. I will continue to model validation if there is no contextual reason to remove the problematic observations. I will build the same model on testing dataset and then check changes in coefficients, predictor significance, model assumptions, and adjusted in test model. If the differences in all coefficients and adjusted aren’t large, the same predictors appear significant, and no new

assumption violations appear, then the final model is likely validated.

**Results**

The following table shows the numerical summaries of each variable.

Detailed introduction of variables is in Appendix.

Table 1: Numerical summaries of each variable

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Minimum | 1st Quarter | Median | Mean | 3rd Quarter | Maximum |
| enginesize | 61.0 | 97.0 | 120.0 | 126.9 | 141.0 | 326.0 |
| horsepower | 48.0 | 70.0 | 95.0 | 104.1 | 116.0 | 288.0 |
| citympg | 13.00 | 19.00 | 24.00 | 25.22 | 30.00 | 49.00 |
| highwaympg | 16.00 | 25.00 | 30.00 | 30.75 | 34.00 | 54.00 |
| price | 5118 | 7788 | 10295 | 13277 | 16503 | 45400 |

Based on the provided statistics for the variables in the dataset, here are the important features to consider: The enginesize variable ranges from 61.0 to 326.0. This feature represents the size of the car's engine in cubic centimeters (cc). It is an important indicator of the car's power and performance. The price variable ranges from $5,118 to $45,400. It represents the price of the car. Price is an important feature as it directly influences the affordability and market value of the vehicle. After making summaries of variables in training and testing datasets, I find that numerical summaries in both datasets are similar. The distribution of car price is skewed to left in both datasets.

The model satisfies conditions 1 and 2 but have violations in normality. Residual plots and a normal QQ-plot are created to assess the assumptions of the regression model. The residual plot displays no discernible pattern, indicating that the assumption of linearity and homoscedasticity is met. The normal QQ-plot shows the residuals are not closely following the diagonal line, having significant deviations on two sides. This is a violation of normality. By Boxcox

transformation, I use log price, log horsepower and log enginesize in transformed models. The transformed model has no improvement in violations and still has violation of normality. Therefore, the original model is the full model in model selection. All predictors have VIF < 5. The stepwise model selection produces the same full model, which means adding or removing any variables will increase AIC in the full model. Then I compare the full model and the reduced model that consists only four predictors in interest. The reduced model has a more significant linear pattern in residual plots; thus it has worse violation in linearity. The residual plots and normal QQ-plot of the final model are included in Appendix.

After deciding the preferred model, I check leverage points, outliers, and influential points, which can be measured by three cutoffs: Cook’s Distance, DFFITS, DFBETAS. There are 6438 leverage points, 21 outliers, 8422 influential points in training dataset, with no contextual reasons to remove them.

Table 2: Parameter estimates for final model (training dataset)

|  |  |  |
| --- | --- | --- |
|  | Coefficient Estimate | Standard Error |
| (Intercept) | -4149.05 | 3455.08 |
| enginesize | 108.66 | 14.47 |
| horsepower | 60.81 | 18.25 |
| citympg | 112.44 | 234.26 |
| highwaympg | -180.07 | 201.13 |

Residual standard error: 3397 on 97 degrees of freedom

Multiple R-squared: 0.8234, Adjusted R-squared: 0.8161

Table 3: Parameter estimates for final model (testing dataset)

|  |  |  |
| --- | --- | --- |
|  | Coefficient Estimate | Standard Error |
| (Intercept) | -2480.24 | 3939.31 |
| enginesize | 64.96 | 21.25 |
| horsepower | 79.30 | 18.39 |
| citympg | -106.25 | 231.35 |
| highwaympg | 42.47 | 187.22 |

Residual standard error: 3273 on 98 degrees of freedom

Multiple R-squared: 0.7229, Adjusted R-squared: 0.7116

From Table 2 and Table 3, although not all coefficients have differences within one standard error (in training model), changes in coefficient estimates aren’t significant. Adjusted

are similar in both models. All predictors are significant in both models. No new assumption violation appears in test model. Thus, the final model is likely validated.

**Discussion**

A multiple linear regression analysis is performed to predict car prices based on the available variables. The following equation represents the final regression model:

Price = -2978.94 + 118.83 \* enginesize + 46.57 \* horsepower + 29.11\* citympg -143.3 \* highwaympg

The model achieved an R-squared value of 0.7975, indicating that approximately 79.75% of the variance in car prices can be explained by the predictor variables. The coefficients of the variables are found to be statistically significant (p < 0.05), suggesting that these variables have a significant impact on car prices.

To interpret the coefficients, with the final model, we expect 118.83 increase in price for one unit increase in enginesize, fixing all other variables. We expect 46.57 increase in price for one unit increase in horsepower, fixing all other variables. We expect 29.11 increase in price for one unit increase in citympg, fixing all other variables. We expect 143.3 decrease in price for one unit increase in highwaympg, fixing all other variables. These findings explain the research question of how these four variables influence used car price, with the present of other possible predictors. The final model is the first linear model that predicts car price and contains all four interested predictors.

However, there exists some limitations of the analysis. The final model doesn’t satisfy normality. Thus, any statistical inferences based on the model would be unreliable. Violation of these assumptions can lead to biased or inaccurate conclusions. All these assumption violations cannot be corrected by Boxcox transformation. The result of stepwise model selection might be unreliable because of these violations. There are some problematic observations that cannot be removed, which might impact estimates of final model. Also, the accuracy and reliability of the data used in the analysis is crucial. If the dataset contains errors, inconsistencies, or outliers, it can significantly impact the results.

**References**

Enis Gegic, Becir Isakovic, Dino Kečo, Zerina Mašetić, Jasmin Kevrić. (2019). *Car Price Prediction using Machine Learning Techniques*. Retrieved from <https://www-ceeol-com.myaccess.library.utoronto.ca/search/viewpdf?id=761902>

Ramgiri Siva, Adimoolam M. (2022). *Linear Regression Algorithm Based Price Prediction of Car and Accuracy Comparison with Support Vector Machine Algorithm*. Retrieved from <https://iopscience-iop-org.myaccess.library.utoronto.ca/article/10.1149/10701.12953ecst/pdf>

Manish Kumar. (2019). *Car Price Prediction Multiple Linear Regression*. Retrieved from <https://www.kaggle.com/datasets/hellbuoy/car-price-prediction>

**Appendix**

Table 4: Introduction of variables

|  |  |  |
| --- | --- | --- |
| Variable | Variable type | Description |
| enginesize | numerical | Size of car |
| horsepower | numerical | Horsepower |
| citympg | numerical | Mileage in city |
| highwaympg | numerical | Mileage on highway |
| price | numerical | Price of car |

Figure1:

A picture containing text, screenshot, diagram

Description automatically generated A picture containing text, diagram, screenshot, plot

Description automatically generated