**Modeling – Predicting the Price of Cars**

Isabelle Supandji 505400335

Kaggle Rank: 8

Kaggle R2 Score: 0.96564

Predictors Used: 11

Betas: 12

BIC: 22678.71

**Modeling – Predicting the Price of Cars**

1. **Abstract**

The main objective of this research study is to advise Chinese Automobile Company, Geely Auto as they prepare to enter the US Car Market. To aid Geely Auto’s financial and pricing strategy in this expansion, this study aims to create a model that best predicts new car prices on a set of independent variables. Based on a data set of 1500 observations of different cars and 21 predictors, this paper will delve into how different variables affect the price of cars and which have the strongest effect on car pricing.

The results of this exploration have been submitted to Kaggle under the nickname Isabelle Supandji. The final model created used 11 predictors which were selected both manually and through stepwise regression. The R-squared score for the training data set generated through R studio was 0.9671 and the R-squared score for the testing data set generated through Kaggle was 0.96564. The final predictions of this model ranked 8th in Kaggle. Through this paper, the understanding behind the predictors used will be further delved into.

1. **Introduction**

Cars have many features that make them appealing to consumers. Some cars have features, such as safety, size, and speed that justify their hefty price tag. Though cars serve the main purpose of quickly transporting people from point A to point B, there are a variety of car types that suit different lifestyles and demographic. An example of this is how a minivan that can comfortably seat 7 people better serves a large family as opposed to a smaller sedan or sports car that can only comfortably seat 4 to 5 people. For different ages, safety features like airbags and cameras often play a huge part in determining the value of a car.

Apart from size, there are many other factors that affect the price of the car, such as its origin. According to Autotrader, European cars tend to be pricier due to higher labor costs in Europe and quality of materials due to “exhaustive engineering and pricy suspension components” (Demuro) implemented in the car manufacturing process that earns these brands a higher value. However, in order to predict a car’s price, it is not enough to categorize cars by their origin country. Take for example the German automobile corporation, the Volkswagen Group owns 10 different car brands under its umbrella. On their website, their own name brand Volkswagen cars are categorized as “Volume” or everyday cars, which sell at affordable prices. The company also owns the luxury sports car brand Lamborghini, which they list as a “Premium” car, with its cheapest model selling at a base price of $200,000 dollars (Gorzelany). This dynamic is present in other car companies and countries such as Toyota and their luxury subsidiary of Lexus. The relationship between price and brand reputation comes as no surprise as we often associate the price of a car with its brand.

With this background knowledge in mind, the purpose of this paper is to examine these variables and their effect on the variability of price. For the purpose of this study, we are using the SummercarsTrain.csv data set to explore the relationships between the independent variables and car price. The data set includes 21 numerical and categorical variables and 1500 observations of different cars. Using this data set, we aim to create a linear regression model that can best predict the price of cars with the least number of predictors possible. Through the creation of this linear model, we are able to examine the relationship of car price with various variables such as horsepower, origin, and type.

1. Text

   Description automatically generated with medium confidence**Methodology**
2. Narrowing Down Numerical Variables

The first step taken in creating this model was narrowing down the potential numerical predictors to be used, through creating a correlation plot of all numerical predictors in the data set against price. Through these results, I narrowed down to a top 5 list of predictors in order, namely horsepower, weight, fuel tank capacity, engine size, and wheelbase that had the highest correlation with our target variable.

Figure 1 Correlation Plot of Numerical Predictors in Training Data Set.

I also chose to test for multicollinearity amongst the numerical and categorical predictors individually. For numerical predictors, I used the vif() function to check for values above 5. Amongst the top 5 numerical predictors, weight had an extremely high vif score above 14 and a high correlation with all top 5 numerical predictors except horsepower. When testing the vif of only horsepower and weight, dropped to 2. Due to this, I decided to create a linear model with these 5 numerical predictors except weight and a linear model with only weight and horsepower as predictors. The linear models generated similar R2 values, which lead me to decide on removing fuel tank capacity, engine size, and wheelbase, which significantly decreases the complexity of the model.

1. Choosing Categorical Variables and Creating New Variables

In choosing our categorical variables, the process was not as straightforward. By creating a linear model of all available predictors against price and running a stepwise variable selection, I found that the make variable had the strongest influence on predicting car prices. So much so that R declared that make was the only statistically significant variable. The make variable consists of the car’s brand and the car model, such as “Toyota Corolla” where Toyota is the brand and Corolla is the model. A linear model of only make against price generated an R2 value of 0.991 for the training data set. However, make is a categorical variable that has over 93 unique values, meaning that this strong model would have 93 predictors. This would make the model far too complex.

To counter the complexity, I created 8 variables named level1 to level8. These variables were created by finding the average price of each make and dividing them into 8 different tiers based on this average price. For each observation, a value of 1 meant that the observation belongs to that particular level. This significantly helped narrow down the make variable from 93 predictors to 8.

Apart from make, I also considered including the origin, type, and airbags of the car into the model as they logically made sense based on past understanding. They also generated good results in terms of increasing the R2 value. In checking for multicollinearity through the chi square test, I found that origin was highly correlated with make as it generated a p-value of less than 2.2e-16. This indicated strong multicollinearity that would defy the assumptions of multiple linear regression. Thus I chose to remove origin from the potential list of categorical predictors.

1. Creating the Model

In creating the final model, I used the lm() function using the following predictors: horsepower, weight, level1 through level8, type, and airbags with no interaction terms. The reason this was done was to reduce the complexity of the model as interaction terms would increase the number of predictors. This raw model generated an R2 score 0.9687, which is very good. However, the list of predictors was quite lengthy. To combat this, I attempted a stepwise selection of variables to weed out any insignificant predictors. However, no variables were deemed insignificant. Thus, I chose to scan the variables individually to see which predictors did not have a satisfactory p-value. In doing so, I chose to remove the type variable as it added 5 predictors to the model, some of which did not have a p-value below 0.05.

1. Diagnostics and Plots

*VIF Scores*

Table

Description automatically generatedThe VIF scores of the variables can be alarming at first glance. However, it is important note that the variables level1 through level8 are essentially 1 variable. These 8 variables were generated from the make variable and thus are highly correlated to one another. Though they appear as 8 separate variables, they better resemble a single variable with 8 factors. Thus, I chose to overlook these high VIF scores. The other variables all have good VIF scores below 5, which suggests that no multicollinearity appears in this model.

Figure 2 VIF Scores of the Final Model

Chart

Description automatically generated*Diagnostic Plots*

Figure 3 Diagnostic Plots of Final Model

The diagnostic plots showed no visible defiance against the assumptions of linear models. The residuals vs. fitted plot does not resemble a fan shape. The normal Q-Q plot does not deviate far from the diagonal line. The scale-location plot arguably stays relatively horizontal and does not resemble a fan shape. Lastly, the residuals vs. leverage plot does not show any bad leverage points. Though outliers exist, none of the have leverage above 0.4, indicating that no bad leverage points are influencing the model.

1. **Results**

The final model submitted to Kaggle had a total of 11 predictors namely horsepower, weight, level1 through level8, and airbags. The model was used to predict the prices of the observations in the testing data set. It generated an R2 square for the training data set of 0.9671 and for the testing data set of 0.96564. This R-square score is extremely high as these 11 predictors explain over 96% of the variability in car prices. Through the steps of the process, we can see that the categorical variable with the strongest influence is make, which is similar to what we hypothesized in the introduction: that the brand and model of the car highly influence the price of cars. Additionally, horsepower has the largest effect on the price among numerical predictors. This also can be explained by the higher price of luxury sports cars.

1. **Discussion**

Overall, the model created has opened up many findings in terms of the relationship of the price of cars and the various variables. We were able to see how make is the strongest influencer of car prices, indicating that the brand of the car is highly correlated with the prices of their automobile products. The high correlation between price and horsepower also indicates how the car market highly values faster cars. In addition, it reveals how faster cars can be more expensive to produce due to specialty parts, cutting-edge technology, and complex engineering (William).

Apart from the relationship of variables with the target variable of price, our findings showed how many metrics of a car are highly correlated with one another. An example of this is how weight was highly correlated with wheelbase. With my own background knowledge of cars, it is easy to see the relationship between these variables as a car with a higher wheelbase variable would be longer, thus larger and heavier.

1. **Limitations & Conclusions**

Overall, I believe that our final model does not have significant limitations. However, there are many other variables that could have been considered that would make it better applicable to our current times. With an increasing move towards clean energy, it would be interesting and helpful to see the relationship between the prices of hybrid, electric and regular gas cars with price. Especially with the presence of Tesla and Lucid Motors, who boast relatively high prices for their cars.

Overall, we can conclude that the make of the car is the most viable predictor of car prices. Our final model heavily relied on this model to predict prices. The final model was able to receive an R2 score of above 0.96, indicating accurate and very good predictions of car prices.

1. **References**

*Brands of the Volkswagen Group*. Volkswagen. (n.d.). Retrieved August 2, 2022, from https://www.volkswagenag.com/en/group/brands-and-models.html#

Demuro, D. (2015, December 10). *Why are European cars usually more expensive than American cars?* Autotrader. Retrieved August 2, 2022, from https://www.autotrader.com/car-shopping/why-are-european-cars-usually-more-expensive-than-american-cars-247052#:~:text=Labor%20costs%20in%20Europe%20aren,which%20helps%20keep%20prices%20high.

Gorzelany, J. (2022, July 19). *How much does a Lamborghini actually cost?* Motor1.com. Retrieved August 2, 2022, from https://www.motor1.com/features/344294/how-much-is-a-lamborghini/

William, J. (2021, September 13). *Why sports cars are so expensive*. Sydney Motorway. Retrieved August 2, 2022, from https://www.sydneymotorway.com.au/what-makes-sports-cars-so-expensive/