Predicting Car Resale Prices

Executive Summary & Analytical Findings

Jamie Liu, Aileen Garcia, Neil Collins, Sam Al Qarzi

Executive Summary

Our data set provides information on a collection of used car listings in Western Europe, and we are interested in performing a regression analysis to look at the relationship between our target variable (car price) and a set of variables. Statistically significant correlations resulted between the top 4 most popular car manufacturers (Opel, Ford, BMW, Volkswagen) and our chosen variables. We tested six regression models using only these statistically significant variables while also testing for multicollinearity. We selected model 6 as our final model which tested against the used car's year produced, amount of listing photos, listing duration, red exterior brown exterior, silver exterior, blue exterior, green exterior, gas fuel type, Ford, Opel, Volkswagen, BMW, all-wheel drive capability, engine capacity liters, lift back, minivan type, SUV, minibus, van, and odometer value.

Our regression model indicated the following variables **positively** impact the resale price of a used car: $3,945 for all-wheel drive, $3,796 for a minibus, $1,816 for a lift-back car, $1,673 for a van, $1,430 for brown exterior, $1,071 for an SUV. Variables that **negatively** impacted the resale of a car and most decrease car selling price include the following: -$3,384 for an Opel Car, -$3,277 for a Ford, -$1,994 for a Volkswagen.

Business Understanding

The business problem that we are focusing on is how to determine what factors most affect the sale price of a used car. Recognizing which factors most determine the price of a used car will dictate what needs to be taken in consideration for pricing and ensure maximum return on every used car sale. The question we want to answer with our analysis is '*What factors significantly influence the sale price of used cars and can be used in a price predicting model?'*

Determining the most influential variables of used cars will allow dealerships or private sellers to effectively price used cars depending on the predictive variables each car has. Sellers will benefit from this and will have direct information on pricing cars. This analysis will also benefit individual sellers of their cars, who may not have experience with pricing cars and have less access to car sales data. The variables we determine will likely correlate to higher prices, which will help with determining the highest price possible for a used car.

Data Understanding and Preparations

Our data contains collective records of sold used cars on various websites across Western Europe, obtained from Kaggle. The dataset was framed with approximately 40 thousand observations and 30 variables. Each observation was mapped with various features that a particular car had whether tangible features such as (e.g color, transmission type, etc…) or intangible ones (e.g engine power, brand, engine capacity, etc...). Our main objective of exploring this dataset as mentioned above, was to examine any impact of those diverse features in determining the price of a particular car. Initially, we began by checking for missing, null, or duplicate values, which fortunately were just a few. We also dropped 6 columns that contained no data. Then, we check data types and set up each variable to its correct data type.

As we decided to carry out our preliminary statistical testing methods, we were presented with our first challenge which required us to transform the majority of those features to numerical parameters as they were in their raw categorical state. Moreover, each single feature enfolded so many labels that produced even a harder challenge for us to what decoding method we should use. We thought of several decoding techniques such ordinal coding which labels each category with an integer, but as we are given so many categories in each variable, we feared that it might cloud our regression results in the modeling stage. We decided on one-hot encoding or "dummy variables" which transforms each category into its own column with binary instances 0-1. As a result, we exceeded the original count of the variables by 12 more; however, with almost all variables as numeric, we were enabled to test for correlations and multicollinearity between all independent variables and our dependent variable (car price). We also decided to exclude the model name variable altogether as it had 900+ unique values and decided to instead focus on a broader analysis of a few car brands.

Our tests indicated much lower correlations between the dependent variable and the majority of the independent ones. This favored us with an opportunity to narrow down the dataset and focus our attention on the most decisive variables. Nevertheless, as we performed our visualizations to test for linear relations between the price and the remaining variables, we noticed that this dataset was filled with outliers that muddied our results. Therefore, as we narrowed down the dataset to a few informative variables, we considered focusing only on the most four popular manufacturers in the dataset (Volkswagen, Opel, BMW, Ford). This last step of data preparation framed out our final dataset to perform our regression models.

Modeling

When conducting our Multiple Regression Analyses on our narrowed down data set of 12,178 rows, we began with two approaches. The first approach began with a correlation analysis. This tested for variables that had a significant correlation to our target variable, price_usd. The variables with a significant correlation value were each put into the first regression model as predictor variables. After the first model was created, we analyzed the p-values and removed variables with p-values greater than 0.05 which indicated that they were not statistically significant for our model. The remaining variables were inputted to a new model, and this process was repeated until only significant p-values remained. As less variables were dropped and we approached our final model for Approach 1, we ran a VIF test to drop variables that had multicollinearity and ran stepAIC tests to further drop any variables that were deemed insignificant and created a new model with the remaining variables. This approach resulted in 12 regression models, with model 9 having the highest R^2 score of .72.

The second approach began with a linear regression using all of the predictor variables to predict price_usd. A stepAIC test was conducted to narrow down variables by dropping the ones deemed insignificant by the test. A second model was created using the variables from the stepAIC test and the p-values were examined. Any variables with p-values above 0.05 were dropped and the remaining variables were put into a new model. We used VIF to test for multicollinearity, and variables indicating high multicollinearity with scores above 5 were dropped from the model. We continued the process of running stepAIC tests and dropping insignificant p-values to create a new model until our model no longer improved. We decided to move forward with this approach, as it produced 6 regression models which resulted in the highest overall R^2 and lowest AIC scores compared to Approach 1. Although model 2 had the lowest AIC score of 229699.3 and highest R^2 value of .748, we concluded that the differences in scores were too minimal and insignificant to use as an indicator of a better model. We ultimately selected model 6 as our final model with an AIC score of 230140.2 and R^2 value of .739, as model 2 contained insignificant p-values and model 6 was the improved version of the model that accounted for multicollinearity and all possible insignificant p-values.

Based on the coefficient values, our final regression model indicated that these 11 characteristics of a car had a significant impact resulting in the following increase in resale prices: $457 per newer year produced, $72.77 per photo included in the listing, $1.47 per additional day of the listing duration, $408.60 for a red car, $1,430 for a brown car, $3,945 for a car with all drive terrain capability, $560.20 per increase in liters of engine capacity, $1,816 for a liftback car, $730.60 for a minivan, $1,071 for a SUV, $3,796 for a minibus, or $1,673 for a van. These 8 characteristics of a car had a significant impact resulting in the following decrease in resale prices: -$0.005 per one unit of odometer value, -$812.10 for a silver car, -$365.80 for a blue car, -$389.40 for a green car, -$709.10 for a car that has a gasoline engine fuel type, -$3,384 for an Opel, -$3,277 for a Ford, or -$1,994 for a Volkswagen car.

Conclusion and Discussion

Our results can be utilized heavily in the resale industry; as a partner firm could use our model to predict the optimal return on investment. If a vehicle has a lower resale value (such as Opel Cars, which tend to be worth approximately $3,384 less), the firm would know to spend less money to acquire the vehicle for resale. On the other hand, brown SUVs might cause the firm to pay more for the vehicle as it has two qualities that positively affect the resale price ($1,430 for brown exterior, $1,071 for an SUV). Theoretically, this should limit overspending and maximize profits for the company beyond what might be achieved via reference materials such as Kelly Blue Book when deployed across multiple locations. Locations may even choose to offer targeted promotions to entice owners to trade in cars that historically lead to better return on investment, meaning that they might offer perks such as free oil changes or cash incentives to those who trade in high value vehicles.

However, our model is not perfect, and there are a number of different factors that could significantly shift our predictions. An R squared value of .74, while usable, is not ideal when making business decisions, and in the long term we would want to aim to have an R squared value upwards of .90. A number of different factors within our model resulted in minimal shifts in the overall model's AIC value as well as in the R squared value. While we removed features that were statistically insignificant and aimed to improve our AIC and R values as much as possible, some of the features that were removed could play a larger role in price shifts than anticipated. With a data set this large, there is also increased likelihood of slight variance and error due to outliers.

Furthermore, our model is limited by the data itself, as a number of variables such as time of the year as well as the fact that our team had to filter the data down to only the top 4 highest occurring car manufacturers as utilizing the full data set provided very few statistically significant insights. However, that concept in itself could serve as an insight as it points towards the need for more in depth, targeted analysis on a per-manufacturer basis, or even on a per-bucket basis (bucketing by style or "tier" of car as could be defined by more analysis utilizing a method such as k means clustering). But without further analysis, our team only managed to find significant correlations within a limited sample size. Because of this, we do not know the full reaching effects of other car brands, and we do not know if the positive features and negative features hold true across the board. While brown vehicles may sell better for the four tested manufacturers, it may actually have a negative impact on resale value given an untested manufacturer, something that has to be taken into account when examining our model.

Finally, the lack of any locational data in our base dataset heavily limits personalization of data application at different used car dealerships. While the data is confined to Western Europe, there is a broad range of socioeconomic clusters within that broad area, and while one vehicle may sell for a higher amount or in a higher quantity in more affluent regions, it may fall short or be too expensive in more disadvantaged regions. As such, further data collection will need to be conducted with more complete features so that return on investment can be truly optimized.

In conclusion, while our model provides a statistically significant correlation between certain key variables in the top 4 highest reselling manufacturer groups and resale price, there are very real limitations on our model, and while it can be used as a general guiding principle at this point, the analysis is not far reaching enough to provide the near-exact precision that would truly maximize the firm’s return on investment.