Prediction of Car Prices in the Used Car Dealers’ Industry

Student Name

Name of Institution

Prediction of Car Prices in the Used Car Dealers’ Industry

**Introduction**

The used car dealers’ industry has showcased significant growth with the entry of online dealerships such as Carvana cars. However, the industry remains highly fragmented with small family-owned non-employing dealerships making up more than 80% of the dealerships and just a little more than half of the total sales (Ibis World, 2021). Due to the high fragmentation of the industry, vehicle pricing is unstandardized across dealers with variant prices for used cars of the same make, model, and condition; consequently, many dealerships misprice their inventory, which has negative implications for sales and customer satisfaction. In the case of overpriced cars, the inventory period is increased as customers are unwilling to pay the marked-up price, reducing the overall profitability of the company. Additionally, customers sold overpriced vehicles will be unsatisfied with their purchase, which further hurts the company’s profitability in the long-run. In the event of vehicle underpricing, the dealer losses out on profits.

Therefore, developing pricing standards for vehicles based on attributes such as year of manufacture, mileage, and body type among other variables would reduce the pricing inefficiencies in the industry. The large volume of vehicle attribute and price data available makes it possible to determine the typical price of a vehicle of certain attributes and to streamline the used car purchase, listing, and sales process. This will in turn improve dealerships’ inventory turnover, profitability, and customer retention. Additionally, the findings of the current analysis can be used to determine the price floor for a given vehicle, which will leverage the operating efficiency of firms in the industry and maintain industry margins as customers are limited in the extent to which they can shop around for a car of certain make and condition. The current analysis will build the groundwork for the development of pricing guidelines for the used car industry through descriptive, correlation and regression analyses of a dataset detailing the prices and attributes of used cars on the Carvana cars website. Descriptive analyses will be used to compare the price distribution of a variety of used cars. Correlation analysis will be used to describe the relationship between car prices and car attributes measured on a continuous scale such as miles per gallon (mpg). Multiple regression analysis will be used to predict car prices using a set of quantitative and categorical dummy variables.

**Data**

The data was retrieved from the Data World website (Data World, 2021). The dataset contains 196 vehicle records across 23 variables that were listed on the Carvana cars website in 2020 and were mined by Crawl Feeds then uploaded to Data World (Data World, 2021). The objective of the analysis was to predict the selling price of used cars based on the car’s attributes. A key advantage of data retrieved from the Carvana cars website is the synonymity of listing price and selling price as unlike other dealerships, Carvana prices its cars efficiently and does not negotiate its prices. Therefore, the listing prices are reliable estimates of the selling prices of used cars. The downloaded data comprised of 330 records and 725 car attributes. However, not all vehicles had complete records and some attributes such as vehicle identification number were irrelevant to the analysis. All irrelevant attributes were deleted from the dataset as were records with missing values leaving a total of 293 observations across 6 quantitative and 1 categorical variable. The quantitative variables include listing price, year of manufacture, mileage, highway mpg, curb weight, and manufacturer’s suggested retail price (MSRP) while body type was retained as the sole categorical variable. For the purposes of the regression analysis, body type was recoded into a dummy variable with four categories – Sedan, SUV and Hatchback – increasing the number of variables to 9.

**Analysis and Results**

**Descriptive Analyses**

Beginning with the descriptive statistics of all cars included in the dataset presented in Table 1 below, the average listing price of all vehicles was $20,513. The standard deviation of listing prices was 2,640.76, which is indicative of fairly high variation around the mean. Vehicle prices ranged from $12,990 and $24,990. The year of manufacture ranged from 2010 to 2021. Since the average year of manufacture has less quantitative meaning than the median year of manufacture, the median value of 2017 is reported: half of the cars were manufactured after 2017. Car mileage ranged from 1,994 to 100,259 miles with an average of 45,670 miles. The wide range and high standard deviation of 21,668 are indicative of wide dispersion of mileage observations. Highway mpg ranged from 22 to 42 mpg with an average of 32.99mpg. The standard deviation was fairly low at 4.71, indicating low dispersion of values around the mean. Vehicle curb weight ranged from 2,018 to 4,913 pounds with an average of 3,316 pounds. The standard deviation was moderately high at 509.05 indicating moderate dispersion around the mean.

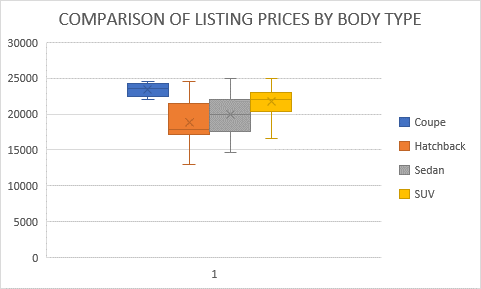
The average manufacturers’ suggested retail price was $26,161, which was $6,161 less than the average selling price of the used cars, with values ranging from $13,385 to $94,525. The variable’s standard deviation (7,935) was considerably higher than the standard deviation of listing prices indicating wider dispersion of MSRP. Of the cars included in the dataset, 56% were sedans, 35% were SUVs, 7% were hatchbacks, and 2% were coupes. All of the described variables are to be used in a regression model to predict the prices of used cars. The range statistics provide useful information on the range of car attribute values for which the regression equation is applicable: for instance, if the prediction equation includes mileage, dealerships should not use the model to predict the prices of used cars with mileages outside the 1,994 and 100,259 range.

Table 1

*Descriptive Statistics of Cars Included in the Sample*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Price | Year | Mileage | Mpg highway | Curb weight | MSRP |
| Mean | 20,513.29 | 2,016.75 | 45,670.12 | 32.99 | 3,316.44 | 26,161.12 |
| Standard Deviation | 2,640.76 | 1.97 | 21,668.03 | 4.71 | 509.05 | 7,935.24 |
| Minimum | 12,990 | 2,010.00 | 1,994.00 | 22.00 | 2,018.00 | 13,385.00 |
| Maximum | 24,990 | 2,021 | 100,259 | 42.00 | 4,913.00 | 94,525.00 |

The difference in car prices across the four body type categories are compared in the side-by-side boxplots in Figure 1 below. While the minimum and median price for coupes was distinctly higher than for the other body types, the number of coupes in the sample was limited with a total of 5 observations limiting the reliability of the median estimate. For the most part, the hatchback, sedan, and SUV price distributions were overlapping. However, the median and minimum price for SUVs is somewhat higher while hatchbacks have the lowest median and minimum prices.



*Figure 1. Comparison of listing prices by body type*

**Correlation Analysis**

Correlation analysis of the quantitative and dummy coded variables was conducted and the results are summarized in Table 2 below. Price was weakly to moderately correlated with the 9 variables as the values of the correlation coefficients range between 0.15 and 0.44. The variables that were most strongly correlated with price were MSRP (0.44), curb weight (0.43), SUV (0.32), highway mpg (-0.32), and year of manufacture (0.30). Price is positively correlated with MSRP, curb weight, year of manufacture, and SUV: all else equal, an increase in any of the variables is associated with an increase in price. Price was negatively correlated with highway mpg; thus, an increase in highway mpg is associated with a decrease in price. While the correlation coefficients are low from an effect size perspective, the sample size is large and all the coefficients are statistically significant at an alpha level of 0.01, as determined from the critical value table of Pearson’s correlation coefficients. Therefore, all the variables can be used in the regression model.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *Price* | *Year* | *Mileage* | *mpg highway* | *curb weight* | *msrp* | *Sedan* | *SUV* | *Hatchback* |
| Year | 0.30 | 1.00 |  |  |  |  |  |  |  |
| Mileage | -0.20 | -0.45 | 1.00 |  |  |  |  |  |  |
| mpg highway | -0.32 | 0.27 | -0.31 | 1.00 |  |  |  |  |  |
| curb weight | 0.43 | -0.31 | 0.43 | -0.75 | 1.00 |  |  |  |  |
| msrp | 0.44 | -0.43 | 0.43 | -0.57 | 0.74 | 1.00 |  |  |  |
| Sedan | -0.26 | 0.01 | -0.02 | 0.56 | -0.33 | -0.15 | 1.00 |  |  |
| SUV | 0.32 | 0.06 | 0.06 | -0.62 | 0.47 | 0.22 | -0.83 | 1.00 |  |
| Hatchback | -0.17 | -0.05 | -0.11 | 0.12 | -0.27 | -0.16 | -0.31 | -0.20 | 1.00 |

**Regression Analysis**

Having ascertained the significance of the correlation between price and proposed predictors, all categorical variables were included in the multiple regression model and the results are summarized in Table 3 below. The regression was highly significant (F = 69.18, p < 0.0001). The R squared and adjusted R squared statistics were equal to 0.66 and 0.65, respectively. Combined, the predictors explain 65% of the variation in listing prices after adjusting for the inclusion of multiple independent variables in the model. Individually, all variables other than highway mpg were significant at an alpha level of 0.05. The non-significance of highway mpg despite its relatively high correlation with car price suggests that the variable explains the same component of price that is explained by another predictor variable. From the correlation coefficients table, highway mpg was strongly correlated with curb weight; it is, therefore, likely that the explanatory value of the two variables overlap.

Table 3

*Preliminary Regression Model Coefficients*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -1321383 | 113860 | -12 | 0 |
| Year | 662.07 | 56.49 | 11.72 | 0.00 |
| Mileage | -0.04 | 0.01 | -8.64 | 0.00 |
| mpg highway | 62.33 | 34.84 | 1.79 | 0.07 |
| curb weight | 1.63 | 0.35 | 4.61 | 0.00 |
| msrp | 0.20 | 0.02 | 10.95 | 0.00 |
| Sedan | -4352.20 | 726.15 | -5.99 | 0.00 |
| SUV | -3625.95 | 728.09 | -4.98 | 0.00 |
| Hatchback | -4228.49 | 795.70 | -5.31 | 0.00 |

The regression was rerun while excluding highway mpg to arrive at a more reliable prediction model comprising of individually significant variables only. The results are summarized in Table 4 below. The overall regression was highly significantly (F = 78.00, p < 0.001) with the variables explaining an adjusted 65% of the variation in listing prices. All the variables included were individually significant at an alpha level of 0.01. The prediction model is given as:

Price = -1,343,208 + 674.37 \* year of manufacture – 0.04\*mileage + 1.34 \* curb weight + 0.20 \* MSRP – 4,146 \* Sedan – 3,665.73 \* SUV – 4,151.19 \* Hatchback

Table 4.

*Final Regression Model Coefficients*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | -1343208.72 | 113642.05 | -11.82 | 0.00 |
| Year | 674.37 | 56.29 | 11.98 | 0.00 |
| Mileage | -0.04 | 0.01 | -8.60 | 0.00 |
| curb weight | 1.34 | 0.32 | 4.25 | 0.00 |
| MSRP | 0.20 | 0.02 | 10.77 | 0.00 |
| Sedan | -4146.04 | 719.72 | -5.76 | 0.00 |
| SUV | -3665.73 | 730.57 | -5.02 | 0.00 |
| Hatchback | -4151.19 | 797.60 | -5.20 | 0.00 |

From the regression equation, moving the year of manufacture one year forward is associated with a $674.37 increase in price while holding all other variables constant: on average, cars with a given body type, mileage, curb weight, and MSRP that were manufactured in 2021 cost $674 more than cars with the same attributes but manufactured in 2020. A unit increase in mileage was associated with a 0.04 decrease in price while holding all other variables constant. A unit increase in curb weight and MSRP were associated with a 1.34 and 0.20 increase in price, respectively, while holding all other variables constant. The dummy variables were created with reference to coupes: the Sedan variable was coded 1 for Sedans and zero for all other cars, SUV was coded 1 for SUVs and 0 for all other cars, and hatchback was coded 1 for hatchbacks and 0 for all other cars; thus, 0 represents coupes as the only remaining category. On average, Sedans were priced $4,146 less than non-sedans, SUVs were 3,665 less expensive than non-SUVs, and hatchbacks were $4,151 less expensive than non-hatchbacks when holding all other variables constant.

**Conclusion**

The primary objective of the current analysis was to develop a price prediction model that could be deployed in car pricing in the used car dealerships industry. The regression model generated in the analysis was highly significant as were the three quantitative variables – year, mileage, curb weight, and MSRP – and dummy coded body type variable, which demonstrates the usefulness of the attributes in predicting car prices. The caveats of the current analysis include the relatively small sample size that is unrepresentative of the population of used cars sold in the US and the failure to control for interactions between the dummy and quantitative variables. If the regression model is to be used in car pricing for dealerships across the country, a more representative sample with wider ranges for the predictor variables should be used to develop the regression model. It is also unlikely that used car prices are standard across the country given the differences in car supply and demand as well as the disposable income of residents: it is, therefore, advisable to develop regression models that apply to a state or a region with similar industry conditions.

References

Data World (2021). Carvana cars details dataset. Retrieved from:

https://data.world/crawlfeeds/carvana-cars-details-dataset/workspace/file?filename=carvana\_data\_2.json

Ibis World (2021). Used car dealers’ industry in the US - market research report. Retrieved from:

https://www.ibisworld.com/united-states/market-research-reports/used-car-dealers-industry/