Final Report

GM car price model

Izaguirre, Diego

2023

Contents

[Appendix A 8](#_Toc136884639)

[Appendix B 9](#_Toc136884640)

[Appendix C 10](#_Toc136884641)

[Appendix D 11](#_Toc136884642)

[Appendix E 12](#_Toc136884643)

[Appendix F 13](#_Toc136884644)

**Introduction**

A linear regression model will be built to predict the price of 2005 year General Motors cars based on data gathered from Kelly Blue Book. The dataset has 804 observations with 10 independent features. There were originally 11 but one called “Trim” was removed due to redundancy with another variable called “Model”. The variables in the dataset are Mileage, Make, Model, Type, Cylinder, Liter, Doors, Cruise, Sound, and Leather. Mileage designates the number of miles the car has been driven. Make is the manufacturer of the car. Model is the specific model of each car manufacturer. Type is the type of body the car has. Cylinder is the number of cylinders in the engine. Liter is a measure of engine volume. Doors is the number of doors. Cruise designates whether the car has cruise control or not. Sound indicates whether the car has a sound system. Finally, leather indicates if the car has a leather interior or not. Based on these dataset features the model will attempt to predict price as the dependent variable. The data will use an 80/20 split for training and testing respectively. The full model, final model, and 2 predictions will be created using the training set. The testing set will be used to validate the model’s performance. My hypothesis for this model is that the most important predictors of price will be Type and Mileage because mileage is a good indicator of the wear on the car and different types of cars are bigger or considered more desirable than others.

**Analysis, Results, & Findings**

The first step is to explore the data. We will start with some descriptive statistics on the numerical variables to get a preliminary sense about the distributions (see appendix A, Figure A1). For the categorical variables I opted to get the frequencies of the categories instead (see Figures A2, A3, A4). Looking at the number of categories for some of these variables it is apparent that there is a problem with dimensionality. If we were to make dummy variables for all these variables, we would not have enough observations to make a proper model. So, to make an accurate model I opted to drop some of the features which would allow us to work with the observations on hand. Make and Model were dropped in favor of Type because those two features already fall under the General Motors company brand. Since this model is already being trained on GM cars, it is preferable to focus on what body type these cars have rather than whether it is a Chevrolet or a Saturn. This simultaneously simplifies the model and helps alleviate this issue with dimensionality. I would recommend that either a separate model be made to study the impact of the removed features or obtain more observations in the future.

Looking at the distributions of the numerical variables (see appendix B Figures B1, B2, B3 for histograms) we see that distribution for price is right skewed. To correct this a log transformation was applied to price which shifted the distribution close enough to a normal distribution (Figure B4 for transformed histogram). The distribution for Mileage was normal and the distribution for Liter was multimodal having a class imbalance since it indicates that most of the engine volumes were 2.1 or 3.9 liters.

The features Type and Doors need to be converted into dummy variables before a full model can be fit. Type has 5 categories and Doors has 2. Therefore, we need 4 dummy variables for Type and 1 for door which will be designated as follows:

Type\_Converti 1=true 0=false

Type\_Coup 1= true 0=false

Type\_Sedan 1=true 0=false

Type\_Wagon 1=true 0=false

Hatchbac will be used as base level.

d\_doors 1 = 4 doors 0 = 2 doors

So, after the creation of the dummy variables, we are left with 13 variables in the dataset to fit the full model along with VIF statistics (see Figure B5 for preliminary full model). After fitting it though there is an issue with the Type\_Coup and d\_doors variables. Type\_Coup conflicts with the d\_doors variable because a coup already designates a car with two doors. Therefore, that makes the parameter estimate for Coup biased. To preserve the integrity of the Type parameter I opted to remove d\_doors. In terms of multicollinearity Liter and Cylinder are colinear. After making a correlation matrix with the newly created variables Liter and Cylinder appear to be collinear with an R value above 0.9 (see Figure B6). Both also have VIF values of above 10, therefore, to avoid inaccuracies with the parameter estimates one of the variables Cylinder was removed because according to the standard estimate values Liter held greater influence in the model.

After removing those variables, we see we no longer have a multicollinearity issue nor redundancy issues and the full model runs fine (see appendix C Figure C1 for full model). The full model was also run with influence and outlier statistics to identify and possibly remove influential points. After running the outlier and influential points statistics on the full model there were several outliers at observations 341-250 and 650 (see Figure C2). These points were skewing the distribution of price, so I opted to remove them to get closer to a normal distribution. After removing them there was an increase in the adj-R^2 as well as substantial increases in some of the parameter estimates.

Once we have an adequate full model a model selection procedure must be run to find the best model for this data. For this project two model selection procedures were run: Backward and Adj-R^2 selection. Backward model selection involves starting with a full model and gradually removing variables that may be statistically insignificant. Adj-R^2 selection involves choosing the combination of variables that yield the highest adj-R^2 value. So, after running both selections (see Figures C3 & C4) the best model (see appendix D for final model) maximized the adj-R^2 while minimizing the complexity of the model. Type\_Coup and Sound were removed in the final model indicating that these variables were not statistically significant. Since we applied a log transformation the coefficients in the final model need to be retransformed in the final model equation using the following equation (exp(x)-1) which yields the correct coefficients. To get the percent increase of these coefficients simply multiply them by 100. Therefore, the final model equation is:

log(lnPrice)= 7767.248503 + Mileage\* -8.37996489e^-6 + Liter\*0.2421577645 + Cruise\*0.2224413919 + Leather\*0.1341347212 +Type\_Converti\*1.51559662 + Type\_Sedan\*0.1686275195 + Type\_Wagon\*0.6725509903

Now that we have a final model it is important to check assumptions using residual plots (see appendix E Figures E1, E2, E3, E4). So, we can assume constant variance because the points are all evenly spread out on the plot. We can also assume independence because the points are spread out along the center line. We can also assume linearity because the points are spread out in a straight line along the center line. Normality is also assumed because on the QQ plot the plotted line is approximately straight. Since we meet all our assumptions, we can go ahead and do a model analysis using GOF, R coefficient, standard estimate, and RMSE to determine model efficiency and make the predictions.

The model is a good fit for the data because with an F-Value of 214.39 and with an alpha of 0.05 and a p value of <0.001, we can reject the null hypothesis and conclude the model is a good fit for the data. The adj-R^2 value is 0.70, which means that this model explains about 70% of the variance in the data. The other 30% may be explained by some of the variables that we left out or perhaps some other unexplored variables. Based on the standard estimates, the top 3 predictors of the price of a GM car appear to be Liter, Type\_Converti, and Type\_Wagon. This indicates that engine size and car type are the 2 biggest indicators of how expensive a car will be. Looking at the RMSE, which has a value of 0.21845, the error on this model is low when considered at the scale of the log transformation of the independent variable. It is also important to go through the parameter estimates which give an idea of how each feature impacts the dependent variables. Based on the parameter estimates for Mileage they indicate that for 1 unit increase in Mileage there will be a -8.37996489e^-4% decrease in price. This means that for every mile the car is driven, the price will decrease slightly. For every 1 unit increase in Liter, there will be a 24.21577645% increase in price. So as engine size increase so does price. If the car has cruise control, there will be a 22.24413919% increase in price. Else, there will be no increase in price. If the car has a leather interior there will be a 13.41347212% increase in price. Else there will be no increase. If the car is a convertible there will be a 151.559662% increase in price. If the car is a sedan there will be a 16.86275195% increase in price. Finally, if the car is a wagon there will be a 67.25509903% increase in price. These last three indicate that convertibles and wagons fetch a higher price than sedans.

The two predictions will be done using the training set, which will then be validated using the testing set. Prediction one will be done using the following values: 34210 for mileage, 4.6 for liter, 1 for cruise, 1 for leather, 0 for type\_converti, 0 for type\_sedan, and 1 for type\_wagon. Prediction two will be done with the following values: 10023 for mileage, 3.8 for liter, 1 for cruise, 0 for leather, 1 for type\_converti, 0 for type\_sedan, and 0 for type\_wagon. Using these data lines, we get predictions for the price of these cars at the 95% confidence level (see appendix F Figure F1 for predictions). Of course, since we did a log transformation on the dependent variable these predictions need to be transformed as follows…

Prediction 1: $13489.94 with 95% certainty that the prediction lies between $12482.71 and $14578.45.

Prediction 2: $18457.01 with 95% certainty that the prediction lies within $16888.69 and $20170.97.

Now that we have the predictions the testing set can be used to validate them. So, using the testing set predictions are run on the current lines in the dataset (see Figure F2 for testing predictions) which then can be cross validated with the actual values. From that the validation statistics can be calculated through which the difference in performance between testing and training can be seen (see Figure F3 for validation statistics). The cross-validated R^2 is -0.15 which indicates that there was an absolute 0.15 difference between the two, in which the testing R^2 being higher than the training. This means that there is no overfitting problem, but there may be an underfitting problem since testing performed better than training. RMSE has a value of 0.22444 which is the standard deviation of the difference between actual and predicted values. This value is small in comparison to the scale of the data; therefore, we have a small error rate. The MAE has a value of 0.17305 is a measure of the average of all the residuals. Again, this number is small in comparison to the scale of the data, therefore we have a small average error rate. The adj-R^2 which is 0.8553 which indicates that the model was able to explain 85% of the variance in the training data.

|  |  |
| --- | --- |
| Training | Testing |
| * Adj-R^2: 0.7053 * R^2: 0.7020 * RMSE: 0.21845 * GOF: ok * Residuals: ok | * Adj-R^2: 0.8553 * R^2: 0.8566 * RMSE: 0.22444 * MAE: 0.17305 * CV-R^2: abs(-0.15) |

In conclusion, the linear regression model for predicting car prices has demonstrated excellent performance. After cross validation, the model has been shown to be fairly accurate at making predictions, with low RMSE and MAE scores. While there is slightly better performance on testing, which may indicate underfitting, the difference is not staggering. Overall, I am confident in the model's ability to accurately predict car prices. Though there is room for improvement. For example, I wish I could have kept more of the initial features in the model. I would like to make another model using more data or retry the model using the model of the car instead of type to see if there is improvement in the model.

# A screenshot of a computer screen Description automatically generated with low confidenceAppendix A

A screenshot of a computer

Description automatically generated with low confidenceA picture containing text, screenshot, font, number

Description automatically generatedA screenshot of a computer screen

Description automatically generated with low confidence

Figure A2

Figure A3

Figure A4

Figure A1

# Appendix B

A picture containing text, diagram, plot, line

Description automatically generatedA picture containing text, diagram, line, plot

Description automatically generatedA screenshot of a graph

Description automatically generated with medium confidenceA picture containing text, diagram, screenshot, line

Description automatically generated

Figure B4

Figure B3

Figure B2

Figure B1

# A screenshot of a computer Description automatically generated with low confidenceAppendix C

Figure C2

A screenshot of a computer

Description automatically generated with low confidence

Figure C1

A screenshot of a computer

Description automatically generated with medium confidence

Figure C4

A screenshot of a computer

Description automatically generated with low confidence

Figure C3

# Appendix D

A screenshot of a computer

Description automatically generated with low confidence

Figure D1

# A picture containing text, line, plot, diagram Description automatically generatedAppendix E

Figure E3

A screen shot of a graph

Description automatically generated with low confidenceA screen shot of a computer screen

Description automatically generated with low confidenceA picture containing text, diagram, screenshot, line

Description automatically generated

Figure E2

Figure E4

Figure E1

# Appendix F

A screenshot of a computer

Description automatically generated with medium confidenceA screenshot of a computer

Description automatically generated with low confidenceA picture containing text, screenshot, font, line

Description automatically generated

Figure F2

Figure F3

Figure F1