Car MSRP Prediction with Multiple Linear Regression

Beshoy Sadek, Sherleen Lee, Xiao Tong Gan, and Somar Dakak

Computer Information Systems, Cal Poly Pomona

CIS 4321: Data Mining

Dr. Mehrdad Koohikamali

May 14, 2021

Research questions and objectives

The purpose of our group project is to figure out what features have an influence on a vehicle's Manufacturer's Suggested Retail Price (MSRP) in the Car Features and MSRP dataset. By doing research and examining data, we will be able to predict an appropriate MSRP given a specific set of features, and understand how those features affect the MSRP.

Importance of your project

The problem for the manufacturers is to find a balance between what consumers want and pricing the particular vehicle. Knowing which features and specs make the MSRP and which features largely determine the price of the car, allows manufacturers to produce and sell cars that meet market demands.

Data and methodology

Our dataset has 16 variables which are divided into 8 numerical and 8 categorical. The numerical variables are as follows: Year, Engine HP, Engine Cylinders, Number of Doors, highway mpg, city mpg, Popularity, and MSRP. The categorical variables are as follows: Make, Model, Engine Fuel Type, Transmission type, Driven Wheels, Market Category, Vehicle Size, and Vehicle Style. The categorical variables such as Transmission type, Vehicle Size, and Vehicle Style, are transformed to dummy variables to see how that will reflect on the overall analysis. We all agreed that multiple linear regression applies to our project because it allows us to predict one variable based on the known information about other variables and examine how multiple independent variables are related to one dependent variable. The dependent variable we are predicting is the MSRP. After cleaning and normalizing the data, we used recursive model elimination (RFE) to compare models with different sizes and selected Model 8 with the highest adjusted R^2 of 0.75. Then, we compared the performance of the model using the training and test sets. We used the following features for our analysis: Year, Horsepower, Fuel_type_flex-fuel (unleaded/E85), Fuel_type_premium unleaded (recommended), Fuel_type_regular unleaded, Transmission_type_MANUAL, Driven_wheels_rear wheel drive, and Vehicle_size_Compact.

Analysis and results

We were able to obtain unnormalized coefficients for the best model, which allows us to predict the MSRP of a car with a 75% accuracy. We predicted the MSRP of a 2011 car with the following features: 335 hp, premium unleaded, manual, rear-wheel drive, and compact, to be $46,549.31.

Findings and conclusion

We were able to predict a car's MSRP with a given set of features. This allows manufacturers to adjust features according to the preferred MSRP, set budgets early, decide how to market the car, and have a better understanding and control over the features contributing to the MSRP. Consumers can estimate the MSRP for a set of features that they want in a car. Limitations include the lack of numerical variables, unknown values which needed to be dropped individually, and irrelevant variables such as market category, vehicle style, and popularity. Based on the 75% accuracy that the RFE method yielded, we concluded that the RFE is not the best model to predict a car's MSRP. We hope to improve our results with a dataset that has more complete data and more relevant variables. For future projects, we may choose to use a different model for our prediction.

Extra Credit

For the extra credit portion, we chose classification. We wanted to classify the vehicle's engine cylinders to either be 4 or 8. This might be helpful for car shopping websites to classify a user's vehicle listing with a missing engine cylinder value. We used three different classification models and validated them via confusion matrix and cross-validation. The results were extremely similar; all three models yielded an accuracy score of 98% as well as 99% score for the Area Under the Receiver Operating Characteristic (AUROC). The three models are Decision Tree Classifier, Random Forest Classifier, and Naïve Bayes. After looking at the plotted AUROC, we are confident that any of these Classification models can perform excellently to Classify a vehicle's engine cylinders.