

Predictive Modeling for the Manufacturer's Suggested Retail Price (MSRP) of New Cars (2019)

Prepared by:

**Colin Horgan**

University of Delaware, STAT 613

Applied Multivariate Analysis

**Introduction**

In 2019, over 17 million cars were sold in the United States [1]. Although many of these cars likely have not been on the road much this year due to the pandemic, transportation remains a critical component of everyday life for nearly every American. The average American family spends 16 cents of every dollar on transportation, with nearly 93% of that payout being put towards operating a car – the second biggest household expense besides home ownership [2]. That Americans shell out this much money for cars is hardly surprising. Cars have long been both a status symbol and a tremendous convenience since their invention. But where is our money really going and what really determines the cost of a new car? Does spending more money mean that we get a “better” car, or does it mean that we get a car that broadcasts higher social status? Or… are those possibilities one in the same? The goal of this analysis is to analyze the features of cars sold during 2019 to develop a predictive model in order to understand what *really* determines the price of a new car.

**Data Description**

The dataset we will use for this analysis contains (prior to preprocessing) data concerning 56 features of 32,316 cars sold during 2019. The dataset was obtained through the open-source data-sharing website Kaggle.com and the link to the raw data download can be found under the **data-sources** section. The 56 features contained within our dataset provide a rich description of each car sold, encapsulating MSRP, engine type, drivetrain, backseat airbags, miles under warranty, years under warranty, horsepower, and passenger volume – just to name a few. The complete data dictionary is provided below. Features are described where needed.

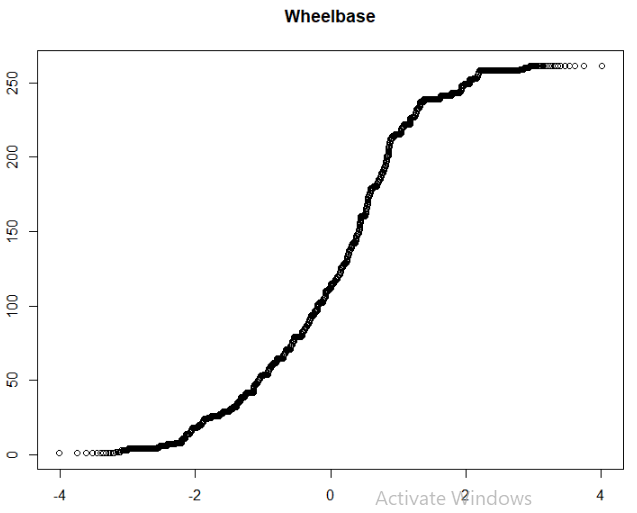
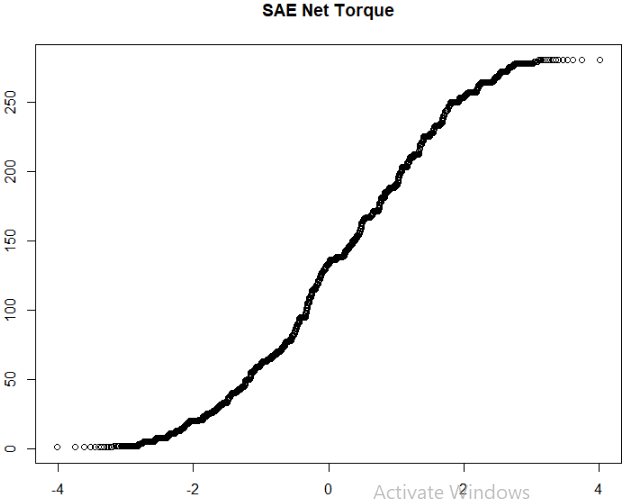
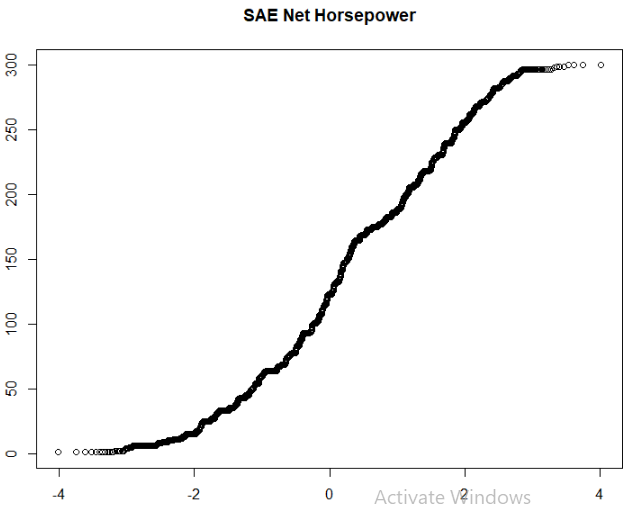
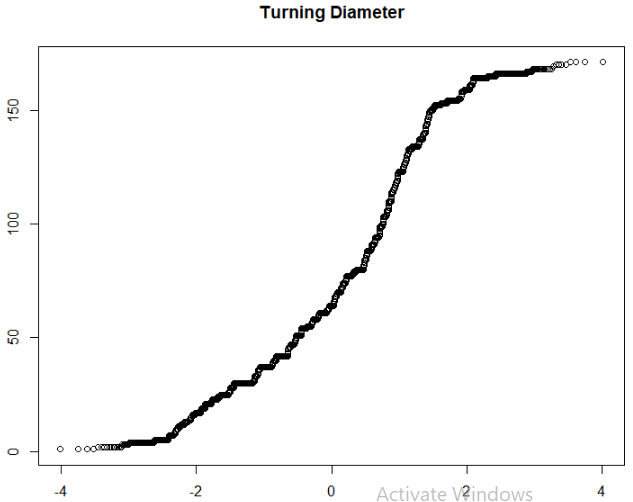
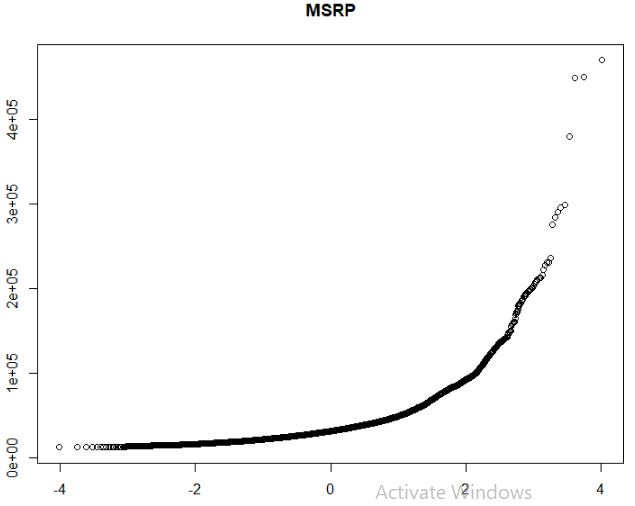
1. MSRP – Manufacturer suggested retail price
2. EPA Fuel Economy Est - City (MPG) – Estimated City mpg
3. Engine – Type of Engine
4. Drivetrain – Type of Drivetrain (FWD, RWD, AWD, 4WD)
5. Passenger Capacity – Number of seats in the car
6. Passenger Doors – Number of doors on the car
7. Base Curb Weight (lbs) – Weight of the car
8. Passenger Volume (ft³) – Volume of the car chassis
9. Wheelbase (in) – Distance between the front and rear wheels
10. Track Width, Front (in) – Distance between the front two wheels
11. Height, Overall (in) – Height of the Car
12. Fuel Tank Capacity, Approx (gal) – Size of the fuel tank
13. SAE Net Torque @ RPM – Optimum torque at average RPM
14. Fuel System – Type of fuel injection used
15. SAE Net Horsepower @ RPM – Optimum horsepower at average RPM
16. Displacement – Measure of fluid moved by fuel cylinders (Liters)
17. Trans Description Cont. – Type of transmission
18. Trans Type – number of gears in the transmission
19. Suspension Type – Front – front wheel suspension type
20. Suspension Type – Rear – rear wheel suspension type
21. Air Bag-Frontal-Driver – binary [0,1]
22. Air Bag-Frontal-Passenger – binary [0,1]
23. Air Bag-Passenger Switch (On/Off) – binary [0,1]
24. Air Bag-Side Body-Front – binary [0,1]
25. Air Bag-Side Body-Rear – binary [0,1]
26. Air Bag-Side Head-Front – binary [0,1]
27. Air Bag-Side Head-Rear – binary [0,1]
28. Brakes-ABS – has anti-lock braking system, binary [0,1]
29. Child Safety Rear Door Locks – binary [0,1]
30. Daytime Running Lights – binary [0,1]
31. Traction Control - binary [0,1]
32. Night Vision - binary [0,1]
33. Rollover Protection Bars - binary [0,1]
34. Fog Lamps - binary [0,1]
35. Parking Aid - binary [0,1]
36. Tire Pressure Monitor - binary [0,1]
37. Back-Up Camera - binary [0,1]
38. Stability Control – binary [0,1]
39. Basic Miles/km – Miles under warranty
40. Basic Years – Years under warranty
41. Corrosion Miles/km – miles under corrosion warranty
42. Corrosion Years – years under corrosion warranty
43. Drivetrain Miles/km – miles under drivetrain warranty
44. Drivetrain Years – years under drivetrain warranty
45. Turning Diameter - Curb to Curb (ft) – 2\*turning radius of car
46. Front Wheel Material – construction material of front wheel
47. Stabilizer Bar Diameter - Front (in)
48. Roadside Assistance Years – years included at purchase
49. Roadside Assistance Miles/km – miles included at purchase
50. Manufacturer – Manufacturer of the car
51. Model year – model year of the car
52. Category – One of SUV, Car, Pickup, Van
53. Front tire width
54. Front tire aspect ratio
55. Front tire speed ratings/cons.type
56. Front tire rim size

**Analysis**

*Data Preprocessing*

Despite the dataset being relatively complete, its raw state needed some cleaning. Many rows contained null values. Due to limits on computational capacity, the imputation methods necessary to fill in these null values (KNN imputation or random forest imputation) could not be performed. Therefore, I decided to exclude all rows which contained null values. I decided to exclude variables where the number of null values exceeded 10,000. The variables dropped were Base Curb Weight, Passenger Volume, Track Width, Height, Fuel Tank Capacity, Stabilizer Bar Diameter, and Roadside Assistance – Years. Prior to excluding these variables, the total number of data points with no null entries was 0. After excluding these variables, the total number of data points was 16,626.

Something that is not immediately obvious from viewing the raw data is that many of the numerical variables in this data set are ordinal. Take for example, City mpg. This data is numerical in nature, but it is not a continuous or discrete random variable. City mpg will be the same across cars with the same make and model, even if their features (like having side air bags) are different. For this reason, City mpg is *not* measuring an independent feature of a random car but can be viewed as an individual cars “score” on a scale of City mpg ranging from 9 to 60. After I interrogated each variable in the dataset, I found there are only 5 features which behave like discrete random variables: Wheelbase, SAE Horsepower, SAE Torque, Turning Diameter, and MSRP. Their respective distributions are presented in Figure 1.



A

B

C

D

E

Figure 1: Normal QQ-plot of discrete random variables. (A=MSRP, B=Turning Diameter, C=SAE Horsepower, D=SAE Torque, E=Wheelbase)

*Statement on Selection of Analytical Methods*

In the submitted plan of analysis, my partner and I outlined several analytical methods we would use including PCA and linear regression. After detecting ordinality in the data, I decided decision trees would be the best way to go for analysis. There are several reasons for this. First, decision trees are capable of handling both numerical and categorical types of data (depending on the algorithm/implementation, more on this later). This means I did not have to one-hot code all the categorical variables which would produce a sparse matrix and make for tedious analysis. Additionally, decision trees are non-parametric and thus only affected by absolute order of data. This means our results will not be sensitive to non-linearity of which (given the ordinal nature of much of our data) there is a lot.

*Selection of Analytical Methods*

Given the highly ordinal nature of the data, analyzing this dataset using decision trees will be the chosen method of analysis. All 49 remaining variables will be used in analysis. The analysis will take place in two stages.

1. *Decision Tree Model*
   1. Some implementations of the CART algorithm (ex: Scikit-learn) cannot consume both categorical and numerical data. Therefore, the decision tree algorithm chosen for this analysis is RPART, an implementation of the CART algorithm that can handle both types of data.
   2. Data will be split into training and test sets containing 70% and 30% of the data, respectively. Once we have trained the model on the training set, predictions made on the test set will be evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). These two metrics were chosen to boost interpretability of fit as RMSE is more sensitive to outliers than MAE, and MAE provides a measure of accuracy on non-outlier data. Both metrics produce an intuitive comprehension of prediction accuracy compared to Mean Squared Error (MSE).
   3. A binary-tree plot will be produced to see on what nodes our tree splits, and how those nodes affect predictions. We will also visualize the model residuals.
2. *Random Forest Model*
   1. To boost the accuracy of our decision tree model, we will use random forests. This model combines the learned rules of decision trees trained on subsets of our features to make predictions with lower variance (also known as bagging). Since this algorithm is computationally expensive, the model will only contain 350 trees.
   2. To analyze this model, I will use feature importance. Feature importance is a measure of how negatively predictions are affected if a given feature is replaced with random noise. In regression, feature importance is computed using MSE.
   3. As above, we will split the data into training and test sets. We will then evaluate model predictions using RMSE and MAE and compare performance and the overall structure of both models.

**Results**

*Decision Tree Model*

The obtained Decision Tree contained 9 splits based on the variables SAE Net Horsepower, Manufacturer, Engine, and Parking Assist (Figure 2). The interpretation of the model structure will be discussed in the **Interpretation** section. This model produces 10 predictions of car price based on the node criteria of each leaf. The model residuals are centered around zero, except for several outliers (Figure 3). The residuals are diagonally oriented due to the way in which our model makes predictions. Our model can only predict the MSRP of a given car to be 1 out of 10 prices. Therefore, many different data points in the test data will be predicted as having the same price. This is made clear in the Actual vs. Predicted plot (Figure 4). Here we can see that our test data is essentially ‘classified’ into 1 out of 10 groups as indicated by the 10 distinct columns of data. We can see that the model does relatively well at predicting the price of cars in a moderate price range between 0 and 80,000 dollars but is not so good at predicting the price of very expensive cars. As we can see in Figure 1, the last split before the highly expensive predictions of $143,000 and $247,000 is on the variable ‘manufacturer’. It is likely that the skew in our predictions at these high prices is due to a high contribution of car manufacturer on model price that is not well captured by the single node in our data. Still, the line of best fit demonstrates a tight relationship between model predictions and the real data with a slope of 1.013 and an r2 value of .8033. For such a simple model it is impressive that we can capture around 80% of the variance in the data!

The RMSE of the model is $9795.835, and the MAE is $6271.504 (Figure 5). Given that RMSE is more sensitive to outliers, and that the model performs worse at higher price points where there is less data, it is expected to be quite high. MAE is about 30% lower since it is more reflective of the average prediction error on non-outlier data which is where the model performs best. The decision tree model has the benefit of being computationally inexpensive as well as being easy to visualize and interpret. However, with a slightly more complex model in a random forest we will almost certainly be able to make better predictions and add some context to the granular ‘rules’ developed by the decision tree model.

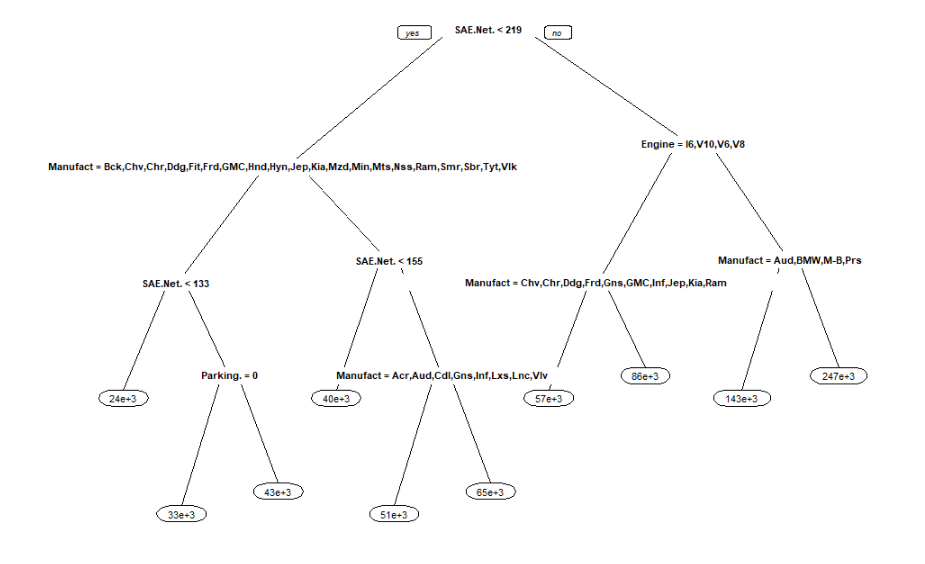
Binary Tree Plot

Figure 2: Plot of Decision Tree Model. Branches to the left indicate an observation meets the node criteria, branches to the right indicate an observation fails the node criteria. Leaves indicate the prediction (in dollars) made for the cost of an observation which meets the given node criteria. SAE.Net refers to the variable ‘SAE Net Horsepower’.

Residual Plot

Chart, scatter chart

Description automatically generated

Figure 3: Plot of residuals from Decision Tree predictions. Residuals were calculated using (actual – predicted).

Actual vs. Predicted Plot

Chart

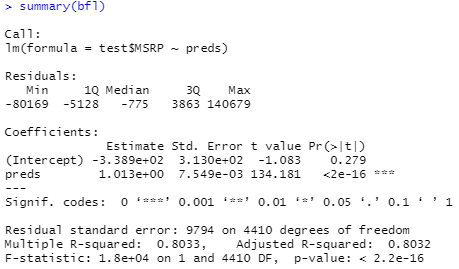
Description automatically generated

Figure 4: (Top) Plot of Actual vs. Predicted MSRP. Line of best fit shown in Red. (Bottom) Summary statistics of line of best fit.

RMSE and MAE



Figure 5: R output of RMSE and MAE calculations for decision tree model.

*Random Forest Model*

The random forest model contained 350 trees; the number of trees being limited by my desktop’s computational capacity. This model was trained using the exact same training and test data as the decision tree model, allowing for direct comparisons between the two. The random forest produced feature importance measures of all 49 variables, the top 15 of which are presented in Figure 6. Feature importance is determined by calculating the MSE of the random forest when the data in each feature is replaced with random noise. We can see the most important features by far are SAE Net Horsepower, Net Torque, Manufacturer, Displacement, and Engine. More will be made of comparisons between this model and our decision tree in the **Interpretation** section, but it is immediately apparent there is a large amount of overlap in terms of the most important features between models.

The residuals of the random forest are tightly centered at zero for cars with an MSRP of up to $100,000. Afterwards, we see the model struggles again with the prediction of MSRP for highly expensive cars (Figure 7). It is important to note that far fewer residuals fall above $50,000 compared to our decision tree model. The Actual vs. Predicted plot (Figure 8) reveals exactly how tightly our random forest fits our data. The line of best fit has a slope of 1.031 and an r2 value of .9743, meaning that our model captures a huge 97.43% of the variance in our data!

The RMSE of our model is $3,609.384, and the MAE is $1,703.58. On these two metrics our random forest fits the data over twice as well as the decision tree model. These two metrics indicate that the random forest exhibits superior performance on not only the average prediction, but on handling outliers as well.

Feature Importance

Chart, scatter chart

Description automatically generated

Figure 6: Node Purity (Feature Importance) of 15 most important variables in the Random Forest Model. Node purity is measured by how much MSE of the model increases if each feature is replaced with random noise.

Residual Plot

Chart, scatter chart

Description automatically generated

Figure 7: Residual Plot of Random Forest Model. Residuals calculated using (actual – predicted)

Actual vs. Predicted Plot

Chart, scatter chart

Description automatically generated

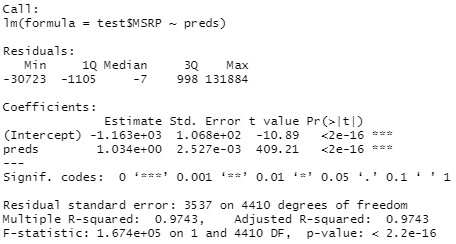


Figure 8: (Top) Actual vs. Predicted Plot of Random Forest Model. Line of best fit shown in red. (Bottom) Summary statistics of line of best fit.

RMSE and MAE



Figure 9: R output of RMSE and MAE of Random Forest Model.

*Interpretation*

Although our decision tree and random forest had differences in their predictive capabilities, the story they tell is largely the same. Our decision tree determined that the most important feature was horsepower. The greater the horsepower, the more expensive the car. The next important feature to MSRP was the engine in the car, followed by the manufacturer. If the car had low horsepower, then the cost was significantly lower, unless it had an additional amenity in parking assistance, or came from a manufacturer with higher status such as Audi, Lexus, Cadillac, etc. (Figure 2). Our random forest determined the most important feature for predicting MSRP was also horsepower, followed by torque, manufacturer, displacement, and engine type (Figure 6). Recall from the data dictionary that displacement refers to the amount of fuel moved by each piston in the engine, meaning that like horsepower and torque it is a measure of engine performance. Thus, across both models the greatest predictors of MSRP are measures of engine performance and manufacturer.

What about the other 44 variables we have yet to discuss? Our results would suggest that although there is sure to be a non-zero contribution from each, we can capture most of the variance using manufacturer and measures of engine performance. To explicitly test this hypothesis, we can train a decision tree on only this subset of features. Indeed, what you will find is that although you get fewer predictions in price (9 instead of 10) you capture approximately the same amount of the variance in the data compared to the full model (r2=.79, Figure 11). The splits obtained in the binary tree-plot are similar to the decision tree trained on the complete data (Figure 10).

So, what determines the price of a car? In the introduction we hypothesized that it could be the social status that the car conveys, or the quality of the car itself. In light the of the current analysis, the answer is emphatically…both. However, this is not some cop out “everything is shades of grey” answer. The number one predictor of MSRP is horsepower. Higher horsepower means a nicer engine, and the higher quality the engine the higher the price and the chance that it comes from a car manufacturer with greater social status. Notice in Figure 2 that the manufacturers in nodes on the far-right side of the tree like Audi and BMW are more socially elevated brands compared to those on the left like Ford, Dodge, and Kia. What do you think about when you imagine an expensive car? Usually a fast one, right? However, regardless of engine performance it is almost always the case that cars from socially elevated brands like BMW, Audi, and Lexus are more expensive (according the reduced model by about $12,000). So, the question is: all things being equal, is the social status of a luxury car brand worth the $12,000 extra you will pay for it? My wallet wouldn’t like the answer.

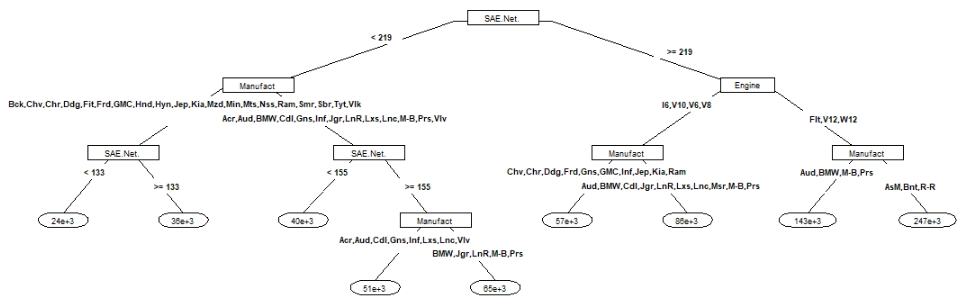


Figure 10: Plot of Reduced Decision Tree Model. Branches to the left indicate an observation meets the node criteria, branches to the right indicate an observation fails the node criteria. Leaves indicate the prediction (in dollars) made for the cost of an observation which meets the given node criteria. SAE.Net refers to the variable ‘SAE Net Horsepower’.

Chart

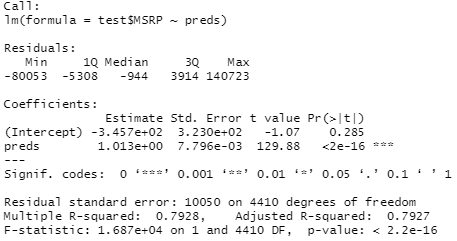
Description automatically generated

Figure 11: (Top) Actual vs. Predicted

MSRP of reduced model. Line of best

Fit shown in red. (Bottom) Summary

Statistics of line of best fit.



**Data Source**

1. https://www.kaggle.com/prassanth/new-cars-price-2019

**Sources**

1. <https://www.statista.com/statistics/199974/us-car-sales-since-1951/>
2. https://www.apta.com/news-publications/public-transportation-facts/