# Introduction

What advantage would a business have if they were able to accurately quantify the different factors that give something value? How could their business strategy be improved if they could accurately predict an items value? In this paper, the data analysis and statistical science behind answering these types of questions is explored through the lens of something most of us use every day, our cars. When a vehicle is produced the auto maker specifies the manufacturers suggested retail price (MSRP) which is intended to be measure of that vehicle’s worth. In part one of this analysis, we tackle the problem of building a model that can quantify by how much we estimate a vehicles worth would change when one of its attributes is changed. Then in part two we trade some of this interpretability for increased predictive performance, and seek to build models that can specify a particular cars value as accurately as possible.

# Data Description

The original dataset contains 11914 observations, each one representing a particular type of vehicle that was manufactured between 1990 and 2017. Each observation consists of 15 attributes that describe various vehicle characteristics, as well as a 16th column containing our target, the MSRP. The descriptive attributes range from common information such as the vehicles Make and Model, to somewhat esoteric descriptors such as a social media derived vehicle popularity score. A data dictionary with descriptions of each attribute is provided in **Table 1**.

Initial analysis found missing information in three of the vehicle attribute columns. Specifically, the engine fuel type, engine cylinders and engine horsepower attributes contain 3, 30 and 69 missing observations respectively. The correct values for these missing records were researched and recorded in a csv file [2], which was then referenced in the data cleaning script to impute the missing observations. The dataset also contains a considerable amount of duplicate information, with 715 observations that are full duplicates across all 16 columns, and another 4563 observations that are duplicated across all 15 vehicle attribute columns, and only differ on the associated MSRP. Since one of the objectives in this analysis is to utilize linear regression to provide inference and prediction on vehicle MSRP’s, the duplicate information was not acceptable as it would likely lead to a violation of the model’s assumption regarding independence of errors. This was remedied by averaging the MSRP values for duplicate vehicles, and only retaining one observation for each unique vehicle in the dataset.

The dataset also contains some clerical errors, such as a 2017 Audi A6 being documented as having 354 highway miles per gallon (MPG), when the true highway MPG for that vehicle is listed as 35. All data cleaning related activities, including those that are described in detail later on, were performed in a single, repeatable, data cleaning script [3]. This script generates a data cleaning report [4], which provides a more detailed overview of the preprocessing steps. A set of summary statistics for both the original version of the dataset and the version after all preprocessing steps were applied have been provided in **Table 2** and **Table 3**.

# Exploratory Data Analysis

Exploratory analysis involved investigating the univariate distributions of all 16 columns in the dataset, as well as the joint distributions of each predictor with the response. For continuous variables, this was implemented using histograms and scatter plots. For the categorical variables, this was done with count plots, and means plots that show the average and variability in responses for each level of the factor. Through this process a number of insights were gained, the most important of which will be summarized next.

The histogram of the response in **Figure 1** shows a significant right skew, which is typical of sales related data that cannot take on negative values. We also see that the log transformation does a decent job at reigning in the large MSRP values in the right tail of the distribution. This influenced the decision to predict log MSRP rather than MSRP directly in some of the later modeling activities.

The count plots for the Make, Engine Fuel Type, and Market Category features, shown in **Figure 2 – Figure 4**, reveal that each of these factors has at least one level that contains only a very small number of observations. This can lead to a number of issues when splitting the data into training, validation and test sets and evaluating models. For example, if a factor level is absent from the training set, a regression coefficient will not be fit for that level, and subsequent predictions involving that factor level will result in errors. A more subtle issue is that having too few observations in a factor level will result in a poor estimate of the associated regression coefficient, which can lead to poor model performance. For each of these features, this common problem was addressed in a slightly different way.

For the Engine Fuel Type feature, the two “natural gas” vehicles were removed from the dataset, and the nine remaining factor levels were binned to create three more general categories, as shown in **Figure 5**.

For the Market Category feature, detailed inspection showed that the low factor level counts were the result of storing multiple distinct pieces of information as a comma separated list within each cell (e.g., Luxury, Exotic, Performance). This was reformatted into nine separate binary columns, one for each unique value in the original feature, aside from a few consolidations (e.g., high performance and performance combined).

The Make feature contains 48 different levels, with no clear way of merging any of them together. The Model feature falls into this same category, and is even more extreme with 915 levels. With a dataset of this size, it would be unwise to use these features because there are likely not enough observations to properly estimate the number of regression coefficients that they would require. As a result, these features were deemed unusable for the modeling stages of the analysis.

## Defining the Population

This dataset represents a very broad range of vehicle types; however, it is imbalanced in the sense that for many of these vehicle types only a small number of observations are provided. This situation makes building an inferential model that is sufficiently broad, yet still reliable in terms of adequately meeting the models underlying assumptions a very difficult task. In the interest of providing a more robust set of interpretations, an iterative set of decisions were made to further refine the vehicle population under study.

The 11 vehicles with MSRP values above 500,000 were removed from the analysis, as well as the 69 vehicles with electric motors. The 20 rotary engine vehicles where the Engine Cylinders attribute was listed as “NA” were also removed.

The removal of electric vehicles resulted in all but two of the “DIRECT\_DRIVE” transmission type vehicles being removed, creating another instance of the sparse factor level issue described above. These two vehicles were Chevy Malibu’s, and were reassigned to the “Automatic” Transmission Type category. Additionally, the 12 vehicles listed with a Transmission Type of UNKNOWN were treated as incomplete observations and were removed from the dataset.

After examining the errors made a preliminary model, shown in **Figure 6**, it was observed that our set of vehicles still appeared to be members of two distinct populations, specifically those with MSRP values below ~$3,000 and those with MSRP values greater than about $10,000. A clear linear pattern was also observed in the lower price cluster, which is highlighted by the yellow markers in **Figure 6**. These yellow markers all represent vehicles with a true MSRP of exactly $2,000. This led to the final dataset modification of removing all vehicles with an MSRP below $10,000. Narrowing the range of vehicle prices by removing those with the most extreme values also helped reduce the skew observed in the original target distribution, which is evident by comparing **Figure 1** to **Figure 7**.

Accounting for all the decisions described above, the population of vehicles examined during the remainder of this analysis can be described as those with gasoline or hybrid gasoline-electric fuel types, standard engine architectures, and a manufacturer suggested retail price between $10,000 and $500,000. This population is considerably less diverse than the one described by the original dataset; however, it remains broad enough to form the basis of a useful inferential model that is applicable to the majority of vehicles on the road today.

Before proceeding to the modeling objectives, the final dataset with 5122 observations was split into three separate datasets. Specifically, a training set [5], a hold-out cross validation set [6] used for model selection and tuning, and test set [7] used to obtain a final unbiased estimated of the selected models generalization error, were created with 80%, 10% and 10% of the observations respectively.

# Objective 1

### Problem Statement

Build a regression model that captures the key relationships between the various vehicle attributes and its associated MSRP, with a specific emphasis on understanding the importance of the social media derived popularity attribute. Interpret the relationships described by the model in terms of their practical and statistical significance, and analyze any influential points or sources of error that could invalidate these interpretations.

### Model Selection

Based on the relationships observed during the EDA, the model building process began with a set of 20 candidate predictors [8]. The best subsets selection method was used to fit all possible models, and evaluate each one on a set of 15 metrics that measure model performance on both the training and validation sets. The predictors used to generate each model and the associated evaluation metrics were recorded in a file [9] for further analysis.

Following a variation of the method outlined in [10], the set of models containing exactly predictors for that received the highest score on at least one of: *Adjusted R-squared*, *Validation set R-squared*, or *Prediction R-squared*, were then set aside for further evaluation. With 20 model sizes and 3 metrics, this final set could have contained up to 60 candidate models. However, after accounting for models that performed better than all others of the same size on more than one of these metrics, only 30 unique candidate models remained.

To select among the final 30 candidates, each model was evaluated again using three distinct categories of metrics; 1. Repeated K-Fold cross-validation, 2. Leave-one-out, 3. Probabilistic selection statistics. Note that none of these approaches to model evaluation benefit from a dedicated hold-out cross validation set. For this reason, the training and validation sets were combined prior to evaluating the last 30 models. A dedicated test set remained set aside to obtain a final estimate of the generalization error for the selected model.

More concretely, the caret package was used to perform repeated 10-fold cross validation with 20 repeats and calculate the cross validation RMSE, R-squared and MAE metrics. The olsrr package was used to calculate the leave-one-out style PRESS and Prediction R-squared metrics, and the stats package was used to calculate the Akaike’s Information Criterion and Bayesian Information Criterion selection statistics. The full set of metrics for each of the 30 final candidate models can be found in [12].

The final evaluation round once again resulted in the largest model (18 predictors) performing the best across all metrics being considered. However, similar to what was observed in the initial best subsets selection search, the amount by which the most complex model out performs its simpler counter parts was minimal. **Table 5** compares the top performing model to a simpler model with 8 less predictors (10 total), and shows that the complex model out performs its simpler counterpart by relatively small percentages across all metrics.

Since the complex model contains the Popularity predictor and the simpler version does not, it is convenient to proceed by analyzing both of these models to further explore the question of interest regarding the importance of the vehicle popularity attribute. Both models were evaluated on the final test set, and as shown in **Table 5**, the complex model outperformed the simpler version in terms of Test RMSE, Test R-Squared, and Test MAE by 4.99%, 0.796%, and 3.68% respectively.

Since the simple and complex models shown in **Table 5** differ across 8 different predictors, a natural question arises regarding how much of the complex models increased performance should be attributed to the inclusion of the popularity variable. We begin to explore this question in **Table 6**, where the test set performance of the complex model and another version of itself where popularity had been removed are compared. **Table 6** shows that the model which includes the popularity predictor out performs the one that does not by 0.413322%, 0.067236%, 0.417942% on the test RMSE, test R-Squared and test MAE metrics respectively. This shows that when the highest performing model no longer accounts for popularity when predicting on previously unseen vehicles, the impact to the model’s ability to provide correct predictions is very minimal.

We can visually inspect the association that exists between vehicle Popularity and the response, after the effects of all other predictors have accounted for, through a plot of the model’s partial residuals, shown in **Figure 8**. The small slope of the least-squares line fit to the partial residuals indicates that after accounting for the other predictors, there is only a slight association between vehicle popularity and the response. Further, the high variability about this line provides visual evidence that with the other 17 predictors in the model, the inclusion of popularity is not very useful at describing the variability left in the response.

This idea is reinforced by inspecting the estimates of the model’s parameters, shown in **Table 7**, where we see that the estimate associated with the Popularity variable is . In the context of our model that utilizes a log-transformed response variable, we would say that a 1 unit increase in vehicle popularity is associated with an multiplicative change in the median vehicle MSRP with all else held constant. The conclusion here is that although the model found Popularity to be a statistically significant predictor, the actual improvement that comes with accounting for Popularity when predicting vehicle MSRP is almost certainly not of practical significance.

In light of the above findings, and due to a preference for a simpler model whenever possible, the following model assumption checks and parameter interpretations will be performed using the relatively simple 10 predictor model discussed above.

### Checking Model Assumptions

**Influential Observations:** The plot of externally studentized residuals versus leverage in **Figure 12** shows four highly influential points in the bottom right corner of the plot. These observations each have an externally studentized residual greater than 4.5, and a leverage statistic that exceeds 2500 times the average (where average leverage is calculated as the number of regression coefficients divided by the number of observations). By comparing to **Figure 13**, note that **Figure 12** has been zoomed out to show these extreme values, because with the standard plot settings used by **Figure 13** these values are off the page. These four points correspond to observation numbers 4320, 4842, 4843 and 4844, and we note that they have also been flagged as influential in the upper right corner of the Cooks D vs Observation Number plot shown in **Figure 14**, as well as the lower right corner of the DFFITS vs observation number plot in **Figure 15**. In **Table 8** we show the entire set of predictor and response values associated with these four observations. By examining the predictor values, we see that all four cars are classified as Exotic, however their MSRP values of around $100,000 are well below the $225,698 average of all vehicles in the exotic category, which may partially explain why the model struggles with these particular points.

The presence of these influential observations means that before attempting to use the results of this analysis to answer any questions of interest, we must perform a second analysis without these observations to determine if their removal alters the answer to the question we are researching. It would be unwise to state statistical conclusions that hinge on four observations.

**Linearity:** From the plot of the externally studentized residuals in **Figure 9** we a reasonable depiction of the ideal random cloud, with no clear discernable pattern, which indicates that the linearity assumption is reasonable.

**Independence:** Also, from the residual plot in **Figure 9**, we do not observe any “tracking” in the residuals. A detailed understanding of how the data was collected is usually required to fully investigate this assumption, however we will proceed with caution as if independence is met.

**Normality:** From **Figure 9** we see that the residuals appear to be evenly distributed around zero, and are not systematically above or below the horizontal red line plotted at zero. The histogram of the residuals shown in **Figure 10** has a clear bell-shaped curve associated with it, however we do note that the distribution has longer tails than would be expected from a sample of this size. The QQ-plot of residuals shown in **Figure 11** is mostly linear which agrees with normality; however, we do note the characteristic S-shaped curve at the edges which is another indicator of the long tails noted from **Figure 10.** Lastly, we note that the points marked in red on **Figure 11** which are contributing to the long tails are influential observations 4320, 4842, 4843 and 4844 that were previously noted above. This is a reasonably large dataset and due to the central limit theorem normality is not a major concern, therefore we will proceed with this assumption being met.

**Equal Variance:** From **Figure 9** we do not see any sharp “funnel shape” or “fanning” that is characteristic of heteroskedasticity. That said, we do note that nearly all of the most extreme values occur on the right-hand side of the plot, which is also much more sparsely populated compared to the left. With fewer samples of expensive vehicles, it seems unlikely that we would observe so many more expensive vehicles with large residuals if the variance of the errors at every set of predictor values were truly equal. With these concerns noted, we will proceed with the analysis cautiously as if this assumption is met.

**No Multicollinearity:** The variance inflation factors for this model are shown in **Figure 16**, and at a first glance there seems to be a clear issue with the Engine Cylinders predictor, which has VIFs for some categories as high as 142.78. Fortunately, however, the situation is not as bad as it seems. As described in [13], this is a situation where the high VIFs are the result of there being a small number of observations in the reference category used by the model. Additional evidence supporting this explanation is provided in **Figure 17**, where for the exact same model substantially lower VIFs were calculated after changing the Engine Cylinder factors reference level. This situation means we should expect higher than normal p-values on one or more of the Engine Cylinder dummy variable coefficients, however it does not invalidate the rest of the model.

### Parameter Interpretation

For brevity, an interpretation of a single regression coefficient is provided here as an example. For a complete set of interpretations of the model’s coefficients, please refer to **Table 10**. *Age* has a coefficient estimate of . Since this is a log-linear model, we interpret this coefficient by saying we would expect a one unit increase in the cars age to be associated with a multiplicative change in the median vehicle MSRP, with all else held constant.

# Objective 2

## Problem Statement

The goal of the second portion of this analysis is to build more complex models that emphasize accurate predictions over inferential ability. The first model will build upon the linear regression model investigated in part one of the analysis, and the second will be a non-parametric random forest model. In the conclusion of this objective, the full set of models from part 1 and part 2 will be compared in terms of their cross validation and test set performance.

## Approach to building complex regression models

The approach to building the complex regression models was to review the exploratory analysis and the diagnostics from previously created models to see if the data suggested that the model would benefit from the inclusion of polynomial or interaction terms. Since partial residual plots provide a way to visualize the relationship between a predictor and the response after accounting for the other predictors in the model, this was the tool chosen to research whether or not a squared term on a particular predictor may be a good thing to try. Using the 10-predictor model from part 1 of the analysis, the partial residual plots with respect to Engine HP and Age are shown in **Figure 18** and **Figure 19**. By adding City MPG to this same model from part 1, the partial residual plot with respect to City MPG was generated, and is shown in **Figure 20**. Out of these three plots, the one with respect to City MPG seems the most promising due to the clear bend that begins at around an MPG of 32. Although it is much more subtle, there could also be an argument made for some light curvature on the right side of **Figure 18** and **Figure 19** as well.

Categorical means plots were used to investigate potential interaction terms. In **Figure 21**, **Figure 22** and **Figure 23**, we show the categorical means plots for the Engine Cylinders factor with the three binary factors of Luxury, Diesel and Exotic respectively. Based on these plots it seems plausible that the change in response associated with a change in Engine Cylinder levels could be different based on the level of the three binarized Market Category features, particularly in regions where Engine Cylinders transitions from 8 to 10.

Based on the advice in [14] that cubic and higher-order polynomials are rarely useful, it was decided not explore any such terms in this stage of the analysis. Therefore, the following six terms were the candidates considered for adding additional complexity to the model in part 1 of the analysis:

1. Interaction of Engine Cylinders and Diesel
2. **Interaction of Engine Cylinders and Exotic**
3. **Interaction of Engine Cylinders and Luxury**

With the 10-predictor model from part one as the baseline, the addition of all combinations of the terms outlined above became the hypothesis space in which we searched for a high performing model with added complexity. Once again, the method of comparing these models was with repeated 10-fold cross-validation with 20 repeats via the caret package. For each model the cross validation RMSE, R-Squared, and MAE statistics were recorded, and saved for further analysis [15].

Using cross validation RMSE as the ranking metric, the highest performing model was the one that added to the original model terms 1, 2, 3, 5 and 6 from the list above, shown in bold. This model had a cross validation RMSE of 0.1777650, however when evaluated to the final test set only received a RMSE of 0.1835966. This reduced performance on the previously unseen data indicates that despite the use of cross validation during model selection, some amount of overfitting still occurred.

### Description of the non-parametric approach

Since a random forest is an ensemble of decision trees where each tree is trained on a different bootstrapped version of the dataset, with the addition of an element of randomness added to the split predictor selection, a full description naturally begins with a discussion on decision trees. A decision tree is an intuitive model that simply looks for the predictor which, when split at some location, does the best job at dividing the observations into buckets with similar values of the response. This process of selecting a predictor and split location is repeated greedily until a stopping condition is met, such as each leaf node (bucket) containing fewer than a certain number of observations. The main issue with decision trees is that they become overly complex, and as a result they don’t generalize well to new data. In other words, decision trees have high variance, meaning that they fit the training data so closely

that relatively small changes in the data can significantly change the resulting tree. A method of reducing the variance of decision trees is by bagging, where multiple decision trees are trained on bootstrapped (randomly sampled with replacement) versions of the datasets, and the predictions from each tree are averaged to generate the model’s final prediction. This helps reduce variance because in general, the average prediction of many trees will be less variable than the prediction of a single tree. The final issue, that brings us to the formulation of the random forest is shown below:

The above equation shows that if and are highly correlated, that the variability of their average will be greater than if they are not correlated. If each tree is allowed to choose from the entire set of predictors when creating their splits, then despite the use of bootstrapped datasets we find that in many cases a large number of trees still choose the same, strongest, predictor for the trees initial split. The result of this is high correlation amongst the trees. The *random* in random forest comes from the fact that we help break this correlation by only allowing each decision tree to select from a *random sample* of the predictors when creating each split.

## Comparison of results

LightGBM [16] was originally developed by Microsoft as a powerful and efficient machine learning library for tree-based models. It has implementations for random forest, as well as other popular tree-based models such as gradient boosted decision trees [17] and DART [18]. All three of these models were experimented with during this final stage of the analysis, and the interested reader can view the models feature importance scores in **Figure 23**, **Figure 24**, and **Figure 25**.

The cross validation and final test set RMSE values for all three models are compared to the two regression models from earlier in the analysis in **Table 11**. Note that the metrics for the linear regression models are on a log scale, where as the metrics from the tree-based models are in the original target units (dollars). The difference in scales makes a direct comparison between the parametric and non-parametric regression models difficult. An interesting observation is that most models exhibited some level of overfitting, evidenced by the larger error on the final test set as compared to the cross-validation data. The one exception to this was the DART model, which actually performed better on the final test set than it did during cross validation.

### Final Summary

Throughout this analysis we explored the different factors that help us quantify a vehicles worth. In part one this was accomplished using highly interpretable linear regression models. Through this process we saw that great care must be taken to adequately define the population under study so that when inferences are made it is well understood in what areas they are applicable, and in what areas they are not. We also saw that the price paid for these powerful interpretations is satisfying a fairly rigorous set of model assumptions, therefore a harmony must be reached between what the researcher wants to know, what the dataset can provide, and what the models assumptions will permit. In part two of this analysis, we removed the burden of interpretability and focused purely on the accuracy of our predictions. We saw that by adding complexity, linear regression is again a suitable tool for the predictive goal. Lastly, we showed that non-parametric models such as random forest and its variants can provide very accurate predictions as well as insight through the examination of the models feature importance scores.

# Appendix

Throughout the document **bold face** **text** is used to reference **Tables** and **Figures** stored in the external appendix [1]. The use of an external appendix is intended to make the paper more pleasant to read by avoiding the need to repeatedly scroll to the end of the document to see a referenced item. This paper is most easily read with the text and external appendix open side by side at the same time. If your copy of the appendix has been lost or stolen, don’t worry, a fresh copy is freely available on the [Project GitHub](https://github.com/BradenAnderson/Applied_Statistics_MSRP_Prediction)

# References

\* Indicates that the referenced item is available on the [Project GitHub](https://github.com/BradenAnderson/Applied_Statistics_MSRP_Prediction)

\*\* Indicates file exceeds GitHub maximum file size, and is instead hosted on Google Drive at the following link:

<https://drive.google.com/drive/folders/1ru38fSq4oh4zmnhjPc3s61CT1gXT4t1l?usp=sharing>

\*[1] APPENDIX\_Project1.pptx

\*[2] car\_dataset\_missing\_fillins.csv

\*[3] Applied\_Stats\_Project\_Functions.R

\*[4] Data\_Cleaning\_Report.txt

\*[5] train\_dataset\_0208.csv

\*[6] validation\_dataset\_0208.csv

\*[7] test\_dataset\_0208.csv

\*[8] 02\_Best\_Subset\_Selection\_Search.Rmd

\*\*[9] All\_Subsets\_Selection\_0210.7z

[10] Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. An Introduction to Statistical Learning: with Applications in R. New York: Springer, 2013. **Page 205**

\*[11] Add notebook name for final model evaluations

\*[12] FINAL\_MODEL\_COMPARISONS.csv

[13] <https://statisticalhorizons.com/multicollinearity/>

[14] Ramsey, Fred L, and Daniel W. Schafer. *The Statistical Sleuth: A Course in Methods of Data Analysis*. Boston: Brooks/Cole, Cengage Learning, 2013. Print. **Page 295**

\*[15] added\_complexity\_analysis.csv

[16] <https://lightgbm.readthedocs.io/en/latest/>

[17] <https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf>

[18] <https://arxiv.org/abs/1505.01866>