James Griffin

Arthur Romero

Stat 536

Kelly Blue Book Report

**Introduction**

The used car market generates roughly $40 million dollars a year in the United States[[1]](#footnote-1). These preowned vehicles are bought and sold by dealerships, individuals and corporations. Since the Model T hit the streets in the early 1900’s. Kelly Blue Book (KBB) has been a trusted resource for the values of used cars. The goal of this analysis is to analyze a subset of the KBB data to better understand and model the factors that go into the KBB value of used cars. Secondarily, we want to give a price range so that consumers can have a good idea of what to sell their car for, at which prices to buy, and have a tool that will help the customer know if they are getting a good deal.

**The Data**

Each car in the data set came from a General Motors manufacturer. This had both advantages and disadvantages. Our data will have less noise seeing as we do not have any German or Japanese cars in the data set, which may have depreciated at different rates, leading to different resale values. Thus, our predictive model will have trouble generalizing to the greater car population and should be limited to General Motors makes. Another potentially limiting factor to the generalizability of this model is that each car in the data set was also relatively new when the data was collected. Each car was less than a year old and was considered to be in excellent condition.

Each car came with details about their make, mileage, model and trim, as well as their body type, number of cylinders and leaders in the engine, number of doors, weather it had cruise control, upgraded sound system and leather seats. Lastly, each car came with a resale value, and this was our target variable for the analysis, while the other variables above were our independent features.

1. **What factors lead to higher or lower resale value?**

In trying to determine which factors lead to higher and lower resale values, we knew we needed to examine some of our data columns that were going to be very colinear. For example, the make of a car is very related to the model, and the models are very related to the trims, each being a subset of the other, thus lots of the information inherent in the trim is also inherent in the make and model of a car.

We decided to create interaction terms with makes and trims, because trims were a subset of models, and we wanted to limit our feature set to avoid over fitting, we thought that this would be a powerful way to capture not only what car did a person but, but what level of trim did that person buy.

Our milage variable was fairly normally distributed with a mean of 19,831 and a standard deviation of roughly 8000. This was encouraging for the assumptions of linearity that we wanted to make about the data, which generally held[[2]](#footnote-2). The one variable that showed the most collinearity were cylinders and liters, which again makes sense as liters are a function of cylinders in a car. We decided to keep liters in our analysis and drop liters, thinking that fewer possible values for cylinders would help limit noise, and be more beneficial in interpreting the data. While one cannot marginally add liters to their engine to the tenth of a liter, like the data we had, but they can buy a car with another cylinder.

We also noticed that the outcome variable for price was very skewed, so we transformed it, by taking the log to get more of a normal distribution to hold with our assumptions. This would provide a more normal variable to help reduce the errors.

Having satisfied the assumptions we needed for linear techniques, as the trained economists that we are, we felt comfortable using Ordinarily Least Squares from the statsmodels.api package in python. We left our one of each categorical variable that we had dummied, Make, and Type, in particular, we were able to run the regression.

The results are shown in Figure 2 of the appendix.

Each of the variables selected was statistically significant. This was achieved in a not so scientific manner, where we ran first a regression with all of the variables included except for the ones we mentioned above which we were excluding. We then trimmed away those who were not statistically significant and did this iteratively until we were only left with variables which were all statistically significant.

The factors for which influence the price of the car are the inputs listed on that table, and those which negatively impact are negative, and those which positively impact the resale value, are positive.

1. **What other factors not included in the dataset could help explain the value of a car?**

We believe that the features of this data set are adequate in understanding the resale value of GMC cars that are less than a year old, however having data on any possible rebuilt titles, accident history, if the primary driver was male or female, to see if there is a possible difference in aggressiveness and therefore a difference in depreciation would be interesting. Also, if we could have geographic characteristics of where the car was primarily located, if it was humid, or mountainous, mile per gallon, and age of the primary driver. All of these would add details to the data which are not captured by the data that we have and could potentially be useful in creating a more accurate understanding of the resale market.

1. **Understanding the impact of mileage on resale value**

The interpretation each of the coefficients is however not so straight forward as to simply read off and say that a 1 unit increase in would have a coefficient effect on the price, because we took the log of our price to help control for heteroskedastic errors. So, any coefficient from the table, must be untransformed before it can be interpreted. Thus, our true effects of any coefficient, . For example, the coefficient on mileage, is:

Then undoing the transformation:

This is the true value of our mileage coefficient, and we can now say that for every mile a person drives, the value of their car decreases by nearly $1.00. This, much like the model in general, is quite specific to the data and would not generalize to all cars on the road. We know this because if we had a car that had driven 100,000 miles, then according to this, the value of that car would have decreased by $98,000. Thus, the marginal effect of mileage on output price is fairly linear in this analysis, but we can assume that there is a decreasing marginal effect for miles at some point beyond the first year of driving.

We also wanted to understand if the devaluation of mileage was uniform across all makes. In other words, do some GMC manufacturers depreciate faster than others as they are driven more. We were able to test this by making interaction terms of make and mileage. Because we are working with a categorical variable in Make and a continuous in Mileage, we defined out interaction as follows:

So the regression equation was as follows:

And we compared this with only being a function of mileage by using an ANOVA test and comparing F statistics and probabilities. In ANOVA testing we started with a null hypothesis that all of our Mileage\*Make interaction terms would be 0, in a sense that there is no difference between the depreciation rates of milage between makes. Our F probability was 0.42, so we failed to reject the null hypothesis and cannot say if any one make depreciates at a greater rate than the others.

1. **Which car has the highest resale value with a minimum of 15,000 Miles?**

To answer this question, we grouped all of the cars in the data set by make, Trim, Model, Type, Cylinder, Liter, and Doors and then looked at which had the highest mean resale value. We found that the Cadillac XLR-V8, hard top, two door convertible with 8 cylinders and a 4.6 liter engine had the highest mean resale value at $59,09.95. Interestingly enough, top three cars in terms of resale value were all Cadillacs, and six of the top 7 were Cadillacs. We found that in our prediction, that the third most important feature in our Random Forest regressor was in fact Cadillac[[3]](#footnote-3).

1. **What is a reasonable resale value for a Cadillac CTS 4D Sedan with 17,000 miles, 6-cylinder, 2.8-liter engine, cruise control, upgraded speakers and leather seats?**

As we approached this question, we wanted to be able to learn from the data and train a predictive model. We took the columns that we found to be significant from the OLS regressions, and those were our features for training. We split our data set in training, validation and test sets, and used a dictionary of hyper parameters with different minimum splits, maximum depths, number of trees, and the minimum samples per leaf, and did a randomized search across the parameters with a 5 fold cross validation.

We then tested our different validation sets by plotting our validation predictions against our true validation resale values, with a best fit line and the . Our = 0.98, which means that we capture 98% of the variance in the data with the columns and model that we used. We also have values for our mean absolute error, and our root mean squared error, which are:

The reason we included both of them, is that root mean squared error and mean absolute error measure slightly different things. Because the errors are squared before they are square rooted, the *rmse* guards against missing big, so if the square root of that was very different than the *mae*, then that would be cause to go back and retool the features or hyper parameters, but because they are very comparable we felt good giving a prediction to the example posed in the lab specs.

Because we had 35 features, we needed to create a vector that had 35 entries, and then reshape that in such a way that the model could process those data points. Having done that, we could use the Random Forest infrastructure to feed it our numpy array and it gave us a prediction of $31,560.39. If we subtract and add the *rmse* from that number, we get the following range:. We believe that this is a reasonable range for the resale value of the car in question. We are confident because our training has been validated and we can see from the histogram in Figure 5 of the Appendix, that our predictions are very close to the actual resale values, and that as the resale value of the cars increases, our errors are still very homogenous. Thus, we can be confident that we did not over fit and the car we are predicting on fits the data that we trained on. If we were to say if we are under predicting or over predicting on this resale value, it is possible, that we are slightly under the true value of the car. Looking at Figure 5, we can see that most of the values are below the best fit line, meaning that we slightly under predicted the true resale value. Therefore, we are more likely to slightly under predict this value seeing as the predicted value is above $30,000, which is where this trend begins to appear.

Appendix

**Figure 1**

A picture containing shoji, building

Description automatically generated

**Figure 2**

Shape

Description automatically generated with medium confidence

**Figure 3**Table

Description automatically generated

**Figure 4**

**Chart

Description automatically generated**

**Figure 5**

**Chart, scatter chart

Description automatically generated**

1. Wagner, I. (2020, September 30). U.S. new and used car sales 2019. Retrieved January 28, 2021, from https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/ [↑](#footnote-ref-1)
2. Please See Figure 1 in the Appendix for graphs which gave us this confidence [↑](#footnote-ref-2)
3. For more information please see Appendix Figure 3 and Figure 4 [↑](#footnote-ref-3)