**Used Car Listings Analysis**



**Group 22: William Leonard, Susan Mikhail, Matthias Ronnau, Lokesh Nandni Sood, Parsa Tahmasebi**

**Data and Programming for Analytics-212**

**December 8, 2020**

Table of Contents

[1. Introduction 2](#_Toc58875449)

[1.1 Project Overview 2](#_Toc58875450)

[1.2 Background/Business Question 3](#_Toc58875451)

[2. Data 3](#_Toc58875452)

[2.1 Data Cleaning 4](#_Toc58875453)

[2.2 Exploratory Data Analysis 6](#_Toc58875454)

[3. Analysis 7](#_Toc58875455)

[3.1 Linear Regression 7](#_Toc58875456)

[3.2 Naïve Bayes 8](#_Toc58875457)

[3.3 Random Forest 9](#_Toc58875458)

[4. Results 10](#_Toc58875459)

[5. Conclusion 11](#_Toc58875460)

[Appendix 12](#_Toc58875461)

[Appendix I: Table of Missing Values in the Initial Dataset 12](#_Toc58875462)

[Appendix II: Table of Missing Values in the Initial Dataset 13](#_Toc58875463)

[Appendix III: Heatmap for Correlation 14](#_Toc58875464)

[Appendix IV: Average Price of Listings by Manufacturer 15](#_Toc58875465)

[Appendix V: The Number of Listings Across Each State 15](#_Toc58875466)

[Appendix VI: Mutual Information Table 16](#_Toc58875467)

[Appendix VII: Scatterplot of Price and Odometer 17](#_Toc58875468)

[Appendix VIII: Scatterplot of Price and Model Year 17](#_Toc58875469)

# 1. Introduction

## 1.1 Project Overview

Buying a used car online is easier than it has ever been. In 2019, there were almost 41 million used cars sold in the United States, compared to only 17 million new car sales.[[1]](#footnote-2) Additionally, the coronavirus pandemic has made the used car market surge to levels it hasn’t seen in over 50 years (See Appendix I).[[2]](#footnote-3) Used cars are often more affordable of an investment than purchasing a brand-new vehicle from the dealership, which partially explains why there is consistently a larger volume of used cars purchased per year than new cars; other reasons include purchasing “project” cars, purchasing rare cars that are hard to find, and purchasing older vehicles that are not made anymore. With such a large and growing demand for used cars, our team sought to analyze what features of a used car tend to increase its resale price.

We used car listings on Craigslist in order to build our predictive models and draw conclusions about important features that increase a car’s resale value. The data we used contained a car’s list price (the dependent variable), the car manufacturer, the model year, the odometer reading, and other information about the car. We decided to predict the dollar amount of a used car listing using linear regression, as well as classify a used car as “very low”, “low”, “mid-high”, and “high” priced using a Naïve Bayes and Random Forest algorithm.

Overall, two of our models performed quite well, with the linear regression showing a strong correlation with the predictors and the random forest giving a decent overall accuracy; unfortunately, Naïve Bayes did not perform as well. We attempted using just the most important variables to build our machine learning models (by using the variables that shared the most mutual information with the price category), but this ended up reducing the overall accuracies.

## Background/Business Question

Craigslist is a very popular method of buying and selling used cars, as it can lead to higher sales prices (compared to trading in to a dealer), it is relatively cheap to list a car, and it opens your listing to a huge audience.[[3]](#footnote-4) Craigslist offers individuals the ability to go in person to purchase something, something that is quite important when purchasing a car, and something that other platforms lack. When an individual sells their car on Craigslist, it is up to them to create the listing, and this takes time and effort. The seller may not have all of the information available, or not feel inclined to research the information necessary to make a listing; the more information, however, the better.

We wanted to analyze the different variables that contribute to the price of a used car in order to build models that accurately predict a given car’s price/price range. With increasing demand for used cars due to the effects of coronavirus and the lockdown measures, our team felt that analyzing this data would be pertinent.[[4]](#footnote-5) This would be of use for sellers, as it is often difficult to determine what exactly a used car is worth. Additionally, these models and prediction tools would be useful for buyers, as they could determine if a given car listing is overprice, a good deal, or about right.

# 2. Data

We used the “Used Cars Dataset” from Kaggle, which contains information from used car listings webscraped from Craigslist; the data can be found here: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>.

In total, the dataset contains 25 different variables on 423,857 different used car listings. The variables include information on the make of the car, the size of the vehicle, the transmission type, the reading on the odometer, and other relevant attributes.

## 2.1 Data Cleaning

Our exploratory data analysis began by dropping columns that would be unnecessary for our analysis, such as the listing URL, the car’s VIN, the county the car was being sold in, and the latitude and longitude of the listing. We then proceeded to look at the number of missing values in each column (See Appendix II). The variable “size”, which contained information about whether a car was mid-size, full-size, etc., only contained information for 75.787% of the data, so we decided to drop this column entirely.

We noticed that the column “cylinders”, which listed the number of cylinders for each car, was structured as a string, such as “8 cylinders”, instead of just the integer. To use this column for our analysis, we used regex to strip the text from each entry and only leave the integer value.  
 Upon looking at descriptive statistics of our continuous variables, we noticed that the maximum value of the “price” column was quite high (3,808,256,046.00). Likewise, we noticed that the maximum value for the “odometer” column was abnormally high (10,000,000.00). With such extreme values in each column, we decided to investigate and fix some of these errors. We were able to impute the maximum in the “price” column by reading the description for that listing, where the seller stated that they were asking for $3,000.

By plotting a bar graph that shows the mean price of listings for each manufacturer, however, we saw that there were still some errors in the data. The graph shows that GMC and Dodge have the highest average price of listings for all manufacturers; this did not make sense to us, as these are not known to be “luxury” manufacturers. We took another look at the “price” column and decided to drop a few more rows which contained obvious erroneous entries. We also decided to remove any listing that was above $300,000 or below $100, as these are likely outliers or spam ads.

We then decided to delete the values of the “odometer” column which were above 999,999 miles, as this is the maximum value that a modern odometer will show. Since Americans drive 13,500 miles on average[[5]](#footnote-6), we decided to impute the car’s age multiplied by 13,500 for any value where the odometer was above 999,999 miles (If this ended up being above 999,999 miles, then we just imputed 999,999).

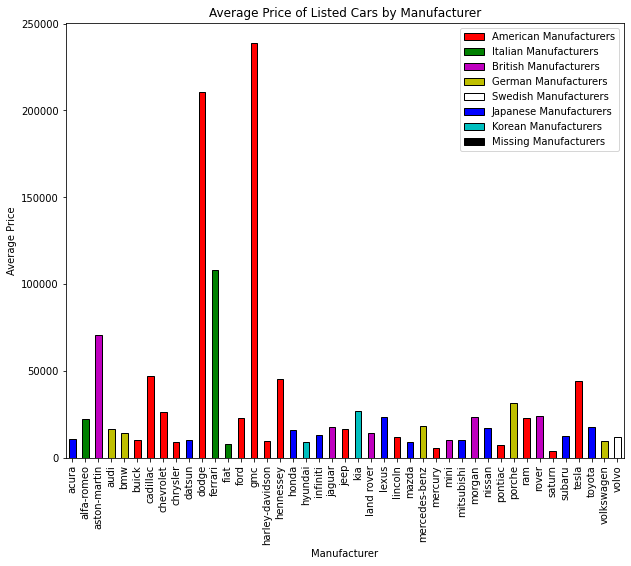


Figure 1: The Average Price of listings for each manufacturer before cleaning.

After cleaning up some the data, we decided to remove all missing values from the original dataset. This accounted for just over 75% of the data, but due to the large size of the original file, we were left with almost 100,00 rows after removal. We checked the distribution of some key variables and found a similar distribution of the variables across the data with missing values as well as the data without missing values. We concluded that removing this data was justified, and we determined that it would not bias our results.

## 2.2 Exploratory Data Analysis

Once the data had been cleaned thoroughly, we conducted some exploratory data analysis to see how our data was distributed and correlated. We found there to be a decent amount of correlation of the numerical variables “odometer”, “year”, and “cylinders” with the dependent variable “price” (See Appendix III). While British cars seemed to have the highest average list price, the most expensive brand on average was the Ferrari (See Appendix IV).

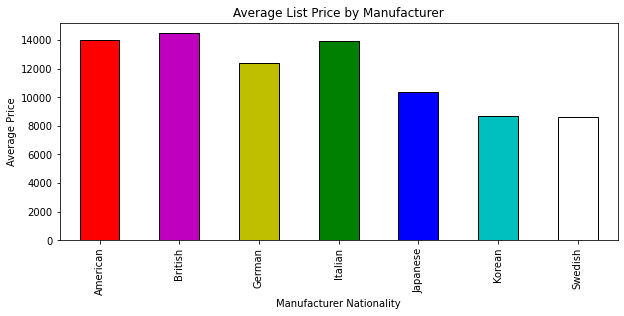


Figure 2: The average list price by manufacturer nationality.

We found that across all manufacturers, the average price of a used car increases the newer it is. This makes sense, as newer cars are “fresher”. Along the same lines, cars that were listed with condition “new” tended to have the highest list price on average. We found there to be a large variation in the number of cars for sale in each state; California, Florida, and New York appeared to be the biggest markets, while Delaware, North Dakota, and Utah appeared to be the smallest (See Appendix V).

# 3. Analysis

## 3.1 Linear Regression

In order to predict the dollar amount of a given car listing, we decided to construct a linear model by regressing the price of each listing on all the independent variables. To do this, we constructed k-1 dummy variables for each categorical variable with k categories by using the pandas method pd.get\_dummies(). We began by looking at the distribution of the listing price for each entry and found that it was heavily skewed right.

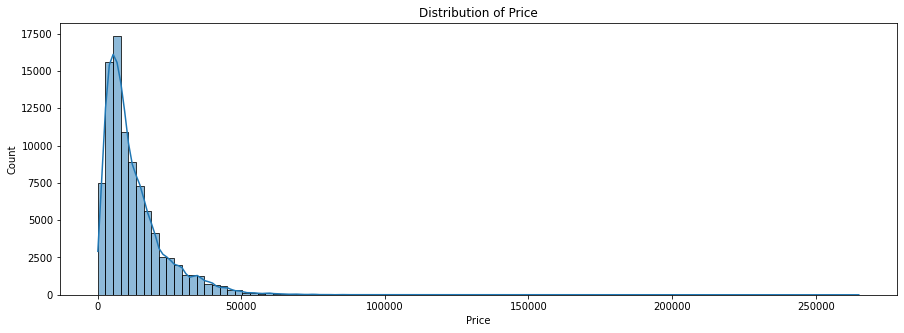


Figure 3: The distribution of price.

After running a linear regression on the data, we found there to be a mean squared error of 32,690,580,880.45, a correlation coefficient of 0.631, and an R2 of 0.399. Based on these results, we determined that we needed to transform our depended variable, “prices”, in order to improve our model results. To achieve this end, we opted for a log transformation, where we simply take the natural logarithm of each price and regress the dependent variables on this transformed data. After completing the transformation, we found the distribution to be much more symmetric.

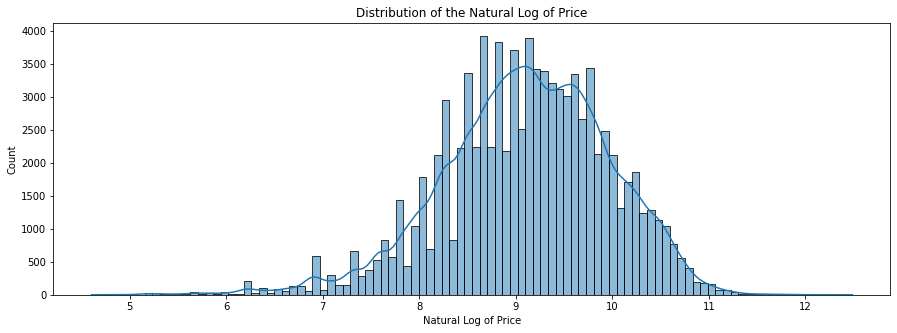


Figure 4: The distribution of the natural logarithm of price.

We reran the linear regression, this time with the natural logarithm of price as our dependent variable and obtained much better results. Our transformed regression yielded a mean squared error of 399.63, a correlation coefficient of 0.883, and an R2 of 0.78. Clearly the results of the regression on the transformed data are much better than the results of the regression on the untransformed data.

## 3.2 Naïve Bayes

For our machine learning models, we opted to use a Naïve Bayes classifier as the “baseline” model to compare our Random Forest model to as it is relatively easy to implement and very quick to run.[[6]](#footnote-7) Naïve Bayes works by calculating the conditional probability of a specific output class given input data; it utilizes Bayes’ Theorem:

In order to run our Naïve Bayes model, we discretized “price” into four bins based on the 25th, 50th, and 75th percentiles. We used 67% of the data for training, and the remaining 33% for testing our model; additionally, we used 10-fold cross validation to evaluate our model performance. We will report the results of our 10-fold cross validation, as we feel these are more accurate and reduce the effects of overfitting.[[7]](#footnote-8) Finally, we ran the model twice: once on all of the variables, and once more after performing feature selection based on the mutual information that the independent variables share with the dependent variables.

On the full dataset, Naïve Bayes yielded an overall accuracy of 45.12% — not great. Unfortunately, after performing feature selection, the results decreased to only 42.69% accuracy. While these results are not great, they are only a baseline, and luckily our results from our Random Forest classifier were much better.

## 3.3 Random Forest

The second of the two machine learning models that we opted for was a Random Forest classification. A Random Forest is an ensemble learning method for classification or regression, and it is easy to implement and gives decent results a lot of the time.[[8]](#footnote-9) Random Forests work by implementing feature bagging, where a randomly selected subset of the entire attributes is selected, and a decision tree is built based on this random subset. Multiple decision trees are made with different random subsets of the features, hence the name *Random Forest*.[[9]](#footnote-10)

Like our Naïve Bayes classifier, we used a 67/33 split for training/test data in addition to 10-fold cross validation; the latter results will be reported. Before feature selection, our Random Forest classifier gave us an overall accuracy of 79.15%, which is quite a big improvement from the Naïve Bayes classifier. However, like Naïve Bayes, the accuracy went down after feature selection, this time to 71.21%. This is not surprising, however, as bagging results in only a subset of features being used, just as feature selection removes some variables from use.

# 4. Results

Our models performed decently overall, with our transformed linear regression model and Random Forest classifier performing the best and our Naïve Bayes classifier performing subpar. While feature selection reduced the performance of both models, it was able to tell us which variables share the most mutual information with the discretized price; in other words, the mutual information tells us how much about one variable we learn by knowing about another variable (See Appendix VI, VII, and VIII).[[10]](#footnote-11)

Our natural logarithm transformation of prices led to a decent linear regression model. This tool could be helpful for sellers who are unsure of exactly what their car is worth. Likewise, this would be useful for buyers, as they can determine if a given listing is over or undervalued. To get a range of possible values, each could opt for using the Random Forest classifier, as instead of giving a set number, it reports a label that corresponds to a range of values.

Unfortunately, there are limitations to our analysis. Firstly, while media reports state that the coronavirus pandemic has led to an increase in demand for used cars[[11]](#footnote-12), we don’t have a way of assessing that in our models. The data we have does not contain any information on the date the listing was made, the length of time the listing was up, or if the car was sold or not. Additionally, we do not have data on web traffic, so we cannot see if there has been an increase in visits to Craigslist ads for used cars.

Secondly, there are bound to be spam ads in the data we have. It would be very tedious to go through each listing to see if it is an ad or not, which is why this was not done in our analysis. Some of the missing values we checked during data cleaning were spam ads, so there are bound to be more. These might bias our results, but unfortunately, we are not able to assess that.

Finally, we do not know when exactly this data comes from. The listing on Kaggle specifies that the data “is updated every few months”[[12]](#footnote-13), but there is little to go on besides this. This drawback ties into our initial one, namely the effect that coronavirus has had on the used car market. Without adequate information about the time these listings were made, it is tough to draw certain conclusions about the impact on the market.

While we believe that all information is necessary to make an honest listing, as well as make an informed decision about a purchase, we found that the mileage on the odometer and the model year to have the greatest impact on the value of a used car. Intuitively, this makes sense, as an older car with more miles is likely to have more wear and tear and ultimately be worth less than a newer car with less miles on the clock. Sellers should list this information when selling a car, and it is pertinent that buyers pay attention to these fields.

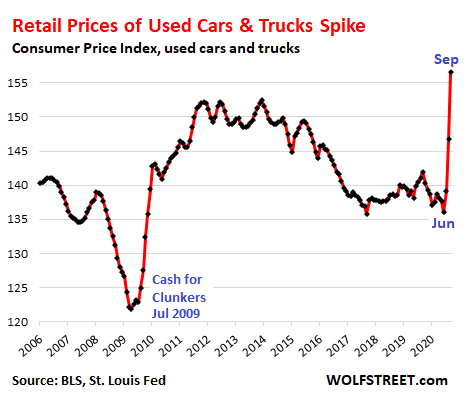
# 5. Conclusion

For our final project for BANA-212, Data and Programming for Analytics, our team sought to analyze used car listings from Craigslist in order to assess the most important factors that affect the listing price of a used car. We constructed a linear model as well as a Naïve Bayes and Random Forest classifier in order to predict/classify the price of a used car listing. The natural log linear regression performed quite well, as did the Random Forest classifier. Overall, we had a fun time using the skills that we learned in class to analyze a relevant topic that so many people encounter in their lives.

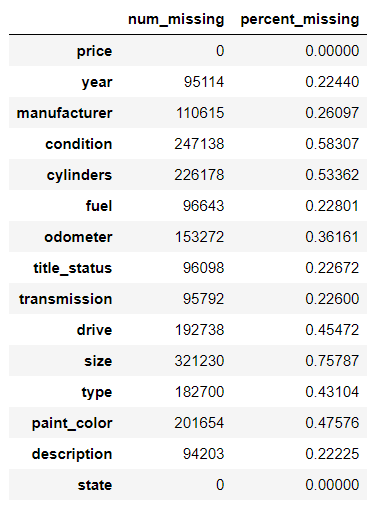
# Appendix

## Appendix I: Table of Missing Values in the Initial Dataset

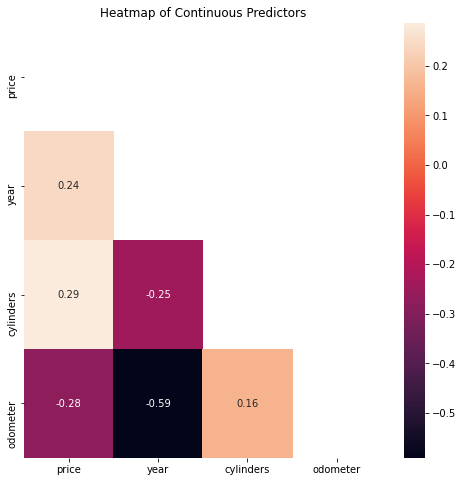
Note: This is not a graph we made, nor it is based off data that was available to us. We found this graph from an online article we used for our research, wolfstreet.com.



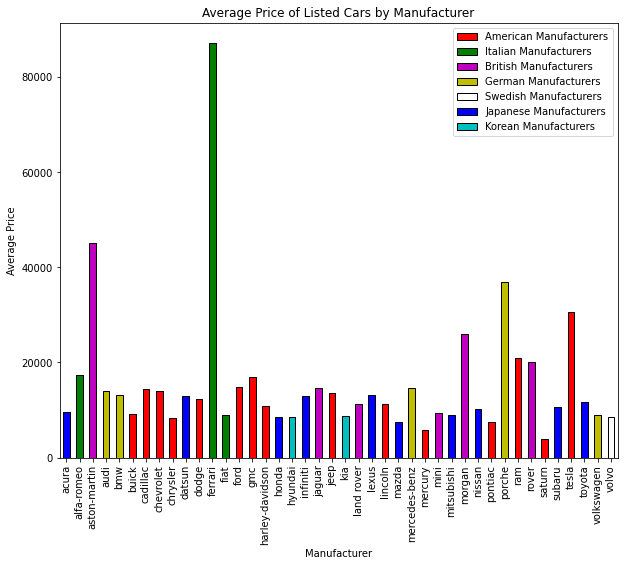
## Appendix II: Table of Missing Values in the Initial Dataset



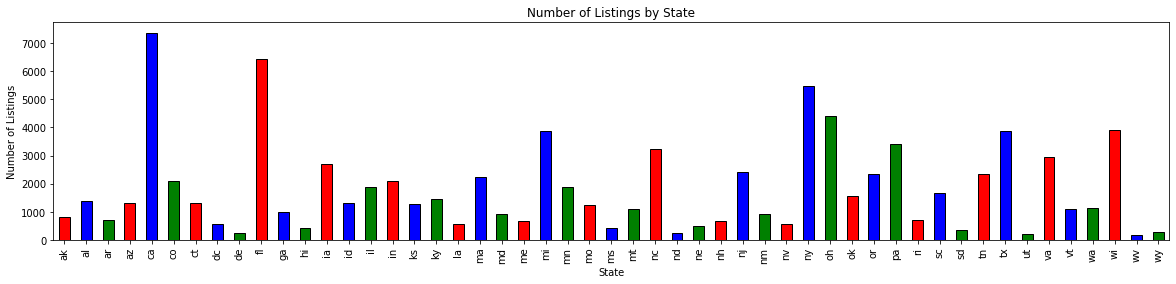
## Appendix III: Heatmap for Correlation



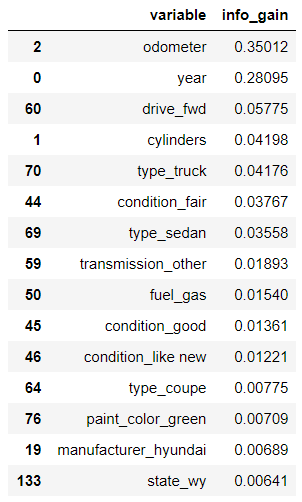
## Appendix IV: Average Price of Listings by Manufacturer



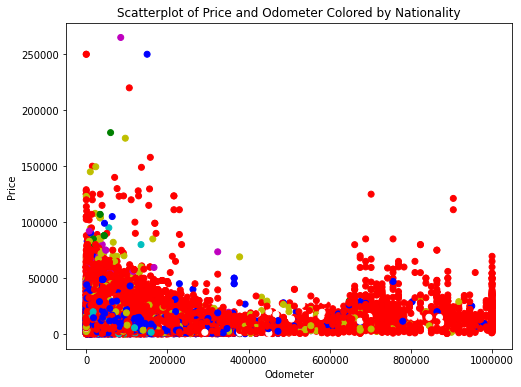
## Appendix V: The Number of Listings Across Each State



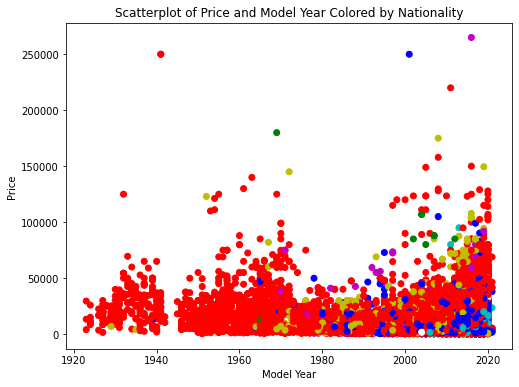
## Appendix VI: Mutual Information Table



## Appendix VII: Scatterplot of Price and Odometer



## Appendix VIII: Scatterplot of Price and Model Year



1. I. Wagner, “U.S. New and Used Car Sales 2019,” Statista, September 30, 2020, https://www.statista.com/statistics/183713/value-of-us-passenger-cas-sales-and-leases-since-1990/. [↑](#footnote-ref-2)
2. “Crazy Times Not Only for Consumers Trying to Buy a Used Car but Also for Dealers Grappling with the Distortions,” Wolf Street, November 18, 2020, https://wolfstreet.com/2020/10/13/used-car-truck-retail-prices-spike-most-since-1969-despite-lower-sales-plenty-supply-months-after-wholesale-prices-spiked-most-ever-but-are-now-falling/. [↑](#footnote-ref-3)
3. Ed Reeves, “The Pros and Cons of Selling Your Car on Craigslist,” MotorBiscuit, October 2, 2019, https://www.motorbiscuit.com/the-pros-and-cons-of-selling-your-car-on-craigslist/. [↑](#footnote-ref-4)
4. Erprose, “The Used Car Boom Is One of the Hottest, and Trickiest, Coronavirus Markets for Consumers,” CNBC (CNBC, October 16, 2020), https://www.cnbc.com/2020/10/15/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers.html. [↑](#footnote-ref-5)
5. Hearst Autos Research, “What Is Average Mileage Per Year?,” Car and Driver (Car and Driver, November 10, 2020), https://www.caranddriver.com/research/a32880477/average-mileage-per-year/. [↑](#footnote-ref-6)
6. BANA-273 Class Lecture, October 21, 2020. [↑](#footnote-ref-7)
7. Raheel Shaikh, “Cross Validation Explained: Evaluating Estimator Performance.,” Medium (Towards Data Science, November 26, 2018), https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85. [↑](#footnote-ref-8)
8. Niklas Donges, “A Complete Guide to the Random Forest Algorithm,” Built In, accessed December 9, 2020, https://builtin.com/data-science/random-forest-algorithm. [↑](#footnote-ref-9)
9. BANA-273 Class Lecture, December 3, 2020. [↑](#footnote-ref-10)
10. Jason Brownlee, “Information Gain and Mutual Information for Machine Learning,” Machine Learning Mastery, December 9, 2020, https://machinelearningmastery.com/information-gain-and-mutual-information/. [↑](#footnote-ref-11)
11. Consumer Reports, “Sell Your Car Now. Yes, Now.,” WJXT (WJXT News4JAX, December 10, 2020), https://www.news4jax.com/money/2020/12/10/sell-your-car-now-yes-now/. [↑](#footnote-ref-12)
12. Austin Reese, “Used Cars Dataset,” Kaggle, December 3, 2020, https://www.kaggle.com/austinreese/craigslist-carstrucks-data. [↑](#footnote-ref-13)