Determining Used Car Prices

Zack DeNoto

Bellevue University

DSC 680: Applied Data Science

Dr. Fadi Alsaleem

April 4, 2021

Abstract

The problem that this paper addresses is how does a car shopper know what a good price is for a used car, and what can a car shopper use to help determine a price? This paper will demonstrate if it is possible to estimate the price of a used car based on several factors provided in the dataset, as well as determine what factor(s) most affect the price of a vehicle. Though the main dataset used in the analysis pertained to cars in Europe, the main concepts also apply to those for cars in the US or another part of the world. Many car manufacturers make cars and ship them globally. The design of the project was to find a good dataset or several, clean the data, determine what factors affect price the most, and create the most accurate model(s) possible.

Intro

Each year millions of cars are sold globally. In fact, the number of automobile sales has been increasing significantly with an estimated 74-78 million vehicles sold around the world every year. In the US, the bestselling car sold just under 900,000 units in 2019 (Kopestinsky, 2021). The average price of new vehicles is up 6.8% from 2016 to 2019 as seen in image 4 of the appendix (Statistica, 2021). With the increase in vehicle price and depreciation many more people are buying used cars compared to new cars (Patrick, 2020). Second-hand cars or used cars, however, have increased twofold compared to new cars in the US. With the Covid-19 pandemic hitting in 2020, there was a decrease in the number of new cars being sold. Even though there was a decline in new cars being sold, the used car market increased, especially in the summer due to fear of contracting the virus in public transport or in rideshares from those in the age range of 18-35. According to a chart on Cargurus.com, the price of used cars went down when Covid-19 first hit in early 2020, but then made an upward trend peaking in December 2020 and is currently the highest it has been in over a year by about 10% (Cargurus.com, 2021). Looking specifically at the last month of March 2021 in comparison to March 2020, the used vehicle value index has increased 23.7% in just one year (Manheim.com, 2021). The used car market grew from $77.8 billion in

2009 to $117.9 billion in 2019 (Mitic, 2021). Reading about the increase in the used car market made me want to research what factors affect the price of used vehicles and if I could determine the price of a car based on data.

Data

The main dataset used for this analysis was called Used-cars-catalog from Kaggle.com, retrieved from the link https://www.kaggle.com/lepchenkov/usedcarscatalog. This dataset contains information regarding cars from one of the most popular online catalogs in the country of Belarus. The csv file had 30 columns with fields ranging from manufacturer_name, model_name, transmission, color, odometer_value, year_produced, number_of_photos, and duration_listed. In addition to the main dataset, there were two additional datasets which were used to help compare and find what factors most affect used car prices. The dataset is called US Cars Dataset and was retrieved from the following link: https://www.kaggle.com/doaaalsenani/usa-cers-dataset. This dataset was scraped from an online car auction website and has information on 28 different brands of used vehicles in the US for sale. The dataset has 12 columns, many sharing the fields of the main dataset and contained fields such as brand, model, year, title_status, mileage, and color. The third dataset is called Used Cars Dataset from the following link: https://www.kaggle.com/austinreese/craigslist-carstrucks-data. This dataset has data that was scrapped from Craigslist, which has the world's largest collection of used vehicles for sale. The dataset has 26 fields such as price, manufacturer, model, year, odometer, paint_color, state, and posting_date. Craigslist is a difficult website to scrape data from, and I knew based on the extra fields that this dataset was going to be the hardest one to clean up.

Methodology

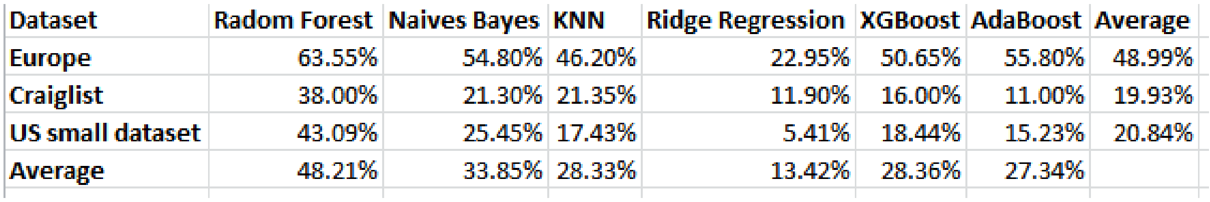
As many people know, a lot of what data scientists do is data retrieval and data cleaning. In fact, this can take up to 80% of a data scientist's time. The first step that needed to be done was to clean the data from each of the datasets. For the main dataset, since the data was from used cars in Belarus, some of the car manufacturers needed to be removed as I wanted to look at car manufacturers that were shared between the 3 datasets. When looking at the car manufacturers in R, some of the data displayed incorrect manufacturers such as "ÐœÐ¾Ñ\u0081ÐºÐ²Ð¸Ñ‡" or "Ð£Ð\u0090Ð—." These were some of the manufacturers that needed to be removed as seen in image 2 of appendix. The odometer readings were in kilometers and had to be converted to miles to ensure all the datasets were using the same odometer readings. Many columns needed to be removed because they were not necessary to the analysis, such as feature_0 through feature_9; they were factors such as air conditioning, window tint, etc. The second dataset on US vehicles was a clean dataset overall and did not require a lot of cleaning. The first step was to examine the data to check for outliers or incorrect data. For example, there were some high odometer readings that did not make sense. To correct this, I only looked at odometer readings if they were less than 500,000 miles. Then unnecessary columns such as the VIN or title status of a vehicle were removed for the analysis. For the third dataset on cars from Craigslist postings, it was very similar to the second dataset in which unnecessary columns were removed such as VIN, posting date, and description. This dataset had some rows with missing years and some odometer readings that did not make sense with values over 20 million miles, so they were then removed. In addition, there were motorcycle postings being included in the dataset which were removed. As someone who has bought and sold many items on Craigslist, I know that many times a seller will use inaccurate information to try to get his ad noticed through tactics such as listing a car with 125,000 miles as 125 miles. To reduce this, I removed rows with mileage less than 999 miles.

Once the data was cleaned up, I looked at the correlation between fields compared to price for every dataset. After correlations were looked at, I used OneHot encoding to turn categorical values into numerical data for modeling. For modeling, Random Forest, Naives Bayes, K-Nearest Neighbor, Ridge Regression, XGBoost, and AdaBoost were used. Due to the very large datasets, I had to take a sample of 10,000 rows for 2 of the datasets due to computational limitations.

Results

Based on the models chosen for this analysis, I did not have any preconceived notions of what model would perform the best or the worst or what the results would be for any of the datasets. I came in with an open mind, hoping to be surprised. I knew that the accuracies were going to be very low going into this analysis based on the data. The datasets chosen have thousands of rows with used car prices, so to accurately predict the price of a used car to the exact dollar was unrealistic. In order to come up with an accuracy for the analysis, I picked a range of +/- 2,000 of the actual price. For example, a model such as AdaBoost may have had an accuracy of 2.75% but within $2,000 the model was 55.8% accurate.

In order of most accurate to least accurate for models were Random Forest, Naives Bayes, XGBoost, K-Nearest Neighbor, AdaBoost, and lastly, Ridge Regression.



Dataset	Radom Forest	Naives Bayes	KNN	Ridge Regression	XGBoost	AdaBoost	Average
Europe	63.55%	54.80%	46.20%	22.95%	50.65%	55.80%	48.99%
Craiglist	38.00%	21.30%	21.35%	11.90%	16.00%	11.00%	19.93%
US small dataset	43.09%	25.45%	17.43%	5.41%	18.44%	15.23%	20.84%
Average	48.21%	33.85%	28.33%	13.42%	28.36%	27.34%	

The image above shows the accuracies of the various models vary significantly depending on the model and the dataset being used. The most accurate model was Random Forest for the main dataset of European used cars, with an accuracy of 63.55%. The least accurate model was Ridge Regression on the second dataset of American used cars with an accuracy of only 5.41%. This dataset was the smallest, but the accuracy was significantly worse than the other accuracies for the dataset. From these accuracies it appears that Ridge Regression is the least accurate model for the datasets compared to Random Forest, which is over three times as accurate.

The results of the analysis were very surprising. I did not expect Random Forest to be so much better of a model compared to Ridge Regression. One of the questions which I had to ask was why was the dataset from the European used cars more accurate compared to the two US datasets? If we look into the correlations from the EDA analysis in R, we can see that the correlation from year and price of car is 0.868. This indicates that in the European used car market, the price of a used car is highly correlated to the amount of kilometers on the car. In the two US datasets, the correlation between year and price was only 0.527 and 0.491, which were significantly lower compared to the main European dataset. This correlation is one of the reasons the main dataset had more accurate models. The dataset on US auction cars was very small, with only around 2,400 rows after cleaning the dataset and removing bad data. This is what helped to create inaccurate models; the other datasets had to be cut off at 10,000 rows. If there was more data in the US auction dataset, I suspect that the models would have been more accurate. I also suspect that the accuracies of the main European dataset and the Craigslist dataset would have been increased if we had more rows used for modeling instead of the 10,000 I was forced to use.

Conclusion

Going into this analysis I did not know if it would be possible to get the price of a used car based on the data, as selling cars can be very subjective. After the analysis and looking at the models I believe that it is possible to get a fairly accurate price range of a used car based on several factors if the right model is used. You need a large clean dataset and, based on the models tested in this analysis, you need to use Random Forest to get the most accurate price. In this analysis the computer system limited the modeling, as all rows could not be tested. If all rows were used in modeling, I believe the accuracies would have been higher than the 63.55% that the highest accuracy achieved. I believe this is a good result and could be used to help determine the price of a used car. In Europe, it appears that the number of kilometers the car has been driven is the highest factor for a used car price. Contrastingly, in the US, mileage does not play as much of a factor and is more of a mix of factors. This result matches a similar analysis by Schibsted.com where they compared the price of a new car and a used car in Europe. Their analysis showed that including the mileage of a car significantly increased the accuracies of the models (Schibsetd, 2020). After using KellyBlueBook.com to determine the price of used cars, I looked at if a car with the same mileage but one year age difference had a larger price difference than two used cars of the same age with the only difference being mileage. The age of a car had a larger impact in price compared to the mileage of car, as seen in image 3 of appendix (KellyBlueBook.com, 2021). This contrasted the information from Investopedia, which stated that the main factor affecting a used car price is the mileage (D'Allegro, 2020). In addition, according to Instamotor, mileage is the number one key factor for affecting the price of a used car and it did not mention the age of the car. In conclusion, it is very difficult to fully determine the price of a used vehicle as the condition of the car on the interior, exterior, and mechanically is too difficult to determine from a dataset. However, based on the factors in the datasets used in the analysis it is possible to create a fairly accurate model using Random Forest with enough data.

References

Kopenstinsky, A. (2021). 20 In-Depth Global and US Auto Sales Statistics for 2021. Retrieved from https://policyadvice.net/insurance/insights/us-auto-sales-statistics/

Cargurus.com. (2021). Used Car Trends. Retrieved from https://www.cargurus.com/Cars/price-trends/

Mitic, I. (2021). US Car Sales Statistics: Figures, Trends, and Historical Data. Retrieved from https://fortunly.com/statistics/us-car-sales-statistics/

Schibsted.com. (2020). Price your car with data. Retrieved from https://schibsted.com/blog/price-car-data/

KellyBlueBook.com. (2021). My Cars Value. Retrieved from https://www.kbb.com/audi/a3/2015/18t premium-sedan-4d/?condition=good&intent=trade-in-sell&mileage=120000&modalview=false&options=6393602%7ctrue&pricetype=trade-in&vehicleid=399177

Manheim.com. (2021). Used Vehicle Value Index. Retrieved from https://publish.manheim.com/en/services/consulting/used-vehicle-value-index.html

Statistica.com (2021). New vehicle average selling price in the United States from 2016 to 2019. Retrieved from https://www.statista.com/statistics/274927/new-vehicle-average-selling-price-in-the-united-states/

Patrick, K. (2020). The top 10 cars that hold their value. Retrieved from https://www.autoguide.com/auto-news/2019/11/top-10-cars-that-hold-their-value.html

D’Allegro, J. (2020). What is the value of your used car? Retrieved from https://www.investopedia.com/articles/investing/090314/just-what-factors-value-your-used-car.asp

Instamotor.com. (2021). The three things that affect the price of a used car. Retrieved from https://instamotor.com/sell-car/car-value/3-key-things-that-affect-the-price-of-a-used-car

Appendix

Image 1. From https://www.cargurus.com/Cars/price-trends/

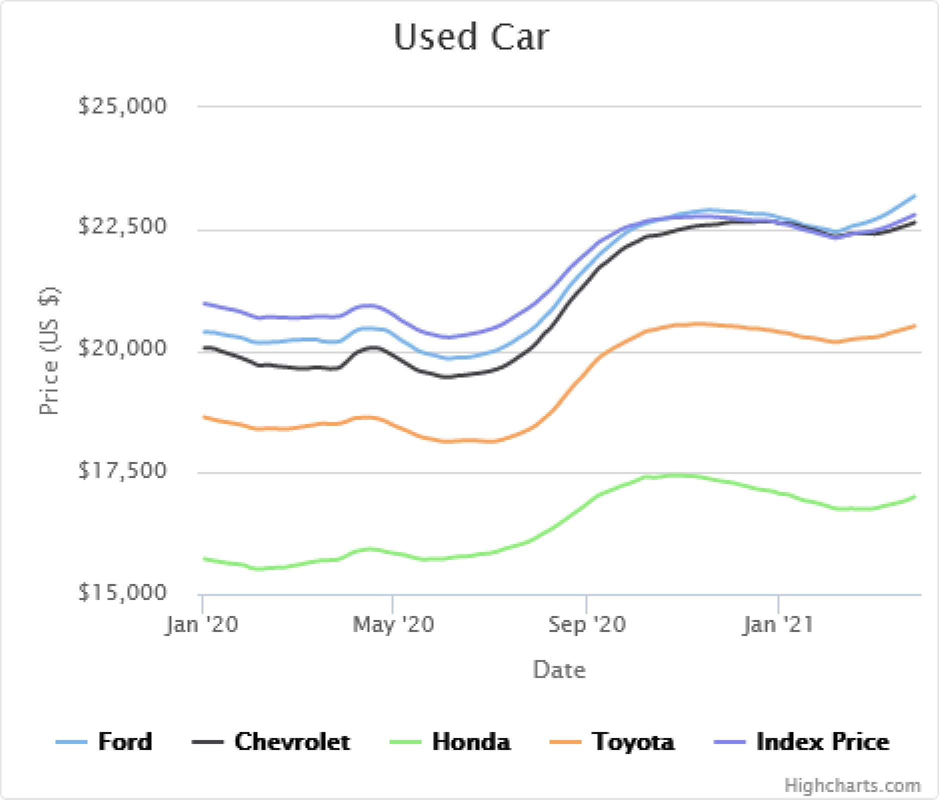
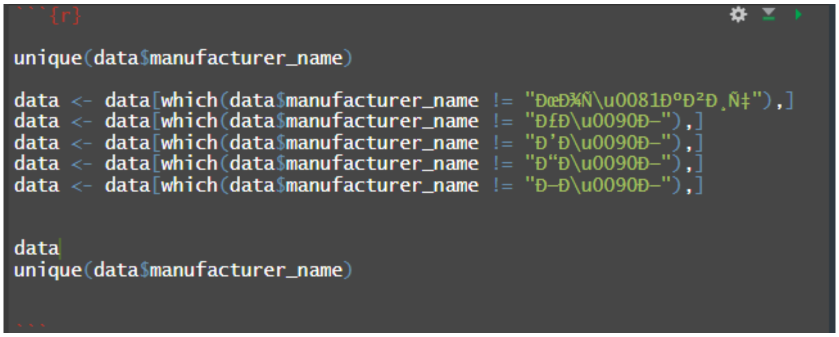
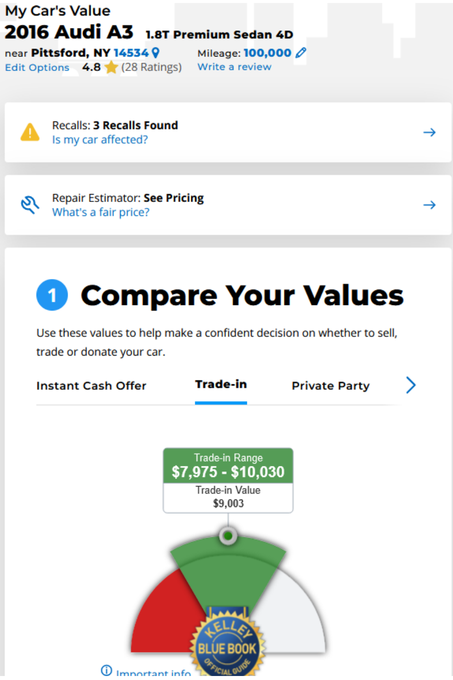


Image 2. From R analysis



~~~~
```{r}

unique(data$manufacturer_name)

data <- data[which(data$manufacturer_name != "ÐœÐ¾Ñ\u0081ÐºÐ²Ð¸Ñ‡"),]
data <- data[which(data$manufacturer_name != "Ð£Ð\u0090Ð—"),]
data <- data[which(data$manufacturer_name != "Ð’Ð\u0090Ð—"),]
data <- data[which(data$manufacturer_name != "Ð“Ð\u0090Ð—"),]
data <- data[which(data$manufacturer_name != "Ð—Ð\u0090Ð—"),]


data
unique(data$manufacturer_name)


```
~~~~

Image 3. From KellyBlueBook.com



My Car's Value

2016 Audi A3 **1.8T Premium Sedan 4D**

near **Pittsford, NY** 14534 Mileage: **100,000**

Edit Options **4.8** ★ (28 Ratings) Write a review

Recalls: **3 Recalls Found**
Is my car affected?

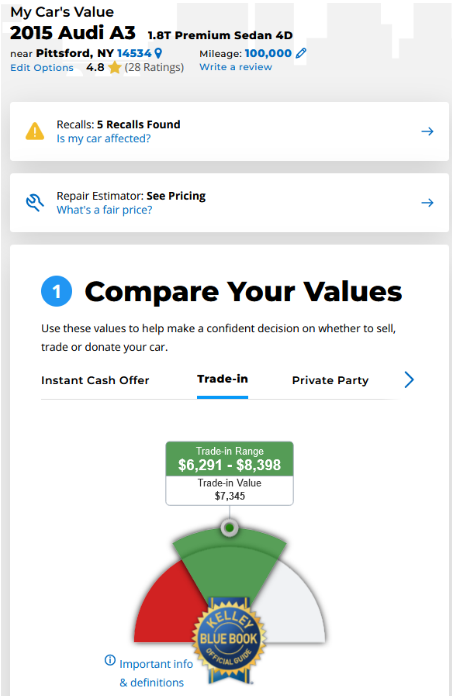
Repair Estimator: **See Pricing**
What's a fair price?

1 Compare Your Values

Use these values to help make a confident decision on whether to sell, trade or donate your car.

Instant Cash Offer **Trade-in** Private Party

Trade-in Range
$7,975 - $10,030
Trade-in Value
$9,003



My Car's Value

2015 Audi A3 **1.8T Premium Sedan 4D**

near **Pittsford, NY** 14534 Mileage: **100,000**

Edit Options **4.8** ★ (28 Ratings) Write a review

Recalls: **5 Recalls Found**
Is my car affected?

Repair Estimator: **See Pricing**
What's a fair price?

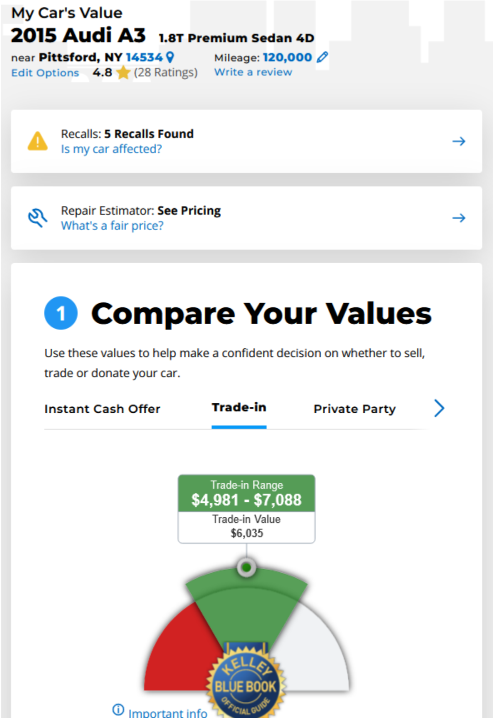
1 Compare Your Values

Use these values to help make a confident decision on whether to sell, trade or donate your car.

Instant Cash Offer **Trade-in** Private Party

Trade-in Range
$6,291 - $8,398
Trade-in Value
$7,345

Image 3(continued).



My Car's Value

2015 Audi A3 **1.8T Premium Sedan 4D**

near **Pittsford, NY** 14534 Mileage: **120,000**

Edit Options **4.8** ★ (28 Ratings) Write a review

Recalls: **5 Recalls Found**
Is my car affected?

Repair Estimator: **See Pricing**
What's a fair price?

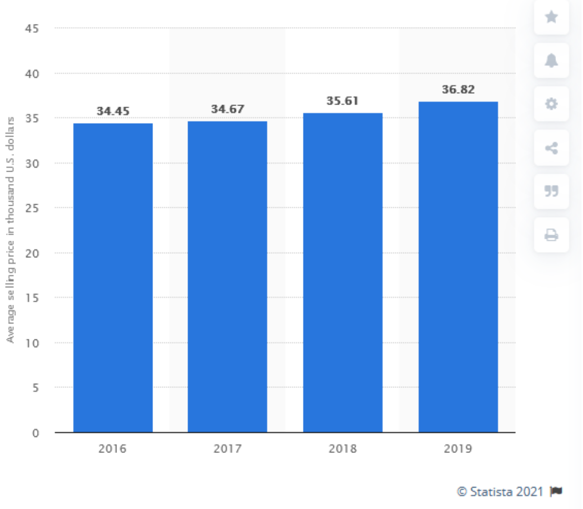
1 Compare Your Values

Use these values to help make a confident decision on whether to sell, trade or donate your car.

Instant Cash Offer **Trade-in** Private Party

Trade-in Range
$4,981 - $7,088
Trade-in Value
$6,035

Image 4. From https://www.statista.com/statistics/274927/new-vehicle-average-selling-price-in-the-united-states/


Average selling price in thousand U.S. dollars
© Statista 2021