**Project Overview**

Dingyu Sun, Irene Lu, Peilin Zhong, Xiaoben Yin

SYRACUSE UNIVERSITY

**Grade: 100%**

Craigslist Used Car Price Analysis

IST 718 GROUP 14

Owning a car is seen as a source of pride and accomplishment in the United States. On average, there are 1.88 vehicles per U.S. household. According to the U.S. Department of Transportation, the percentage of households without a car or light truck came to around nine percent in 2017, meaning that about 90 percent of households had at least one light vehicle at their disposal in that same year.[1] While owning a new car is many peoples’ options, many customers turn to seeking for used cars due to the increasing price of new cars and limited budget. However, the price of a used car may vary since it is affected by many attributes: year of the car, model, region for sale, etc. In this project, we want to study the market price for used cars and help customers identify the actual value of a certain used car.

**Dataset Summary**

From all the car listing websites, we choose Craigslist used car dataset for our study. Craigslist is an American classified advertisements website with sections devoted to jobs, housing, for sale, items wanted, etc. It also includes the ability for users to focus on a specific area or city. [2]

This dataset contains web scraping data on Craigslist from September 2020 containing most all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories. You can find the dataset on Kaggle: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

This dataset has 423,867 rows and 25 columns. Since we are predicting the price of used cars, we identify a list of attributes that contribute to the change of price (predictors). The list of attributes are region, year, manufacturer, model, condition, cylinders, fuel, odometer, title\_status, transmission, drive, size, type and paint\_color.

One interesting fact we found about the dataset is that it has many cars with $0 or $1 value. The distribution of price is extremely skewed and more than 75% of used car price lies under $20,000.

**Data Exploration**

Now it is time to explore the data. We start this by making a statistical summary and histogram of our target variable, price. The price is so skewed that we can hardly see any shapes that lie beneath. While the price range goes from -$158 to $203,947,154, the mean value of price is $29,802. In order to find out the undercover data distribution, we decide to take a subset of good-quality data and do further data exploration.

First, we create a subset with attributes that would contribute to price. We exclude id, url, region\_url, vin, image\_url and description.

Then we drop extreme values in price since they may inhibit the prediction power of our future model. We set the upper boundary for price to be $84,500, which is at the 99.7% percentile and lower boundary to be $1000, 5% percentile. Although 5% seems a lot, we make that decision after researching on several other online car sale websites including Carmax[3] and Kbb[4]. We find out that there are barely any cars under $1000 and the reason why Craigslist has 5% of the price values under $1,000 is because many of the listing prices are $0 and $1.

The final step is to retain a good quality subset of sample data. We only want to look at the record if it has sufficient amount of non-empty data to support the price and remove the record if it contains too many NAs. We set up the non-NA threshold to be 13. Any record that contains less than 13 attributes will be removed from the subset so that we end up with a good quality sample for data visualizations.

With the good-quality sample data, we create a boxplot for price (left plot). You can tell from the graph that the data is extremely skewed as median lies somewhere around $10,000. On the right you can see a scatter plot of “price” vs “year of car”. We can see that there is a positive

Chart, scatter chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedrelationship between year and price. As the year gets closer to 2020, there is an increasing trend for car price.

Chart, bar chart

Description automatically generatedAnother interesting thing to look at is the different car types in the market. On the right is the bar chart of car types. While Sedans and SUVs are the two most popular types of car, trucks and pickups take up a large portion of used cars as well. We may drill down further on types of cars and see if there is overlap between categories or if we can combine some types together.

**Predictions**

As a customer who is interested in buying a used car, one key question they may ask is how to determine whether the listing price is fair or not. To narrow down to more specific predictions, we have:

1. In the Craigslist used car market, is this listing used car under market value, equal to market value, or over market value?
2. What is the accurate expected price for a specific used car on sale?
3. What is a good time for a customer to buy a type of used car and resell it in the future?

**Inference**

From the business questions, we come up with several insights that we can derive from the data:

1. Which factors (predictors) mainly contribute to the used car entry price?
2. What is the relationship between entry price and each factor (predictor)?
3. How can the relationship be adequately summarized？
4. What percentage in entry price is associated with manufacture / type / condition?
5. How long will a used car hedge against depreciation if it is a Japanese car compared to the same features of a Germany car?

**Non Spark Packages**

We do not have non Spark packages other than listed in the specification.

**References**

[1] Wagner, I. “Number Cars per Household in the U.S.” Statista, 28 Apr. 2020, [www.statista.com/statistics/551403/number-of-vehicles-per-household-in-the-united-states/](http://www.statista.com/statistics/551403/number-of-vehicles-per-household-in-the-united-states/).

[2] Wikipedia

[3] <https://www.carmax.com/>

[4] <https://www.kbb.com/>