Name: **Muley, Tushar**
Assignment: **DSC 680 - Week 9 Project Proposal Milestone 1**
Date: **February 13, 2021**

Craigs List Auto Sales

Contents

Topic:

The pandemic has changed the world in many different ways. New and used car prices have fallen off the deep end. Seller asking crazy prices for cars in a normal year would be called junkers. It appears there is no such thing as an affordable car market left. What can we learn about the current and future of the used car market, by performing data analysis and predictive modeling on Craigslist used car advertising site?

Questions to research:

The data was found on Kaggle.com. It contains data for April and May of 2021. I really thought car prices would come down some by now. But it appears every month car prices are going up. I am planning to resolve the below research questions from the data.

1. What model years are most frequent come up for sale?
2. Which manufactures are seeing highest average sale prices by state?
3. What is the average odometer reading?
4. What is the most common condition listed?
5. Predicting the sale price of similar vehicles?
6. What states are seeing the largest used car sales?
7. Which make of cars are be sold?

This next question is a little bit of a stretch since it will be difficult to determine which are dealer advertisement compared to sale by owner listings. I am going to list it as an optional question to be completed if the above six questions and the predictive model are complete in time.

- Determine which listings are from auto dealers compared to sale by owner as the seller of the vehicle?

Dataset:

The data set was obtained from Kaggle.com which was surprising. Another individual doing a project for school created a scrapper to pull prices from Craigslist.org for all vehicles that have been listed for sale in the United States. Craiglist.org created by Craig Newmark back in 1995 is a classified advertisement website that has various section devoted to jobs, vehicle for sale, items wanted, services and other sections. It is based in San Francisco, California. The service is currently available in 70 different countries. The data appears to have April and May 2021 information with posting date of vehicles for sale. It is the most current version of the data which is version 10.

Data obtained from: https://www.kaggle.com/austinreese/craigslist-carstrucks-data/version/10

Methods:

1. Perform analysis for the data sets
    a. Pull samples from the Craigslist data
    b. Determine missing data values
    c. Assess values types
        i. Remove Nulls or updated them depending on which column data is missing
        ii. Removing missing rows
        iii. Drop columns like latitude and longitude as they are not needed. Drop vin, region_url, url, county, and image_url columns
        iv. Split date and time data into their own columns (date and time is in single column)
        v. Update numeric strings to int
        vi. Check frequency of the data (count of the attributes)
2. Perform graphical analysis (histogram, scatter plots and correlation) to better understand what kind of data I am dealing with.
3. Build any visuals that are need for the analysis questions
4. Collect/Connect additional data if needed (possible dates)
5. Based on the data select a few models to predict sale prices.
    a. Multi-variable Linear Regression
    b. Decision Tree Regressor
    c. Random Forest Regressor
    d. Ordinary Least Squares
6. Run model
    a. Review data
    b. Determine changes
    c. Build visuals of the data story telling
7. Pull all my analysis together
    a. Polish visuals with correct titles, axis labels and colors
    b. Prepare PowerPoint presentation
    c. Prepare script for presentation
    d. Complete presentation

Ethical and Other Considerations:

This data is publicly available due to the nature of the advertising. The data does not contain any personally identifiable information. As Craigslist does try its best to warn posters and setup email address to mask actual email address of the posters. Phone numbers being displayed in posts are frowned upon and warns are given by Craigslist. But that sometimes does not prevent poster from providing some personal data in hopes the item can be sold quickly.

When I provide the data for further analysis I will try to remove phone numbers in the description field or remove that column once I have pulled information I need for the analysis.

If more attributes could be pulled like posting removal date or scraping the seller type that would help with the analysis quite a bit. Craigslist is pretty popular compared to other sites as some charge 50 dollars or more to post on their website. Depending the state dealership like to post on Craigslist for additional exposure, even if they have their own sites. This show how popular Craigslist can be.

I believe the data can be used for future price predictions even if limited to two months. This would help seller and buyers alike. On the seller side they could determine better range to sell their vehicle for the most amount of money. On the buyer side this would assist in tempering expectation of what they can pay for a make or model they are interested in.

Dealership could also use this information to predict what range they could offer their customers when they are trading in a vehicle. While dealerships have greater resource to determine this in the current market sites like KBB, commonly known as Kelley Blue Book and North American Dealer Association or NADA price predictors are not able to determine expected vehicle prices accurately.

Challenges/Issues:

I am actually expecting a lot of challenges with this data. It appears a good job has been done to scrape a lot of good data from the site. It is also missing data fields like number of images, posting removal date and seller type. I don't believe this information is needed but would provide additional information for the analysis. If the data was available we could determine days on market, if image count was available it could be a feature in degerming sale timelines.

The other challenge is the data is limited to two months. More data would be useful and is available but comes with issues. As new versions where added fields have been dropped or added. This becomes a challenge when trying to combine different version to increase the amount of data. I still believe I can use the data to predict future days prices or weekly price changes. While car's prices in the past have not been as volatile as they are in the pandemic car market.

Reference:

1. Reese, A. (June 2021). “Used Cars Dataset” Version 10. From Kaggle.com. https://www.kaggle.com/austinreese/craigslist-carstrucks-data/version/10
2. “Craigslist”. (February 2022). From Wikipedia.org. https://en.wikipedia.org/wiki/Craigslist
3. Hodge, L (January 2022). “Welcome to Used Car Buying Hell”. From Jalopnik.com https://jalopnik.com/welcome-to-used-car-buying-hell-1848381994
4. “13-Month Rolling Used-Vehicle SAAR”. (January 2022). From Cox Automotive.com https://www.coxautoinc.com/market-insights/cox-automotive-13-month-rolling-used-vehicle-saar/

Appendix:

Vehicle Table

Column Name	Column Description
id	A unique number for the listing
url	A unique url for the listing
region	Region in which the listing was posted. Contains city name in most cases
region_url	Region the listing was posted within
price	Asking pricing of the vehicle
year	Year of the vehicle for sale
manufacturer	Manufacturer of the vehicle
model	Type of model the vehicle is
condition	What condition the vehicle is in
cylinders	The number of cylinders of the internal combustion engine
fuel	The type of fuel required
odometer	The milage that appears on the odometer of the vehicle
title_status	The current titling of the vehicle
transmission	The type of transmission in the vehicle
VIN	The unique number call the vehicle identification number that identifies the vehicle within the DMV systems and to the manufacturer
drive	The placement of the drive wheels
size	The class size of the vehicle
type	The type of vehicle
paint_color	The color of the vehicle
image_url	URL of the images provided by seller
description	The description of the vehicle that is for sale
county	The county in which the vehicle is located
state	The state in which the vehicle is located
lat	Short for latitude coordinate of the vehicle location
long	Short for longitude coordinate of the vehicle location
posting_date	The date the vehicle was posted on Criagslist