Dingyu Sun, Irene Lu, Peilin Zhong, Xiaoben Yin

SYRACUSE UNIVERSITY

Craigslist Used Car Price Analysis

IST 718 GROUP 14

Table of Contents

[1. Abstract 3](#_Toc57585083)

[2. Introduction 4](#_Toc57585084)

[3. Data description 4](#_Toc57585085)

[4. Data processing 5](#_Toc57585086)

[5. Visualization EDA 6](#_Toc57585087)

[6. Methodology 8](#_Toc57585088)

[7. Feature encoding 8](#_Toc57585089)

[8. Model Building 9](#_Toc57585090)

[7.1 Random forest 9](#_Toc57585091)

[7.2 Linear Regression Model 9](#_Toc57585092)

[9. Inference 12](#_Toc57585093)

[9.1 Feature Importance 12](#_Toc57585094)

[9.2 Japanese car vs German car 13](#_Toc57585095)

[9.3 Relationship between entry price and each factor (predictor)? 14](#_Toc57585096)

[9.4 What percentage in entry price is associated with type / condition? 14](#_Toc57585097)

[10. Prediction 14](#_Toc57585098)

[10.1 If the listing price is below market value / about the same / beyond market value; what’s the actual value? 15](#_Toc57585099)

[10.2 When is the best time to resell the car? 15](#_Toc57585100)

[11. Conclusion 16](#_Toc57585101)

[Reference 17](#_Toc57585102)

[Appendix - Dataset Description 18](#_Toc57585103)

## Abstract

Craigslist is the world's largest collection of used vehicles for sale. The team found someone who web scrapped used car data from Craigslist from September 2020.[1] Starting from the interests for used cars and the identity of being a customer, the team sets the project goal to explore the current used vehicle market as well as predicting the price of a used car.

While focusing on the needs of customers who do not know car market well, the team comes up with a list of predictions to help customers value their choice including (1) predicting if the listing is under/equal to/above market value; (2) the actual market value of a used car; and (3) how long should the customer keep the car until resell it for minimum value lost.

Also to serve the purpose of exploring and learning about the current used car market, we aim to make several inferences from our data analysis. The inference includes: (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) What percentage in entry price is associated with type / condition; (4) How long will a used car hedge against depreciation if it is a Japanese car compared to the same features of a Germany car?

After defining the goal and scope of our project, the team started with a background research, followed by data processing to get an overview of the dataset as well as pre-clean the data for later analysis. Then we explored our dataset with exploratory data analysis and visualizations to deeply understand the data. From there, we did feature encoding to transform our data into the shape that can be used for data modelling.

To predict the price of used vehicles and understand how each feature is related to the price, we built two models: Random Forest Regression and Linear Regression. And we interpreted the result from both models to answer the prediction and inference questions we came up in the very beginning. In the end, we summarized our findings to achieve our project goal.

In this project we did not successfully use a linear model to predict the price of the car because of the low R-Square value and its limited explanatory power. We conclude the reason behind this issue due to the problem complexity and the existence of outliers. The linear regression model cannot perform very well when the problem complexity and scale of outliers increase.

## Introduction

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.[2] 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.

## Data description

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. [3]

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>

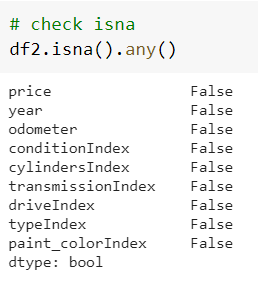
This dataset has 423,867 rows and 25 columns. Since we are predicting the price of used cars, we identify a list of attributes that contribute to the change of price (predictors). The list of attributes are region, year, manufacturer, model, condition, cylinders, fuel, odometer, title\_status, transmission,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

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 processing

Data processing is one of the most important parts of the data analysis project, it is a process that converts the data from raw data to well-organized data. In this process, data can be converted from different units to the same unit, from skewed to normal distribution.

In this part, we did the data processing part from 4 perspectives. The first one is to filter the abnormal data points, and then we dropped some unneeded columns, the third one is that we combined some types into one group, the last step is to find the NA value and use a model to fill it. Due to a lot of missing values in our dataset, we spent great time on this part.

Price and odometer are numerical variables, these two attributes are easy to consider their outlier. After researching online, we found that most of used cat websites set $1,000 as the minimum price when then selling the used car. To find the most appropriate maximum value we decided to use the percentile value of the dataset. We used the 99.7% price as the upper boundary. Although we did not find a good standard to evaluate the odometer, we still filtered the data in which the odometer is less than 0.

The dataset has both numeric and categorical attributes. With categorical variables, some attributes have a lot of unique values, whereas many values only occur a hundred times (aka rare) compared to our vast data size. To make the model more general and prevent overfitting, we combined some rare types into one group and named it “other”. From the summary table, we also found that some columns contain anomaly data. For instance, the “condition” column contains a long string text and it was obviously not a description of the car condition, so we dropped these rows.

After dropping all the rows with more than 13 NA columns, we ended up with only 30% of original dataset. So, we considered that if we drop more data, the rest of the dataset would not be a good dataset for model training, so we looked for a good way to fill the NA value for this dataset. KNN is a good model for our dataset, as it can also fill the NA value for both numerical variables and categorical variables. Following the model building steps, we converted the categorical variables to numerical variables at first, set a number for each type including the NA value type. And then we selected columns which contain NA value into the model, after this process, we converted the cluster number back to the original categorical data and finished the fill NA value step.

Until now all data processing was finished, there is no NA value, no abnormal data, and the final version of the dataset contained 92729 rows and 17 columns.

## Visualization EDA

Chart, box and whisker chart

Description automatically generatedBefore building any models, we did exploratory data analysis on our dataset to better understand our data. We did exploratory visuals for all attributes from our dataset, but we will only show the most important and interesting visuals in this section.

1. Price

The first thing to look at is the attribute “price” since this is our predicted variable. We did both a histogram and a boxplot to price and saw that price is extremely right skewed. From the boxplot you can see that more than 75% of the price lies under $20,000.

1. Year

We also want to know what the manufacturing year of all the listings on Craigslist. So we did the histogram of the attribute “year”. We see that year is left skewed with most of the listings are after year 2000 but the older listing can be dated from 1923. In order to view our visuals in a clear way, the below left graph only shows the partial histogram after 1997.

Chart, scatter chart

Description automatically generatedSo how is the year of manufacture related to the price of the listing? We did another scatter plot of Price vs Year to see how price is affected by year. And the interesting finding here is although after 1980, the newer the car, the higher the price it is, we also see that as the car gets really old, like an antique car, the price can be really high as well. So it is not a linear relationship between price and year.

Chart, bar chart, histogram

Description automatically generated

1. Type

Chart, box and whisker chart

Description automatically generatedChart, bar chart

Description automatically generatedWhen talking about cars, we are referring to different types of vehicles, like sedans, SUVs, etc. So we did the bar chart of the number of listings by car types. On the bottom left you can see that there are 13 categories of car types, where sendan, SUV and truck are the three most common car types on the market. So we did another comparison boxplot between Price and Car types to see how car types affect price. From bottom right we see that sedan, wagon, hatchback and mini-van have the smallest price range as they are smaller cars by common sense. Whereas pickup, bus, and truck have the largest price range because of their large size.

1. Chart, box and whisker chart

   Description automatically generatedYear and Condition

From our EDA with condition, we wonder how the condition is associated with the year of manufacture. The common sense is that the newer the car is, the better condition it is too. So, we did the comparison boxplot between Year and Condition. Yes, it quite follows people’s common sense as the better the condition, the newer the car if we look at the IQR of each condition. These two attributes do have a more linear relationship. But the graph also shows us that even the recent manufactured cars can be differed in conditions. Since all 6 condition categories have cars between 2015 and 2020. And good condition cars can be from old cars too if we look at the lower fence.

1. Chart, box and whisker chart

   Description automatically generatedCar price by country of manufacturer

When we started this project, all the team members have a same common sense that Japanese cars are considered “hedge” in the US market whereas the German cars tend to lose more value once it becomes a used car. We want to test if this is true in our EDA as well, so we built this comparison boxplots between Japanese cars and Germany cars to see how their price changes by year.

And the conclusion cannot be arrived at this point for this question. If we look at how the IQR changes, we see that German car price shrinks more compared to Japanese cars from 2018 to 2020 whereas Japanese brands seem more stable. But if we trace back a bit earlier before 2018, German cars seem to be the winner. So, it is hard to say which country of manufacturer is more hedged.

## Diagram Description automatically generatedMethodology

Our project follows the methodology in data science shown on the right, which consists 8 steps from finding a problem to collecting data, from understanding data to preparing data, from modeling to evaluation, and from deployment to conclusion. This provides a framework for proceeding with the methods and processes to get results. Once completed with individual work, the team collaborate to share observations and finish reports via Google drive, Google Colab and Zoom.

## Feature encoding

After data processing, there’s still one more step to do before building the models, which is feature encoding. As mentioned above, there are 17 columns in our final version of the dataset, out of which there are 5 numerical columns, which are price, year, odometer, latitude and longitude, and 12 categorical columns that can be divided into three types to perform different feature encoding.

The first type is the categorical variables with order, such as title status, condition and cylinder. For these three variables, we assigned numbers from 0 to represent their orders. In the title status column, there are 6 status: clean, salvage, rebuilt, parts only, lien and missing so we assigned 0 to 5 for each of them. Same thing happened in condition and cylinder columns. In the condition column, we assigned 0 to 5 for fair, salvage, good, excellent, like new and new. In the cylinder column, we first dropped text ‘cylinders’ in each row to only keep the numbers, and then assigned 0 to 6 standing for 3 cylinders, 4 cylinders, 5 cylinders, 6 cylinders, 8 cylinders, 10 cylinders and 12 cylinders.

The second type is categorical variables without order, such as fuel, transmission, car type, drive and paint color. We performed one hot encoding on these variables and finally converted them into vectors.

The last type of the categorical variable is the manufacturer for which we found their corresponding countries. In our dataset, all the cars come from UK, USA, Japan, Germany, South Korea, Italy and Sweden. After feature encoding, the dataset was ready for further analysis.

## Model Building

To predict the price of used vehicles and understand how each feature is related to the price, we built two models: Random Forest Regression and Linear Regression. In this section we will explain in detail about these two models and interpret our result.

### 7.1 Random forest

Basically random forest was used in the prediction part in our project. We performed random forest, GBT and grid search cross validation with all converted variables into consideration. For each regression model, we used R-squared, RMSE, MAE and explained variance to evaluate their performance.

For the random forest regressor model, the R-squared was around 71%, RMSE was around 5,628.81, MAE was around 3,705.68 and explained variance was around 58,960,375.60. The scores of GBT model and GBT model after doing grid search cross validation have been improved dramatically. The R-squared of GBT model was around 77%, RMSE was 5,008.32, MAE was 3259.61 and explained variance was 87284231.71. In the grid search for the GBT model, we found the best parameters were as follows: the best depth was 5, the best iter was 30 and the best bins was 40. After applying these parameters, we got an even better result. The R-squared was raised to 78.6%, the RMSE was reduced to 4,850.3, MAE was reduced to 3,144.66 and the explained variance was 88,714,271.25, meaning that in our best model with the smallest RMSE, the variables could explain 78.6% of the variation in price.

### 7.2 Linear Regression Model

Before building the linear regression model, there are some potential problems for fitting a Multi-Linear Regression Model needed to be clarified at the very beginning.

1. Non-linearity relationship of the target variable and predictors.

The linear regression model assumes that there should be a straight-line relationship between the target variable and the predictors. If the actual relationship is too far away from linear, then the validation of this model and any inference / prediction drawn from that are suspect, the explanation power and accuracy would be significantly reduced.

1. Outliers

According to the exploratory data analysis conducted in the previous section, the distribution of price (target variable) and predictors like odometer, are extremely right-skewed. Those extremely large outliers could be the results for incorrect entry during a used car seller filling out the information of cars for sale; or some used car sellers tend to exaggerate the selling price for cars which do not deserve that worth.

1. Collinearity

Collinearity refers to the situation where two or more predictor variables are closely related to each other. In this case, the presence of collinearity would cause problems on the model performance and explanation. It would make it difficult to distinguish the effort of each collinear predictor variable placed on the target variable, because they are closely associated with each other.

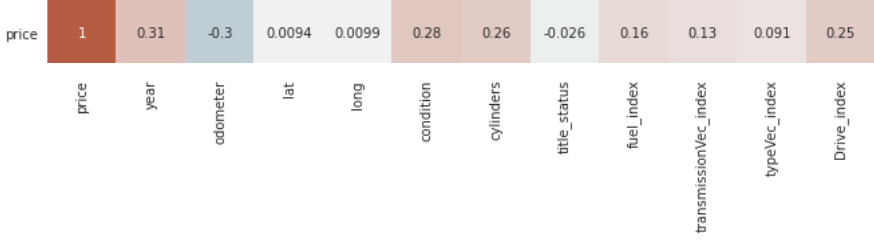
#### 7.2.1 Explore Relationship

In order to examine and avoid those potential problems. A preliminary exploration on the relationship between variables - target variable and predictors - is needed. Before the feature encoding, the pair plot below displays the pairwise relationships between numerical variables in the dataset.

According to the pairplot, The distributions of price and odometer are extremely right-skewed. Thus, transformation on target variable (price) is needed to prevent the nonlinearity. By implementing log transformation on ‘price’, we managed to get the distribution of ‘price’ closer to a normal distribution, which may significantly reduce the effects made by outliers and generate a more linear relationship. Besides, standardizing predictor variables has similar functionality, which may reduce the negative effect brought by outliers and improve the explanatory power of the linear regression model.

The scatter plots of price versus other variables show their relationship in between. In general, the latitude and longitude is closely related to each other, thus including both of them into the linear regression model would cause collinearity. Since these two variables will not be taken into our consideration, there is not further explanation on that.

In order to take a closer look at the relationship among each variable, we generate a heatmap to display the relationship in between. Based on the color pallete, more brick red shows a stronger positive relationship in between; more blue shows a stronger negative relationship in between. Overall, the absolute values of all correlation coefficients are not very significant - not greater than 0.31, a rather weak relationship.

Year, condition, cylinders are positively related to price, thus newer the listing year, better the condition, more cylinders, higher the average entry price of used cars. Odometer is negatively related to price, thus more odometer a used car ran, lower the average entry price of that used car, which matches the common sense.

#### 7.2.2 Multi-Linear Regression Model

Here are the general steps to build the linear regression model:

1. Apply log10 transformation on target variable (price), standardize numerical predictor variables.
2. Include 10 predictor variables into the model: year, odometer, condiction, cylinders, title status, fuel, transmission, type, drive, paint color.
3. Use R-Square, Root Mean Square Error (RMSE), Mean Absolute Error (MAE) to evaluate the model performance.
4. Interpret the coefficients and make inference.

Based on that, we implement a forward stepwise approach to find out the optimal order for predictors adding into the linear regression model, in order to improve the model performance.

What's more, we implement the order of feature importance (from most to least) for predictors adding into the linear regression model. Then compare the model performance with the one with Step-Wise Optimal Order to find out the optimal model.

#### 7.2.3 Forward Stepwise Selection

Forward Stepwise Selection is to find out the best predictors subset. Starts with a model containing no predictors and then adds predictor variables into the linear regression model at a time, until all predictors are included in the model. At each step, the predictor that contributes the greatest improvement to the fit will be added into the model.

In the project, based on the resulting correlation coefficients from the previous section, we begins the stepwise approach with ‘year’ as the first predictor included in the model, since ‘year’ is the most correlated predictor variable with ‘price’, which give out the highest R-Square among any other predictors. Then, we add one additional predictor, consider all models that augment the model performance then choose the best model with the best improvement. In this case, we apply Mean Square Error (MSE) to evaluate the performance of each model generated during the loop process. Following this logic, we define several functions to implement the stepwise approach.

The resulting optimal stepwise order of all predictors are: year, drive, condition, type, odometer, fuel, paint color, transmission, cylinders, title status.

#### 7.2.4 Implement GBT Most Important Feature Order

According to the feature importance outcome from the GBT model above, we have another subset of predictor order: year, odometer, cylinders, type, condition, drive, fuel, title status, print color, transmission. We implement this subset as the input column for fitting the linear regression model, then compared with the previous ones.

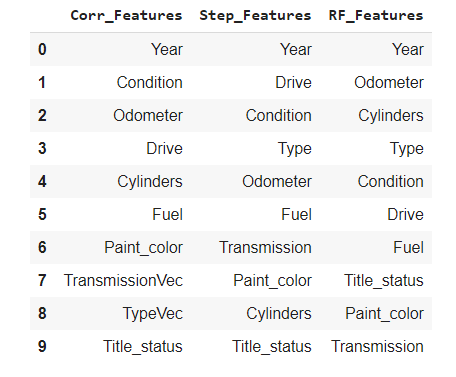
## Inference

### 9.1 Feature Importance

The dataset contains over 10 columns, which means the price value was evaluated together by many factors, such as year, odometer, paint color, etc.

We were also curious about which column is the most important column and which column is the least important column. After getting the result of feature importance we can give some business insights, telling sellers if they want to get a higher selling price they should give more consideration to such factors.

We were planning to evaluate the feature importance from three models, correlation, linear model and random forest.



The first model is the correlation computation, like we did before, we computed the correlation between each column. Correlation is a common way to evaluate how close the two variables related to each other. If two variables have strong linear relations, they have a higher correlation score than the other two columns. In this process, we computed the correlation between “price” and other variables, after absolute the correlation value, we got the feature importance result. The first column of the screenshot represents the feature importance order from correlation.

The second model is the linear model, we got the inspiration from the HW3, the stepwise function. We wanted to find the feature order from our best model, so we created a stepwise function, by evaluating the MSE we add the feature to the result list which has a lower MSE increase compared with other columns.

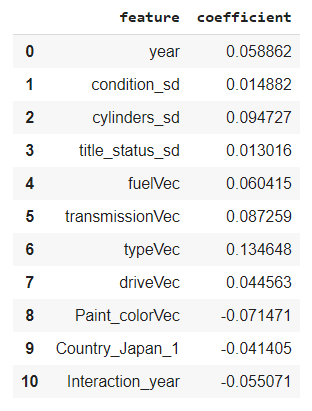
The third model is the GBT model, by calculating how much performance improvement of each attribute split point, we got the result of the GBT model.The third column of the dataframe is the feature importance result from the GBT model.

At first, we thought these three models might create the same result, however, the results were different. We tried to search for the reason, and we thought there are several factors that may affect the result. The first one is they used different ways to compute the feature importance, the correlation model computed the correlation, the linear model computed the MSE value, and the GBT model computed the purity. Thus, different methods lead to different results. The next consideration is which model provides more convenience. The second factor is that different models have different model performance, if a model has a lower performance than others, this model has less credibility than others. From the previous research, we can compute the R-square value and RMSE value, the GBT model has a higher R-square score and lower RMSE score than the linear model.

We can conclude that the “year” feature is the most important feature among all features, it ranked in the first place in three models. The least important features might include “title\_status”, “paint\_color” and “transmission”.

### 9.2 Japanese car vs German car

We want to find a relationship between the car manufacturer and the price, is there any possibility that a car from a specific manufacturer has a higher selling price than the other car manufacturer when we keep other elements as the same.

We supposed that cars from the Japan manufacturer may have a higher selling price compared with cars from the Germany manufacturer, and this suppose we got from our daily life experience. We created a linear model to analyze the coefficient of the “country”, we wanted to set other variables as equal and only consider the effect of the “country”. So we filtered the “Germany” and “Japan” from the country variable and created an interaction variable with the “year” column.

The screen shot in the right side shows the coefficient result of the linear model, from the result we can see that the coefficient of the interaction variable is very small, thus it’s hard to conclude that with the year increase there would be a huge price increase of the car from the Japanese manufacturer.

### 9.3 Relationship between entry price and each factor (predictor)?

We consider this inference from two perspectives, the EDA result and the feature coefficient. The coefficient score of the “year” is positive, which means with the year of car production the selling price will increase, and the EDA plot also shows an increasing tendency. The coefficient of the “odometer” is negative, which means with the odometer of the car increasing, the selling price will decrease.

### 9.4 What percentage in entry price is associated with type / condition?

The screen shot in the right side is the coefficient result of the stepwise linear model. When the condition increases by 1 standardization, the selling price is predicted to decrease 0.006%. When the type increases by 1 standardization, the selling price is predicted to increase 0.099%.

## Prediction

Based on our Random Forest Model, we decided to make a few predictions.

### 10.1 If the listing price is below market value / about the same / beyond market value; what’s the actual value?

The goal of this prediction is to compare our model predicted price and the listing price the customer finds. And we should have a conclusion to whether the listing has a fair price or not. For this prediction, we searched online for new listings with price and stored the information in a separate csv file. We would run the new data frame in our model and make predictions accordingly. Below you can find the result data frame screenshot after we predict the market value for the customer.

Table

Description automatically generated

These listings information are found from another used car platform Caravna [6], known for no dealer fee and high standard car pre-check. Unlike Craigslist, Carvana’s price is non-negotiable so we would expect that the price on Craigslist can be a little higher since it is negotiable.

And yes from our prediction price, 3 out of 5 prices on Carvana is lower than our prediction market value. And these results make sense according to the above explanation. Maybe Carvana is a good platform with a fair price.

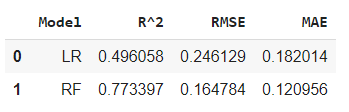
### Chart, line chart Description automatically generated10.2 When is the best time to resell the car?

Imagine a scenario: a customer wants to purchase a used car, and he/she wants to know what the value of this car will be in the following years so that he/she can decide when to resell it for minimum money lost.

Starting from this, we built a function to help the customer visual the price of his/her specific listing for the next 10 years. The function will take in all the listing information as well as the annual mileage that the owner plans to drive. Then the function will predict the price of the cars for the next 10 years and build a visual from it.

From above right you can see a sample result from our function. The price pattern is quite interesting here as the value suddenly drops between year 5 and 6. And at year 6, the value drops below half of the original price. We would recommend our customer to consider reselling the car after 5 years.

## Conclusion

To summarize our project, we have answered all prediction and inference questions with EDA and two models we built. In our model comparison, we used R-square, RMSE and MAE to evaluate the performances. From the graph on the right we can conclude that our random forest model performed better than the linear regression model. With our best model whose RMSE is 0.16, the independent variables can explain about 77% of the variance in price.

As for the inference, we use the regression coefficients to describe the effect of each predictor variable placed on price. By building the linear model we managed to find the optimal feature order and determined factors that greatly affect the price of the car. From the feature importance result and RMSE result, we chose the random forest model as the best model. According to the resulting feature importance of the stepwise linear model, the top 5 most features are “year”, “Odometer”, “Cylinders”, “Type”, “Condition”. Thus, we would consider these 5 predictors as the most important features, and If a customer wants to buy a used car, these 5 factors should be considered first.

As for the prediction, by building the random forest model and using the model to predict a new listing price, we managed to predict the exact value for a certain listing and make comparisons to original price to conclude whether the listing price is fair or not. Beyond that, we also built a function to generate a visual for next-ten-year price to help the customer make decisions on when to sell the car.

We did fulfill our project goal and answer all the prediction and inference questions. But there’s still some limitations in our project. Regarding the prediction result, our audience is the customer. So we should enable them the ability to input data and get the predicted value directly. It would be ideal if we could build a front-end user interface to have our customer interact with our predictions. Another potential direction is to build a recommendation system for cars as our target audience is customers.

At the final of our report, we look back on our project. We made some accomplishments and we also met some problems. When we meet the dilemma we have to overthrow our goal and find an alternative way. In this process, we had a deep and comprehensive understanding of data analysis, we also got inspired when we had a passionate brainstorm, these experiences will help us in our further career and study.

## Reference

[1] Used Cars Dataset

<https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

[2] 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/).

[3]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/) .

[3] Wikipedia

[4] Used Cars Dataset. (n.d.). Kaggle. Retrieved November 17, 2020, from <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

[5] Can feature importance change a lot between models? (n.d.). StackExchange. Retrieved November 17, 2020, from <https://datascience.stackexchange.com/questions/28818/can-feature-importance-change-a-lot-between-models/28831>

[6]carvana. (n.d.). Carvana. Retrieved November 29, 2020, from <https://www.carvana.com/cars>

## Appendix - Dataset Description

|  |  |  |  |
| --- | --- | --- | --- |
| **Attribute Name** | **Data Type** | **Description** | **Details** |
| region | String | Craigslist region | 404 unique values |
| price | Integer | Price of the listing |  |
| year | Integer | Year of manufacture | Range (1932,2021) |
| manufacturer | String | Manufacturer of vehicle | 39 unique values |
| model | String | Model of vehicle |  |
| condition | String | Condition of vehicle | Fair, salvage, good, excellent, like new, new |
| cylinders | String | Cylinder of vehicle | 3,4,5,6,8,10,12 cylinders |
| fuel | String | Fuel of vehicle | Gas, diesel, hybrid, other, electric |
| odometer | Integer | Odometer of vehicle |  |
| title\_status | String | Title status of vehicle | missing, lien, parts only, rebuilt, salvage, clean |
| transmission | String | Transmission of vehicle | Automatic, manual |
| drive | String | Drive type of vehicle | 4wd, rwd, fwd |
| size | String | Vehicle size |  |
| type | String | Vehicle type | Sedan, SUV, trcu, pickup, ... |
| paint\_color | String | Paint color of vehicle | White, black, blue, silver, ... |
| lat | Float | Latitude of car selling place |  |
| long | Float | Longitude of car selling place |  |