**Project Topic**

Used Cars Analysis

**Project Team**

Team 11: Hanzhang Wang, Xi Zhao

**Introduction**

*Background*

* In the U.S., the Covid-19 pandemic has led to an increase in used car sales as people avoid mass public transportation
* Currently, in the spite of economic recession, people are more sensitive to vehicles cost
* Consumers can now find more inventory of used vehicles online compared with before, and dealers are accelerating efforts to sip the piece of this cake
* While large used-car retailers such as AutoNation and CarMax, as well as recent IPO Carvana, have seen stocks rebound strongly since the March bottom in 2020, some online car sales rival such as Shift Technologies, are now eager to planning the new service in the used car market as well
* This demand surge and hot dealer supply can drive up the used cars’ prices

*Motivations*

* During the Covid-19 shutdowns, normal autos dealers are all facing the big challenge that all the open new car business are gone, and the used car market is a good start to get recovered
* Due to the shutdowns, factories, car supply chains are all closed and new car supply cannot be delivered on time, so, a better understanding of the used cars market is crucial for the whole autos industry
* A mass amount of people is encountering unemployment and the auto cost can be a burden to this kind of family since they have less income and few opportunities to drive outside but need to pay the same car loan
* More players are entering the growing used car market and the price would be driven up by the huge demand and dealer competition to be unreasonable or unconvincing, that will disturb the market and hurt both customers and dealers in the end

*Goals*

* To have a full picture of the used car market including car brand share, top-ranked models, colors, and regions, etc.
* To find out the correlation of car features and the price and conclude the insights
* To identify the strong price predictor to be used in the price prediction model
* To develop a healthy and powerful used car price prediction model to handle the market accordingly

**Methodology**

*Descriptive* *Analysis*:

* Clean the dataset by dropping unnecessary data and redundant data
* Transform the data to align to a certain format requirement
* Visualize the dataset by leveraging NumPy, Pandas, Matplotlib, and Seaborn libraries

*Advanced Analysis*

* Correlation analysis: mainly focus on identifying features that are highly related to the price:
  1. Select potential features which potentially impact used car’s price
  2. Rescale data by utilizing preprocessing, metrics, tree, etc. functions in Scikit-learn library as well as MinMaxScaler and LabelEncode from Preprocessor library
  3. Applying linear regression to get the correlation result
  4. Evaluate regression and classification results based on the Pearson’s regression
  5. Identify strong predictor attributes with high effect in regression and classification algorithm
* Prediction by Machine Learning: primarily work on general car price prediction and brand price prediction:
  1. Clean the data by removing unrelated columns recognized from correlation analysis and handle missing values for accurate prediction
  2. Model the data with properly split train and test datasets and fit in prediction models with predictive modeling algorithms such as linear regression, decision tree regression, random forest, etc.
  3. Gain the trend of all kinds of prediction models by calling internal predict function
  4. Evaluate all prediction results generated from all applied predictive models on different scales such as accuracy score, mean squared error(MSE) or root mean squared error(RMSE), etc.
  5. Conclude the best predictive model and summarize the results

**Dataset Description**

* This dataset contains every used vehicle entry within the United States between October 2020 to December 2020, as well as most all relevant information on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories
* All the tabulated information of used vehicles for sale are from Craigslist, which is one of the American largest information exchange platform and classified advertisements website with sections devoted to jobs, housing, for sale, items wanted, services, community service, gigs, résumés, and discussion forums
* The data was scraped every few months by web technology tools
* The owner of the dataset built a scraper for a school project and expanded upon it later to create this dataset
* Dataset link: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>

**Results and Analysis**

*Loading libraries and dataset*

In this project, pandas, NumPy, Matplotlib, and Seaborn were mainly used during the whole project analysis.

Graphical user interface, application

Description automatically generated

After downloading the dataset, we used Pandas to read the CSV file and call the info() function to get the general idea of this dataset. This used car dataset has 458,213 entries and 26 columns in total.

Table

Description automatically generated

*Descriptive Analysis*

After general understanding of the used car dataset, we realized that there were some unnecessary columns that we did not want to include in the descriptive analysis part. Then we selected some useful columns to better describe the used car markets. In this part, we opted region, price, year, manufacturer, model, condition, cylinders, fuel, odometer, status, transmission, drive, size, type, color, state, latitude, longitude, posting date columns, which are 19 columns in total.

Apparently, there is no perfect dataset. Every dataset is incomplete and usually having some null values. So, we just dropped all the null values. Moreover, some data types were not suitable for analysis. Year and posting date were all object data type, so, we converted year and posting date column to the date data type.

Text

Description automatically generated

What’s more, the age of a used car is very important. In order to better analyze this dataset, we added one more column “age” by calculation to better study the used car market. And we only focused on the cars manufactured after 2000. Thus, after selecting specific columns, dropping invalid values, and adding new calculation column, finally we got a new dataset with the shape of (80747, 20).

Text

Description automatically generated

Since it is a used car dataset which reflects used car market, we want to know how old these used cars are. So, we utilized the kernel density estimate (KDE) plot to visualize the age feature of the used cars. KDE is a method for visualizing the distribution of observations in a dataset, analogous to a histogram. KDE represents the data using a continuous probability density curve in one or more dimensions. In the density plot of the age of the used cars figure, we can conclude that 5 to 10 years old cars have the highest possibility to be posted for sale. It means people tend to sell their cars after using 5 to 10 years.

Chart, line chart

Description automatically generated

Further, we used Seaborn library to visualize the dataset to get some general insights of used car market. There are plenty of different attributes of used cars in this dataset. For better visualization, we only selected top ranking items to compare. From the count plot figure, we can see that the top 5 manufacturers are Ford, Chevrolet, Toyota, Honda, and Nissan. Among different car types, the most popular car types are sedan, SUV, and truck. Similarly, among all kinds of car models, f-150, Silverado 1500, 1500 are most common car models. For the car color, the most posted used car colors are white, black, silver, grey, blue, and red.

Chart, histogram

Description automatically generated

Then we used cross analysis to get knowledge of specific popular models from some certain manufacturers. The following heatmap can tell us exactly what we want. Darker the blue color is more popular the certain car type is. After a glance of this heatmap, we can clearly see that the Ford truck, Chevrolet truck, Ford SUV, Chevrolet SUV, Jeep SUV, and Toyota sedan are top posted used cars for sale.

Graphical user interface, application

Description automatically generated with medium confidence

Here is another heatmap about manufacturers with most used cars for sale each year. Similarly, more shadow red represents more used cars for sale in the specific year. From this plot, we can clearly see that Ford has the most posted used cars for sale between 2012 and 2018. And Chevrolet also has a lot of posts from 2011 to 2014.

A screen shot of a computer screen

Description automatically generated with medium confidence

This line plot is more clearly on the trend of 5 top manufacturers over the time. From 2000 to 2020, Ford and Chevrolet both dominated the used car market.

Chart, line chart

Description automatically generated

Price should be the most concerned aspect of a for-sale used car. The following three bar plots illustrate the mean price of different manufacturers, car types, and car colors. We can conclude that Chevrolet used car has the highest price. Pickup is the most expensive used car for sale on average. Red cars are way more expensive price than other colors.

A picture containing graphical user interface

Description automatically generated

Finally, we also analyzed region distribution of all used cars by manufacturers. From this map, we can see that for the top 5 manufacturers, used cars for sale posts were posted more from middle east and east coast. We believe this is also related to the population distribution.

Chart, scatter chart

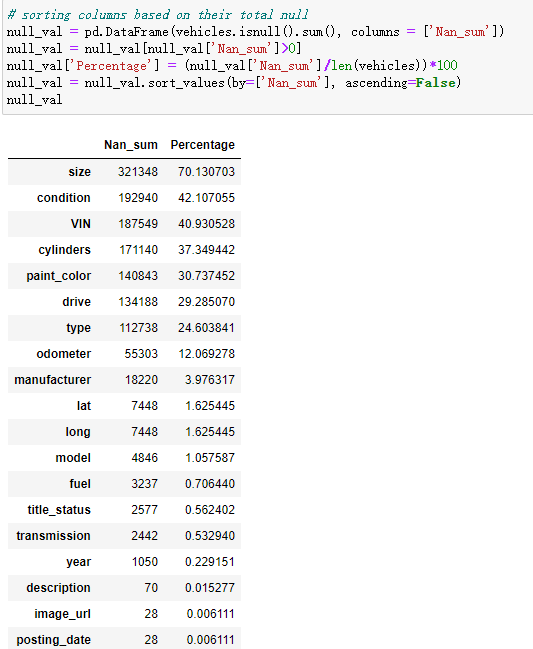
Description automatically generated

*Correlation Analysis*

Many columns are useless in correlation analysis. Without a doubt, we need to clean the data. First, we dropped some irrelevant columns. And then we want to check how many null values in each column.

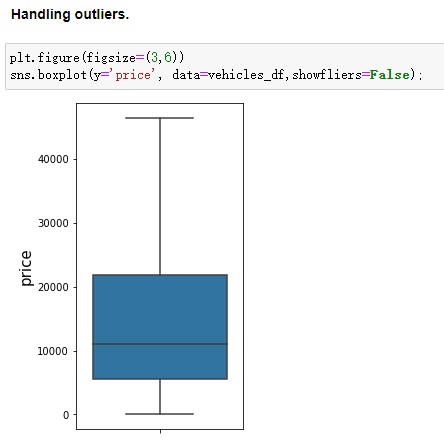


It turns out that many columns have a great number of null values. Then we sort columns based on their total null.



We can treat those columns whose null values differently depends on the project's goal. For the sake of this project, we are only going to keep columns which have less than 40% of missing values. Next, we cleaned remain rows with missing values. Until then, there was only 178410 rows × 15 columns left which is way smaller than the original data size.

Then, we thought about outliers. Outliers can affect the accuracy of correlation analysis and further prediction. Take feature price as an example.

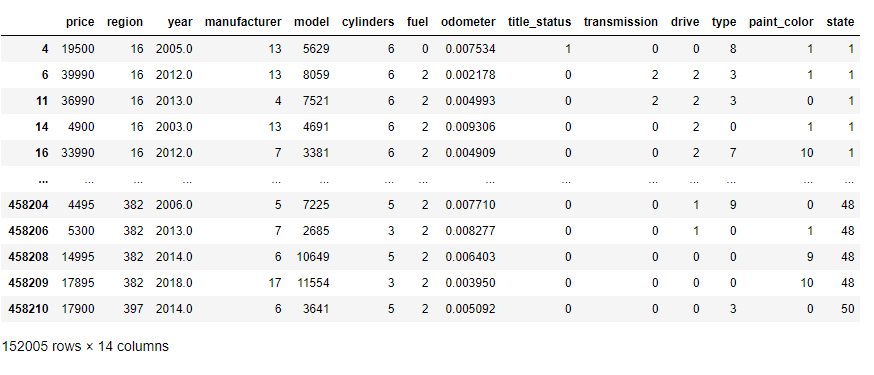


As we can see, most of the values of price are fine. So, we remove rows with price as 0 and take range of value from 5%-95%. 16325 rows have been removed.

Odometer is a feature with larger magnitude. We need to reduce the scale of it to prevent from dominating the prediction model. We applied MinMaxScaler so prediction model will perform better.



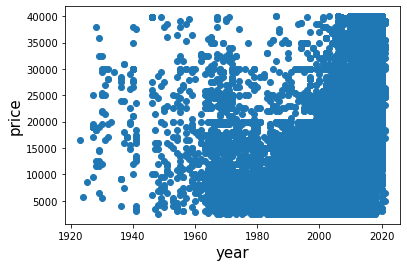
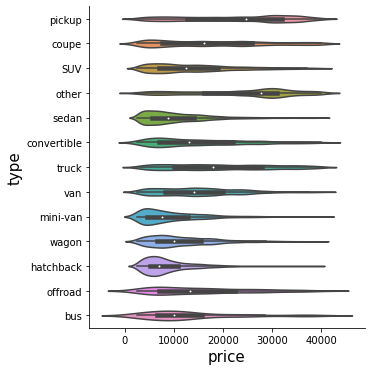
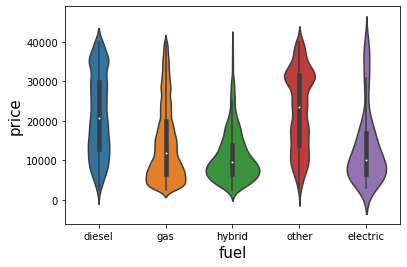
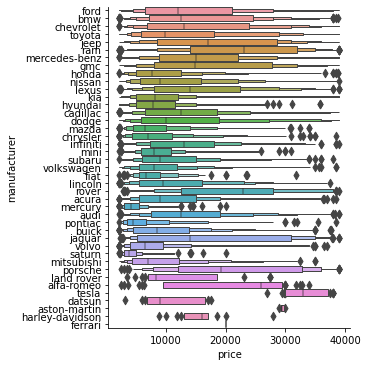
Below is our cleaned data.



We use two methods for features selection, and we will select the one with highest accuracy when we apply both methods in multiple linear regression.

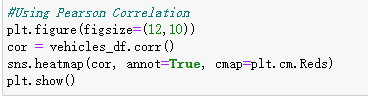
Before we introduce our method, let's check distribution of some predictors with respect to target(price) for general understanding.

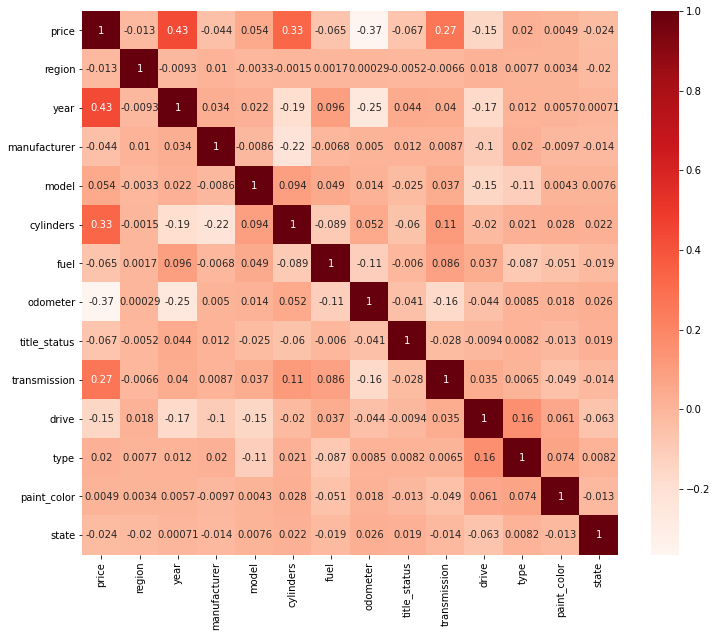
Most manufactures have its unique boxplot. Therefore, manufacture probably plays an important role in influencing price. As for fuel, there are only a few types of fuel and they tend to split into two groups. One with gas, hybrid and electric. The other with diesel and other. Fuel might have some influence on price. In the image of type VS price, most features of types perform same shape. So, type is probably irrelated to price. Year is definitely a big factor affecting price as we can see from the image. The price section from 20000 to 40000 has denser distribution in 2000 to 2020 then 1980 to 2000.



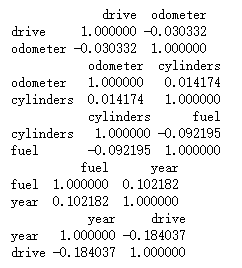
After general understanding of these features’ roles in relation to price, then we decided to use Multiple Linear Regression to get the high relation features with price. There are two prediction methods in this model, which are Filter Method and Backward Elimination Method. In order to get the best result, we used both methods.

The first method we used is Filter Method. We filtered and took subset of relevant features. Then we used correlation matrix with Pearson Correlation.



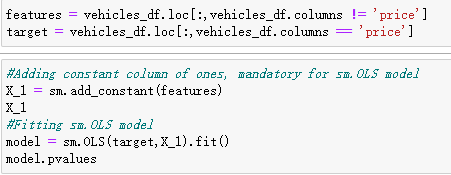


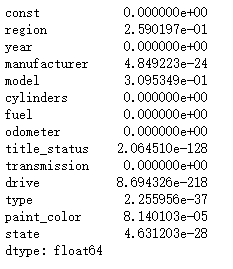
As we can see from the heatmap that year, cylinders, transmission, drive, and odometer have relatively high positive and negative relationship with price. We will use these features for the following procedure. One of the assumptions of linear regression is that the independent variables need to be uncorrelated with each other. We also need to check if these variables are related with each other. From the following calculation result, we can see that these features do not have high relation with each other. Therefore, we can use these key features to predict the used car price.



The second method we used is Backward Elimination. This is an iterative and long procedure, but it is more accurate for most of the time. We feed all the possible features to the model at first. We check the performance of the model and then iteratively remove the worst performing features one by one till the overall performance of the model comes in acceptable range (0.05).

First, we set price as target and every column other than price as features. Fit them into OLS model. Then we got p values of all features.





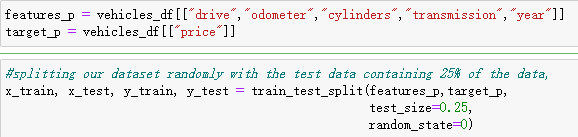
After that, we kept eliminating features and fit model until maximum p values is below 0.05. And top 5 features which have high relationship with price are year, manufacturer, cylinders, fuel, and odometer.

*Price Prediction by Machine Learning*

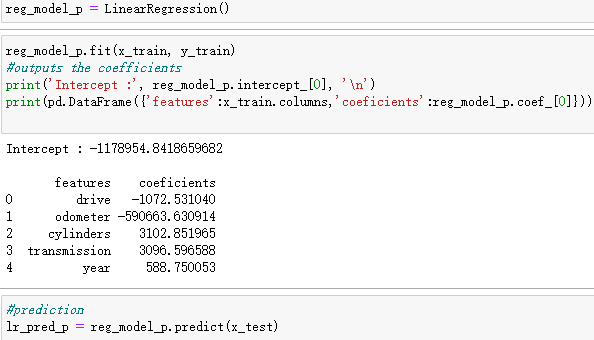
* Multiple Linear Regression Model

After we figured out 5 features which are highly related to used car price, next we continued using the two methods, Filter Method and Backward Elimination Method in Multiple Linear Regression Model to predict used car price.

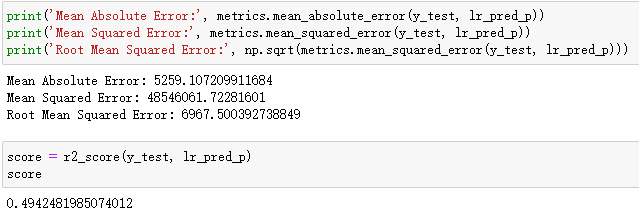
First, we used Filter Method of MLR model. We set target as price and features as year, cylinders, transmission, drive, and odometer. We split our dataset randomly with the test data containing 25% of the data.



Then, we fit the train data to the method to train this model and then called the internal predict method to predict the price of the test data.



After the prediction with Filter Method, we used multiple linear regression to find out this method’s accuracy. We run the regression model with Pearson Correlation method and here is the result.



Secondly, we did the same steps for Backward Elimination Method. We set target as price and features as year, manufacturer, cylinders, fuel, and odometer. We again split our dataset randomly with the test data containing 25% of the data. Then, we called the internal predict method to predict the price of the test data. Lastly, we also used multiple linear regression to find out this method’s accuracy. And here’s the result.

Graphical user interface, text, email

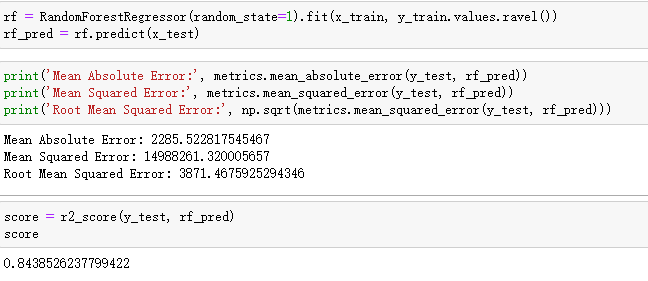
Description automatically generated

Since Backward Elimination method has slightly higher R2 score, we used results from Backward Elimination method to stand for the MLR model results. That is the top 5 features which have high relationship with price are year, manufacturer, cylinders, fuel, and odometer.

* Random Forest Model

The second model that we used for prediction is Random Forest Model. As we can see from the above model, it takes a lot of time to find the correlated features of price. So, we wanted to use a simpler and more intuitive model to predict the price. Random Forest Classifier is the best choice. It consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction and the class with the most votes become the model’s prediction. It suits the used car dataset with multiple features very well.

First, since we already obtain top 5 high relation features with price from MLR, we can skip splitting data step and kept using the previous split data. Then, we fit data into random forest regressor model and make prediction. Finally, we also calculated several scores to evaluate this model. Here is the result.



We can see that the R2 score of Random Forest is 0.84, while MLR model’s is 0.51. Therefore, Random Forest Model is the optimal model for the used car dataset.

**Conclusion**

* 5 to 10 years old cars have the highest possibility to be posted for sale, which means people tend to sell their cars after using 5 to 10 years.
* The top 5 manufacturers of for-sale used cars are Ford, Chevrolet, Toyota, Honda, and Nissan.
* Among different car types, the most popular car types are sedan, SUV, and truck.
* f-150, Silverado 1500, 1500 are most common car models in used car market.
* The largest count of for-sale used car colors are white, black, silver, grey, blue, and red.
* Ford truck, Chevrolet truck, Ford SUV, Chevrolet SUV, Jeep SUV, and Toyota sedan are top posted used cars for sale.
* Ford has the most posted used cars for sale between 2012 and 2018. And Chevrolet also has a lot of posts from 2011 to 2014.
* From 2000 to 2020, Ford and Chevrolet both dominated the used car market.
* Chevrolet used car has the highest price on average.
* Pickup is the most expensive used car for sale on average.
* Red cars are way more expensive than other colors on average.
* For the top 5 manufacturers, Ford, Chevrolet, Toyota, Honda, and Nissan, for-sale used car posts were posted more from middle east and east coast. No clear differences on the region distribution of different manufacturers.
* For this used car dataset, in the Multiple Linear Regression model, Backward Elimination method is more accurate than Filter method. And it is concluded that top 5 features from MLR model, which have high correlation with price, are year, manufacturer, cylinders, fuel, and odometer.
* The R2 score of MLR model is 0.51, while Random Forest’s is 0.84. Therefore, Random Forest Model is the optimal model for the used car dataset.

**References**

1. Rosenbaum, Eric. (2020,October 15). The used car boom is one of the hottest, and trickiest, coronavirus markets for consumers. <https://www.cnbc.com/2020/10/15/used-car-boom-is-one-of-hottest-coronavirus-markets-for-consumers.html>
2. JeongBin\_Park. (2021, March). Craigslist\_Car Data Analysis. <https://www.kaggle.com/jeongbinpark/craigslist-car-data-analysis/data#1.Data-setting>
3. Deep Shah. (2020, November). Used Cars EDA. <https://www.kaggle.com/deepshah16/used-cars-eda>
4. Nurlaida. (2021, April). Exploratory Data Analysis. <https://www.kaggle.com/ideayahya/exploratory-data-analysis-framework-created>
5. Vitalii Mokin (2020, April). Used Cars Price Prediction by 15 models.

<https://www.kaggle.com/vbmokin/used-cars-price-prediction-by-15-models>

1. Aneri Savani (2020, October). EDA and price prediction of used vehicles

<https://www.kaggle.com/anerisavani/eda-and-price-prediction-of-used-vehicles>