Springboard Data Science Capstone Project 1

Predicting the Used Car Price

From Craigslist

Yang Fei
Mentor: Kenneth Gil-Pasquel
Data Science Capstone Project 1, June 2020

1. Introduction

More and more people who plan to sell and buy their used cars tend to choose online platforms since it helps consumers save time and cost efficiently. Craigslist, as the world's largest used car platform, has a great number of trading volumes every year which also provides a large amount of selling data.

Online used-car prices are affected by several factors, including the year, mileage, brand, model, etc. We try to analyze these influential data, build a machine learning model, and make price performance predictions on the most popular Top 5 used cars manufacturers on craigslist.

WHO MIGHT CARE

Buyers who may not have enough knowledge about the used-car market and tend to make their decisions by following the public's ideas. They prefer to get a clear and concise price by observing an average standard.

Used car franchisees who could know about market trends by observing used online used car price performance and consumer psychology. They could make up their marketing plans that may be helpful to increase benefits and decrease risks.

Vehicle manufacturers who plan to conduct market research from the used car market. The used car price performance usually reflects the hedging rate directly. Used car price performance may affect manufacturers' future development and sales strategies.

DATA:

This data is from Kaggle https://www.kaggle.com/austinreese/craigslist-carstrucks-data

It contains most all relevant information that Craigslist provides on car sales including columns like price, condition, manufacturer, latitude/longitude, and 18 other categories. Since I am a beginner in data science, I plan to choose a topic which I could follow other's steps when I meet troubles.

APPROACH:

My first step is to clean and analyze the existing data, in order to conclude a price performance standard for each sample. After that I plan to build the model to predict price performance of certain models. The results should be clarified into three classes: over standard, standard, below standard which could be considered as a Multi Class classification after changing the price variable.

DELIVERABLES:

Codes

b. data cleaning:

a. data acquisition

c. data exploration analysis

d. machine learning model development

Report on the capstone project

Presentation on the capstone project

All materials will be uploaded to my Github repository.

2. Data Cleaning and Wrangling

2.1 Import and check data

I got a general idea of the data to know the size of the data set, the numbers of columns and so forth. This file contains 25 columns; I only kept the useful ones. The purpose of this step is to figure out the columns that are useful for further analysis, as well as to make the data frame easy to run and read.

2.2 Deal with missing and duplicated data

· Firstly I need to figure out how many null values in each column. For those columns that do not have any values at all, I could delete them directly.

· Second step I clarified the rest columns and fill the missing values with any value we choose, and this is the better technique to remove missing values from the data set.

· For the columns which are numeric I replaced the Null by their mean number. For 'manufacturer', 'model', 'paint_color' I replaced the null by 'unknown'. For 'year' I replaced the null by mode number.

· Each row has a specific id, so I check the duplicated ones according to their id.

2.3 Drop the fake data

Based on common sense, I found that some of the data are not real. For example, maximum price in the data set is 3600028900, which is obviously a fake data. By observing the scatter chart, I set a gross range for the data set, for example, the 'price' should be from 0 to 20000, and the 'odometer' should be 0 to 40000.

2.4 Transforming Data Using Function

In order to make the kilometer data more intuitive, I divided the 'odometer' into three classes, namely 'low odometer', 'medium odometer' and 'high odometer'. The purpose of this step is to prepare for the comparison box chart which shows the relationship of price and odometer.

2.5. Detecting & Filtering Outliers

A large number of outliers can be found in the box chart from the last step. I used the **IQR rule** and the **boxplot** to detect and extract the outliers from the data set.

3. Data Exploration

3.1 Introduction to the clean data

The clean data has 485862 records of used vehicles on Craigslist from 1900 to 2020, with 485862 rows. According to the research goal, I added one additional column, the age of the used car. I took 20 years as a research period, and chose the five manufacturers with the highest market share as research objects.

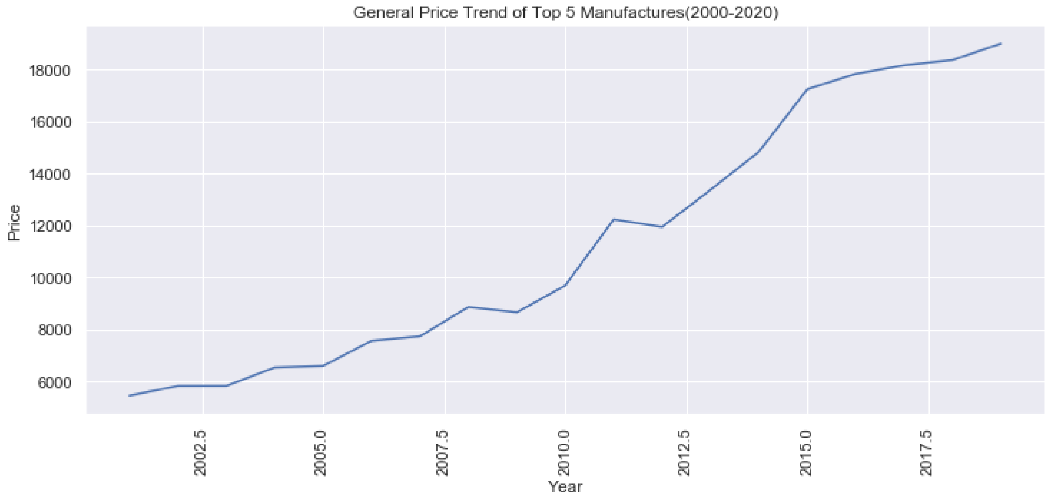
Target Variable = Independent Variable = Response Variable: Price

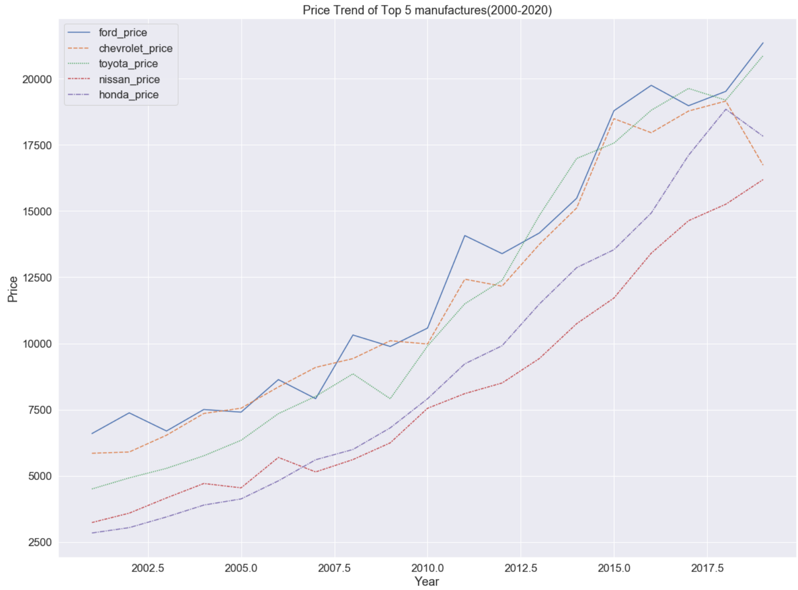
Features = dependent Variable = Explanatory Variable: odometer, year, age, paint color, condition, models.

I will classify all variable types and create analysis strategies which may explore their correlation with price.

3.2 year(age)

During the last 20 years, the average price of all used vehicles went up year by year. It may be related to the economic situation.





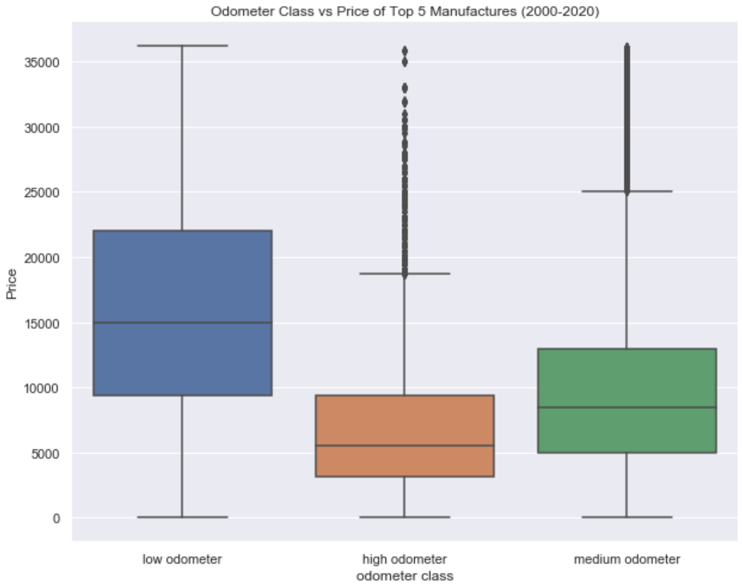
Ford, with the highest average price, was surpassed by Chevrolet around 2006 and by Toyota between 2012 and 2015.

The Nissan average price was overtaken by Honda in 2006.

Chevrolet suffered a rapid decline in 2017, as well as Honda .

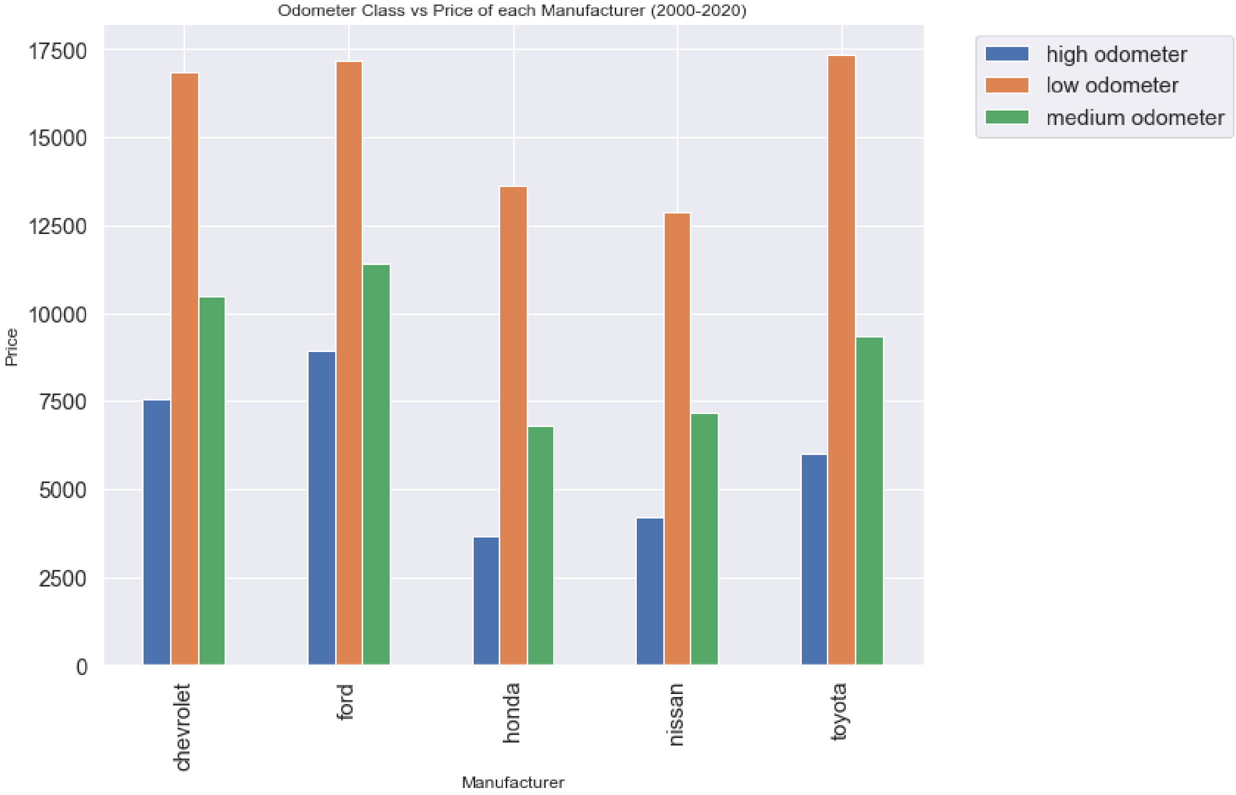
3.3 Odometer

In general, the vehicles with low odometers usually have a higher mean price,followed by vehicles with medium mileage. But vehicles with low odometers also have a wider price range.

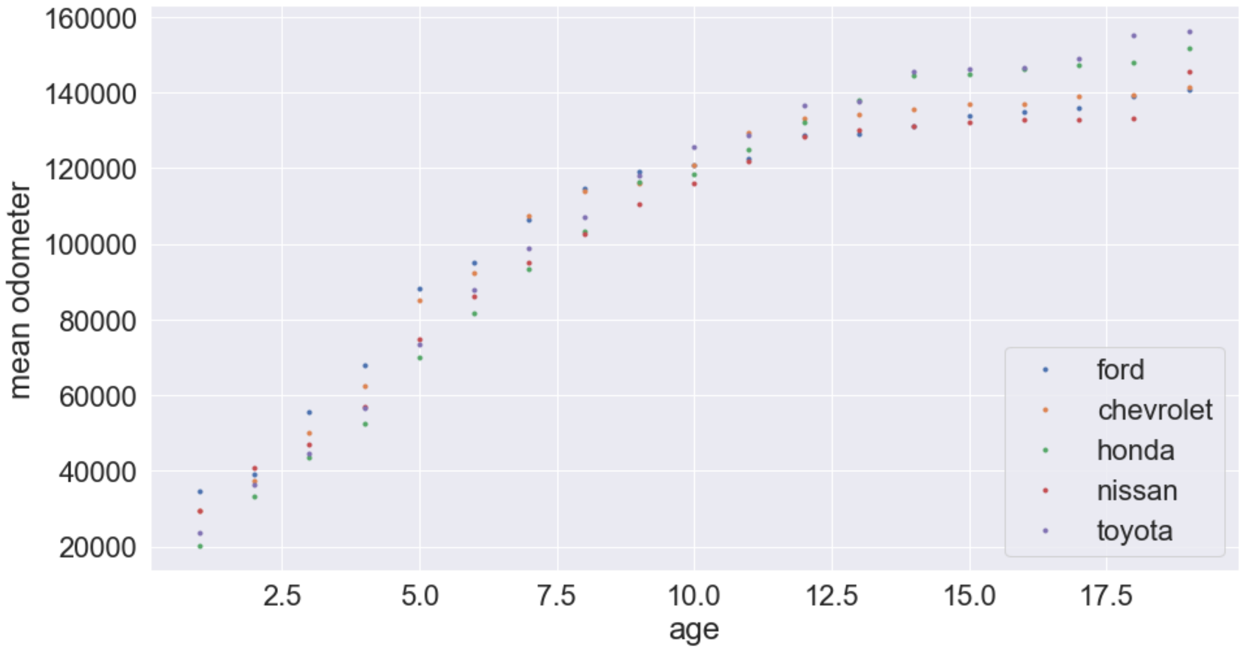


For different manufacturers, the relation between price and odometer is different. Because the performance of different manufacturers may differ as the odometer goes up.

Toyota has the highest mean price of low kilometers(less than 10,000km). The second is Ford's low odometer.



I took Ford as an example. On one hand, Pearson value between ‘odometer’ and ‘price’ is -0.3, which cannot be considered as a strong strength in correlation. Combining the scatter plot, I should say it is a nonlinear relation. In the next stage I need to find out other methods to test their correlation.

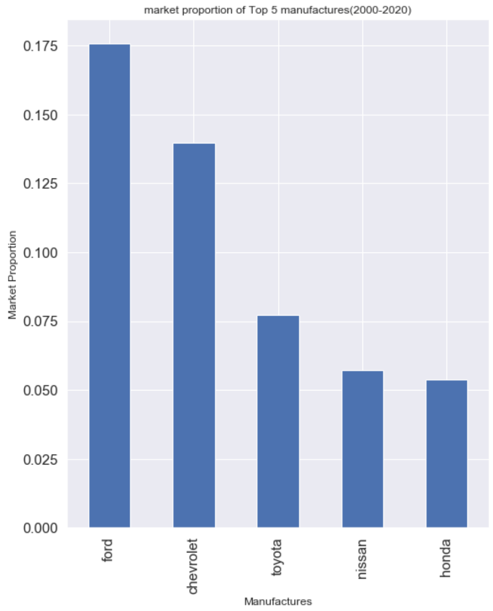


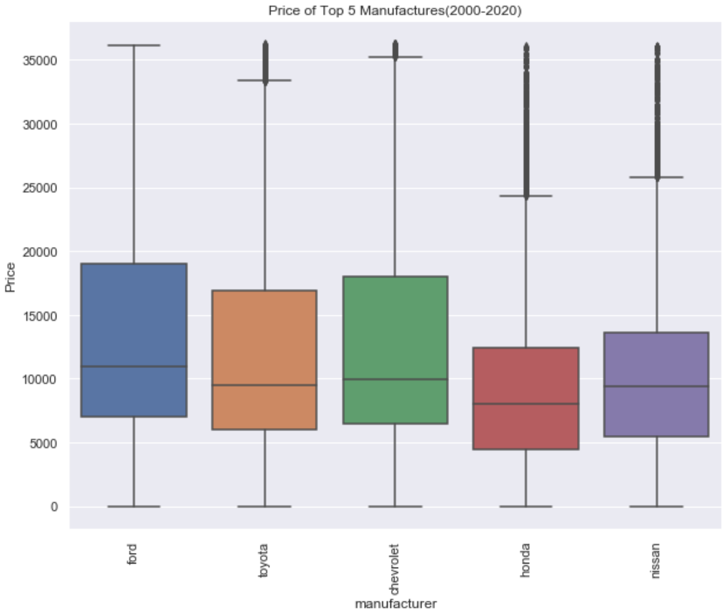
On the other hand, The scatter plot and Pearson 0.5 indicate a positive linear relation between odometer and age. What is remarkable is that the growth rate of odometers slowed down significantly among the vehicles which are over 12.5 years old.

3.4 Manufacturers

Because of the long time of data coverage, many manufacturers have merged and brands have disappeared.So I chose the top five producers with the largest market share, which can illustrate most of the market.

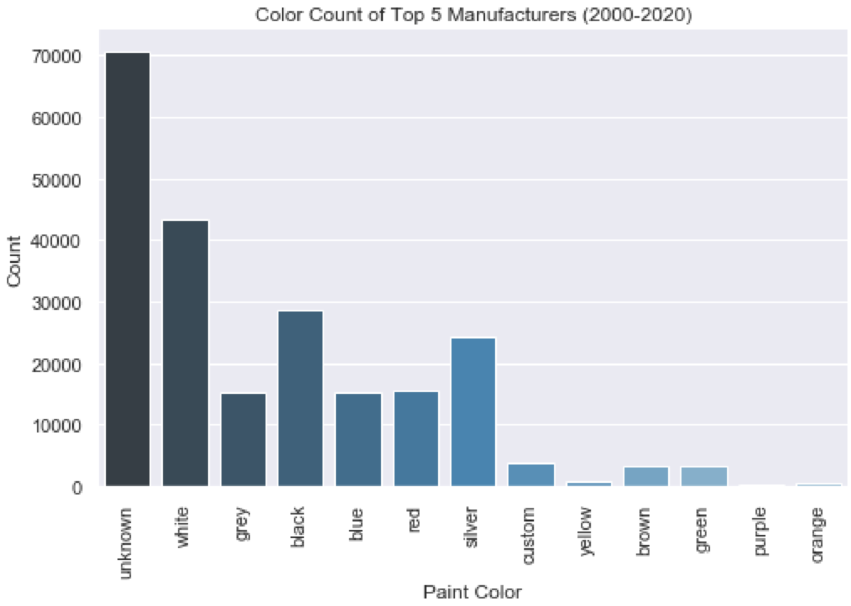
Ford had the highest average price, followed by Chevorlet, Toyota, Nissan and Honda.

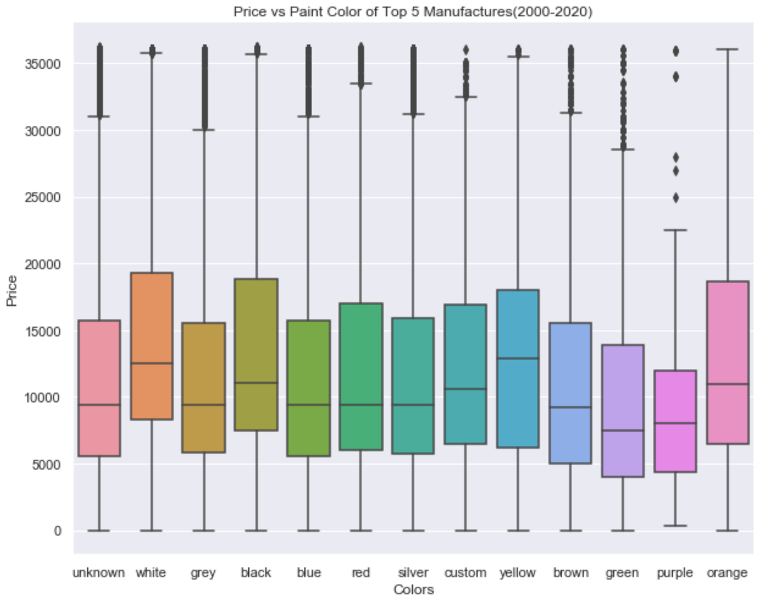




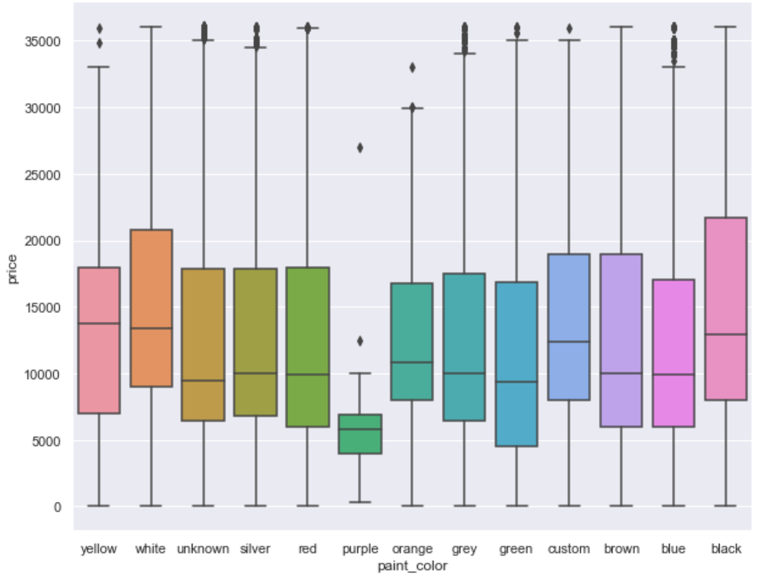
3.5 Color

Generally speaking, white, orange and black usually could be sold at a higher price than other colors. what should be noted that the top three colors in market share are white, black and silver. Probably because the orange is a popular color for customers, however without enough supply.



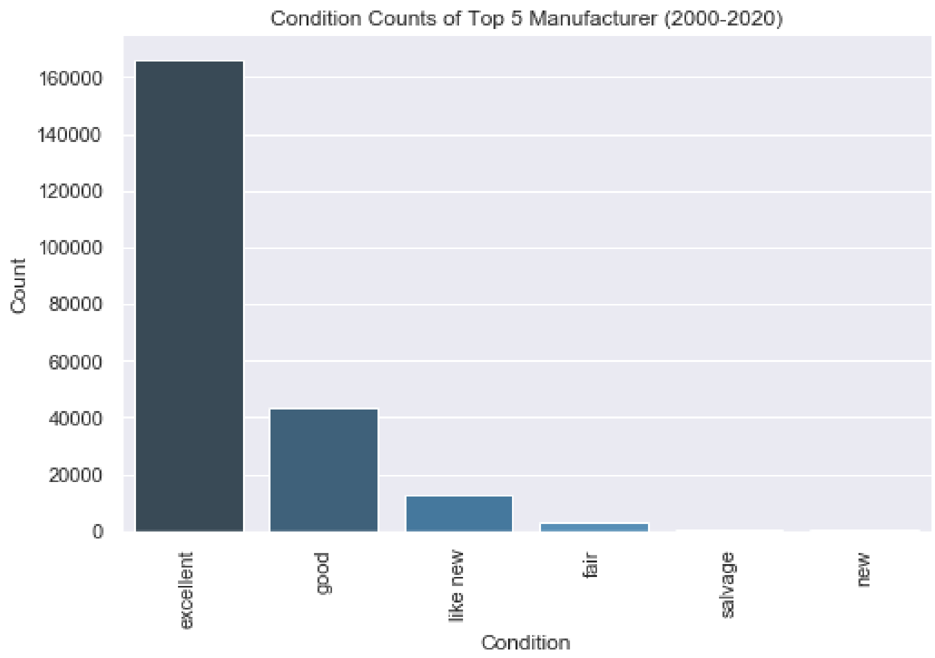


For Ford, White, black and yellow vehicles have the highest median prices.Presumably because Ford's popular models are trucks. Trucks have a great number of business customers who prefer black and white.

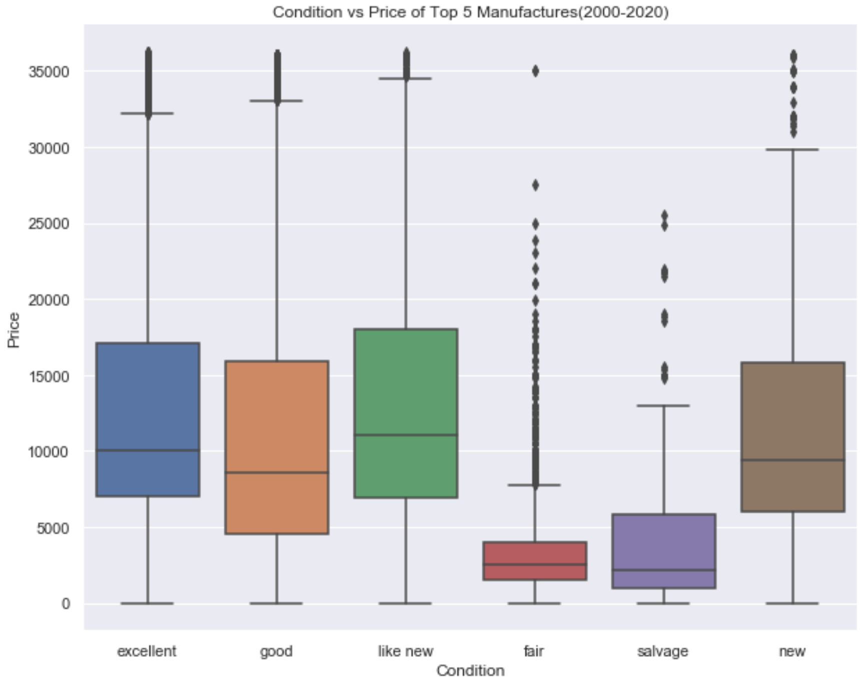


3.6 Condition

Most vehicles on sale are described as ‘excellent condition.’

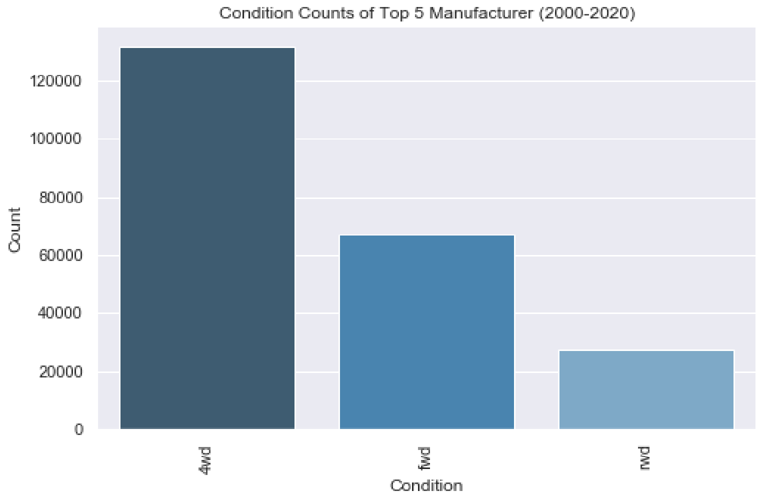


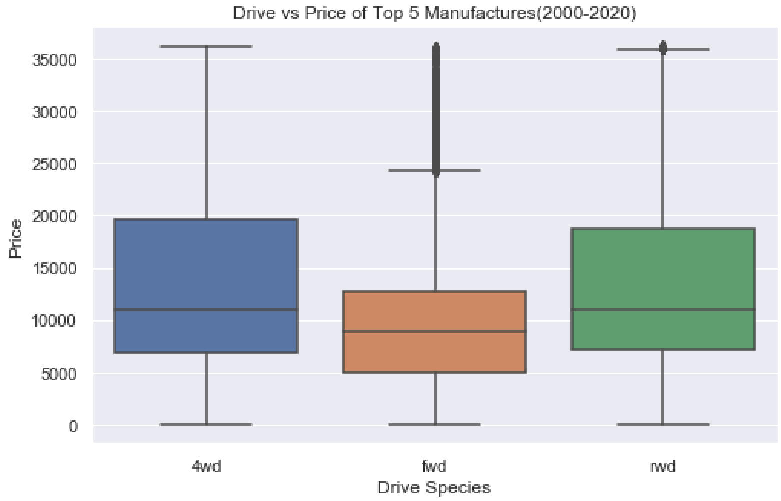
The vehicles labeled as ‘Like new’ and ‘excellent’ condition have the highest median prices. The ‘fair’ and ‘salvage’ ones have the smallest median prices as well as their price range.



3.7 Drive

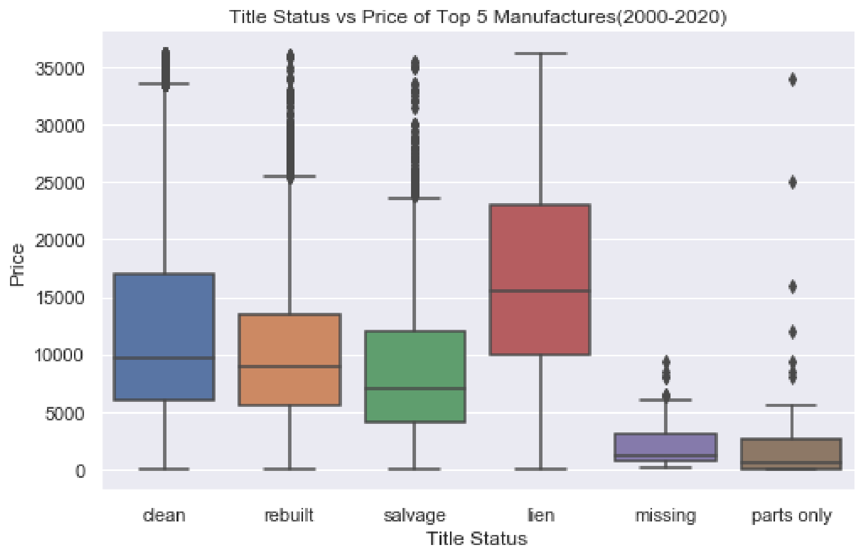
4wd trucks seem to have a bigger share of the market, which could explain why Ford and Chevrolet are so popular. 4wd vehicles always have a higher median price than other two categories.





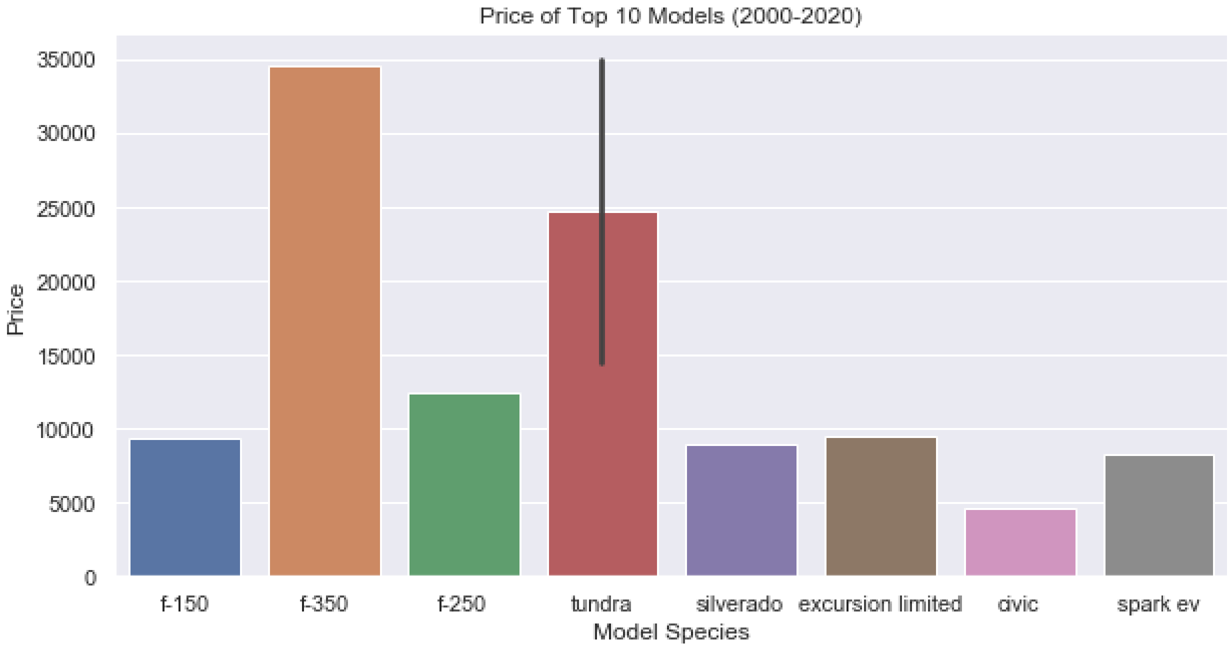
3.8 Title status

Lien vehicles have the highest median price of all, which also have the largest price range.



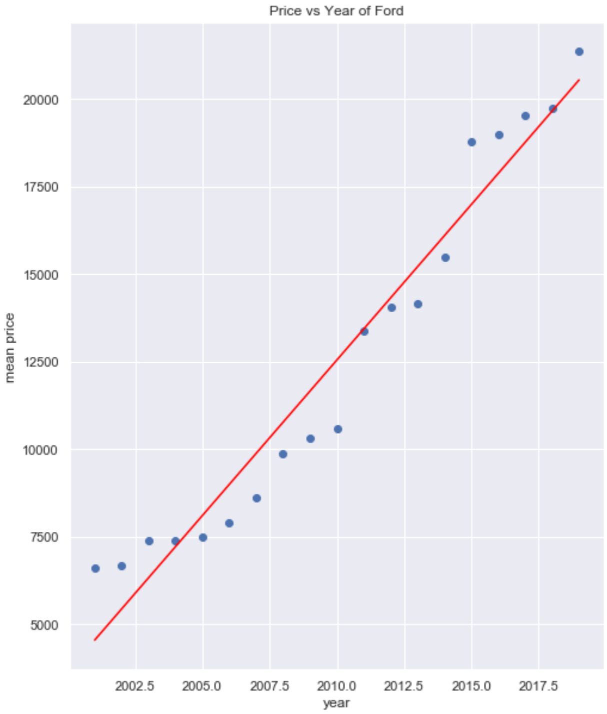
3.9 Model

The F-350 is the most expensive model, followed by the Toyota Tundra.As can be concluded from the plot below, truck price is generally higher, followed by 4WD SUV.



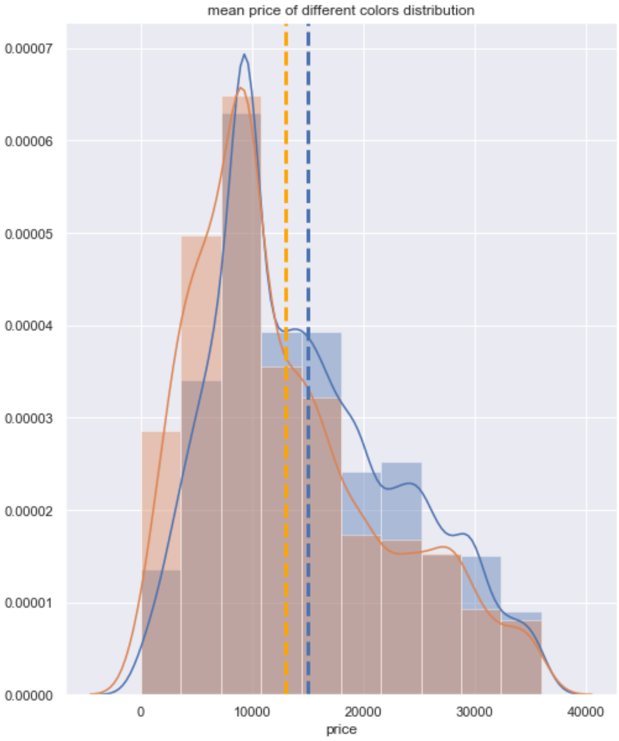
4. Statistical Data Analysis

4.1 Linear Regression



The histogram of Ford’s price showed that they are right skew normally distributed.

After accomplishing the scatter plot of vehicle’s price and age, I could say that they have a strong linear correlation which can be inferred by a Pearson value of 0.5.



4.2 Hypothesis Testing

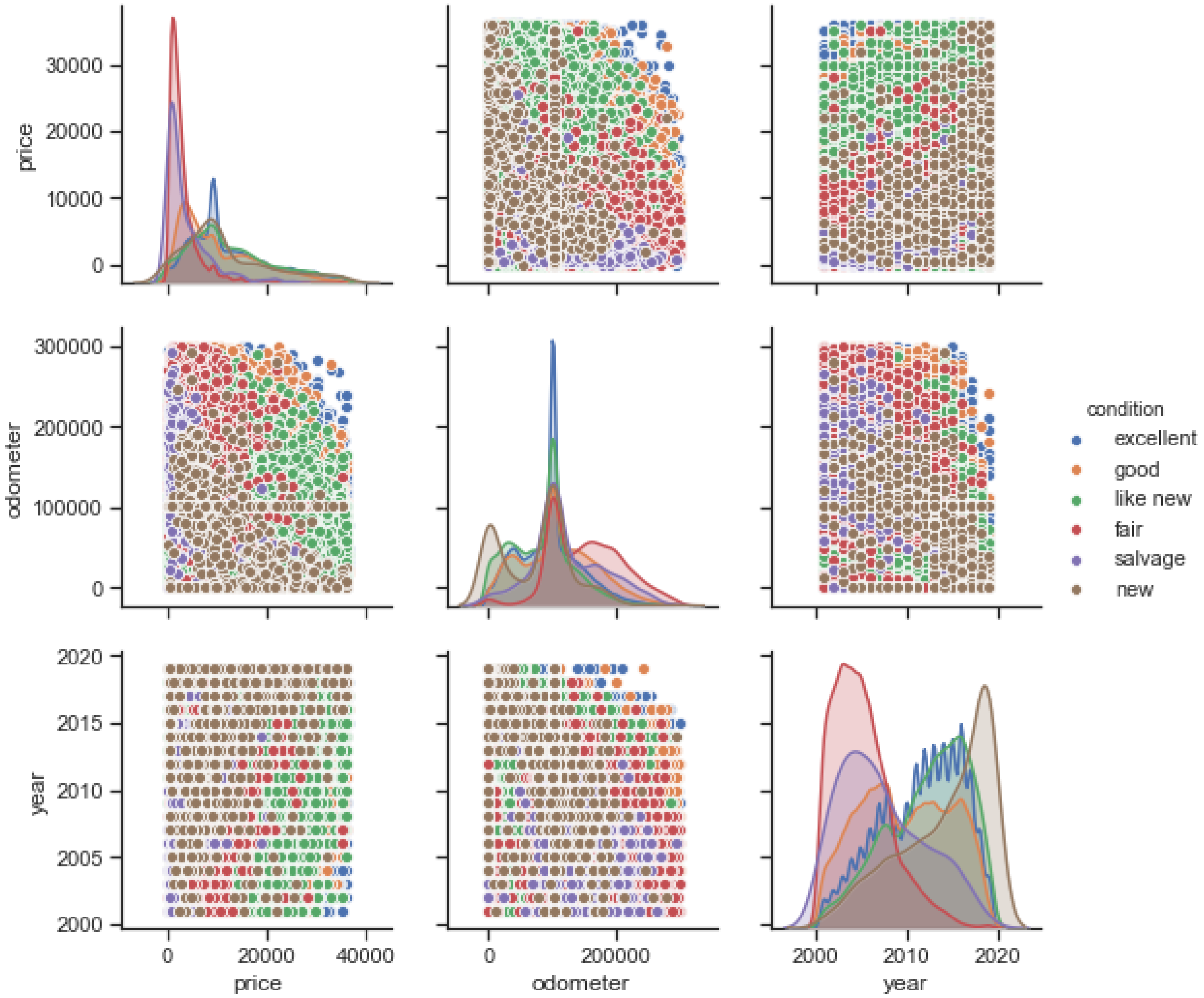
Through the calculation of P-value, I found that among the Ford, the mean price of white cars and silver are different. Meanwhile, it cannot be proved that the prices of white cars and red cars are different.Therefore, there is no correlation between the mean price of different colors.

By the same method, there is no link between market share and average prices.

5. Modeling

The target of this project is used vehicles price prediction of the top 5 manufacturers. It could be considered as a regression topic which would be solved by supervised machine learning algorithms. Since the outcomes are determined by both numerical and categorical variables, I decided to use different modeling ways to compare their performance.

5.1 Prepare for the modeling:



Firstly, I chose the most relevant categorical variables which have an obvious significant effect on the target variable.

Label encoding

I need to convert each text category to numbers in order for the machine to process them using mathematical equations. Since the dataset had several kinds of categorical variables, there are more than two classes in one column. I decided to use the Label Encoding method.

Data Splitting

After that, I split the data set into training and test datasets. Generally the dataset is balanced, I set the training set size as 0.5.

5.1 Dummy Regressor

Dummy Regressor is a modeling method with the simplest rules. I choose it as a baseline for other modeling methods which could show the difference clearly.

5.2 Linear Regression

Since the variable ’year’ and the ‘price’ has a strong linear correlation. I used Linear Regression as my first method. The RMSE of this model is 6873 and r2 score is 0.31, which was not considered to fit well.

5.3 Random Forest Regressor

The reason that I chose Random Forest as a method is that it can provide in general a good predictive performance, low overfitting, and easy interpretability. It could work both for numerical and categorical variables and draw a conclusion about the importance of all variables. Since my data set is too big to run, I use a sample way to do the feature selection work. The RMSE of this model is 6837 and r2 score is 0.31, which was not considered to fit well.

5.4 Gradient Boosting Regressor

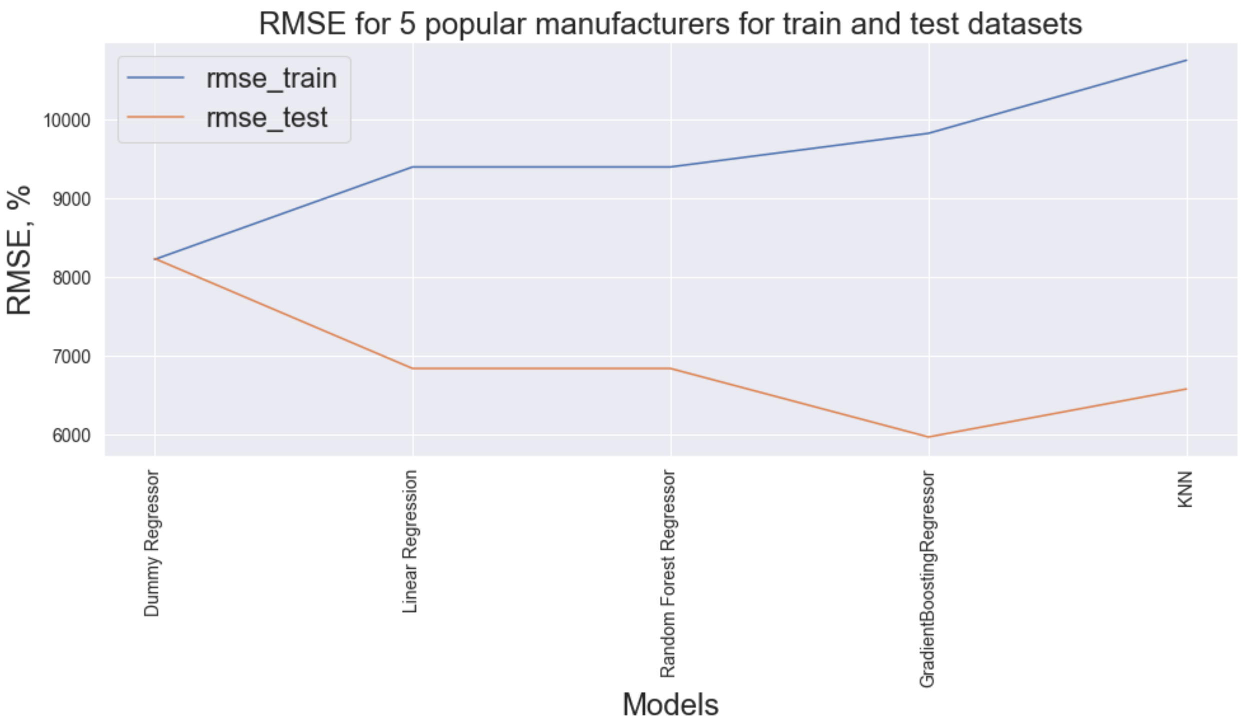
RMSE of Gradient Boosting Regressor is 5966 and r2 score is 0.47, which performed well in model fitting.

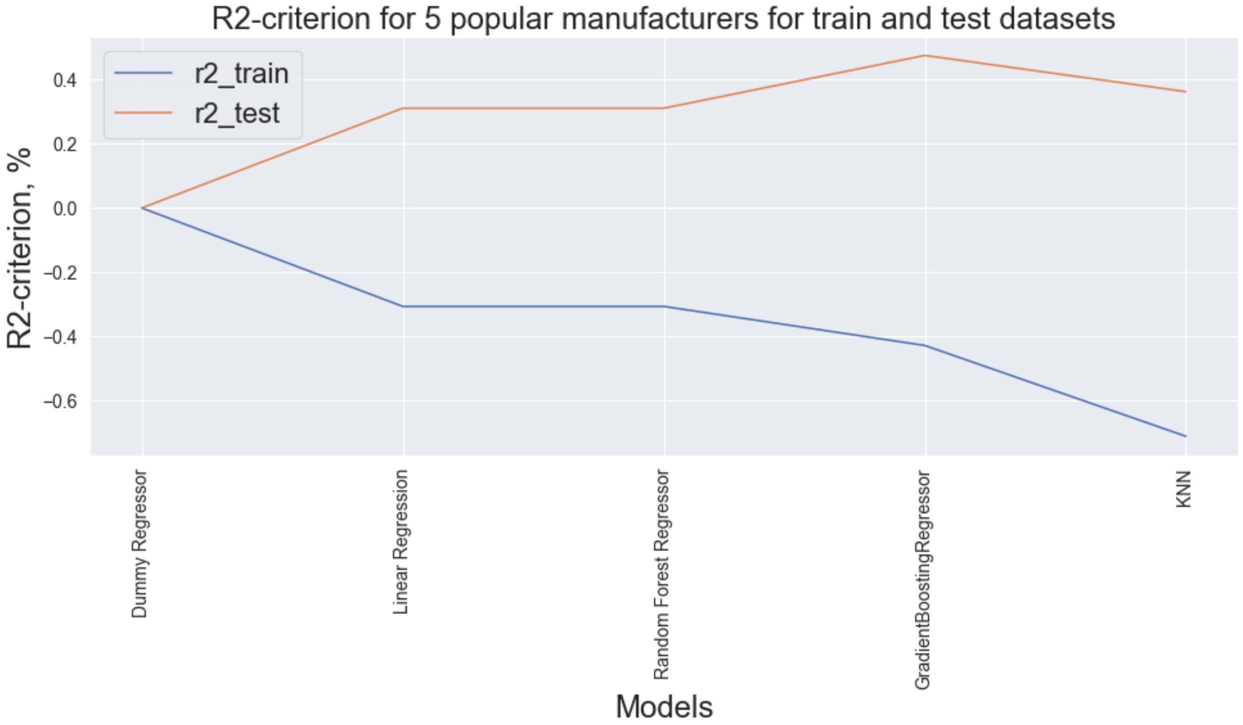
5.5 KNN Regressor

For KNN regressor, firstly I used MinMaxScaler Transform features by scaling each feature to a given range. Then I compare all the RMSE and r2 scores of different K values. From the curve plot of RMSE , the K-value 3 are the best predictors in this model.

5.6 Comparison of all models

Out of 5 models, the best model by the RMSE is Gradient Boosting Regression.


RMSE for 5 popular manufacturers for train and test datasets
rmse_train
rmse_test
RMSE, %
10000
9000
8000
7000
6000
Dummy Regressor
Linear Regression
Random Forest Regressor
GradientBoostingRegressor
KNN
Models


R2-criterion for 5 popular manufacturers for train and test datasets
r2_train
r2_test
R2-criterion, %
0.4
0.2
0.0
−0.2
−0.4
−0.6
Dummy Regressor
Linear Regression
Random Forest Regressor
GradientBoostingRegressor
KNN
Models