greatlearning
Learning for Life

Project report on Craigslist used cars sales price prediction

Submitted towards partial fulfilment of the criteria for award of PGPDSE by Great Learning Institute of Data Science

Submitted By Group No. 2 [Batch: Pune 2022]

Group Members:

1. Sumeet Dalal
2. Vikrant Khadatkar
3. Jatin Choudhary
4. Ritika Pal
5. Sakshi Ghuge

Project Mentor:

- Ankush Bansal

CERTIFICATE OF COMPLETION

I hereby certify that the project titled craigslist car price prediction was undertaken and completed under my supervision by Group 1 of Post Graduate Program in Data Science and Engineering (PGPDSE).

Date:

Signature of the Mentor

Mentor: Ankush Bansal

Acknowledgment:

I would like to express my sincere gratitude to all those who have contributed towards the successful completion of this project on second-hand car price prediction using machine learning.

First and foremost, I would like to thank my project supervisor for providing me with invaluable guidance, support, and constructive feedback throughout the project. His expertise and dedication were invaluable in shaping my understanding of the subject matter and ensuring that the project met the highest standards.

I am also grateful to the team members who worked tirelessly on this project, contributing their skills and knowledge towards achieving the project's goals. Their collaborative efforts and teamwork played a significant role in the successful completion of the project. We certify that the work done by us for conceptualizing and completing this project is original and authentic

Sincerely,

[Group 1]

Abstract:

The second-hand car market is a rapidly growing industry, with consumers looking for affordable and reliable options. However, pricing used cars can be a challenging task, as it is influenced by various factors such as vehicle condition, mileage, age, make, and model. In recent years, machine learning techniques have been used to predict second-hand car prices, offering a more accurate and efficient approach to pricing used cars.

This project aims to develop a machine learning model for predicting second-hand car prices. The project's dataset consists of various car features, such as make, model, age, mileage, engine capacity, and fuel type. The dataset also includes the car's selling price, which serves as the target variable.

Keywords: Data cleaning, Missing data, Outliers, Distribution, Descriptive statistics, Data visualization, Box plot, Correlation, Heatmap, Data transformation, Feature engineering, Data normalization, Dimensionality reduction, Data exploration, Features, Training data, Test data, Validation data, Model, Hyperparameters, Loss function, Feature scaling, RMSE, MSE, r2_Score

greatlearning
Learning for Life

Table of Contents

Industry Review

The second-hand car market is a large and growing industry worldwide. In many countries, the market for used cars is actually larger than the market for new cars, and it is estimated that around 70 million used cars are sold globally each year.

One of the main drivers of the second-hand car market is the cost savings that can be achieved compared to buying a new car. Used cars are typically much cheaper than new cars, and buyers can often get more features and options for their money.

Another factor driving the growth of the second-hand car market is the increasing reliability and durability of modern cars. As cars have become more advanced, they have become more reliable and can last longer than ever before. This means that buyers can purchase a used car that is still in good condition and has plenty of life left in it.

In recent years, the second-hand car market has also been influenced by technology and e-commerce. Online platforms such as Craigslist, Autotrader, and Carfax have made it easier than ever for buyers and sellers to connect and conduct transactions. Additionally, advances in data science and machine learning are being used to provide more accurate pricing information and to help buyers and sellers make more informed decisions.

However, there are also challenges facing the second-hand car market, such as the potential for fraud and the difficulty of assessing the condition of a used car. Buyers and sellers must be vigilant and take steps to protect themselves from scams and ensure that they are getting a fair deal.

Overall, the second-hand car market is a dynamic and growing industry that offers many opportunities for both buyers and sellers. As technology continues to evolve, it is likely that the market will become even more efficient and transparent, making it easier for people to find the right car at the right price.

Literature Survey

The first paper is Predicting the price of Used Car Using Machine Learning Techniques. In this paper, they investigate the application of supervised machine learning techniques to predict the price of used cars in Mauritius. The predictions are based on historical data collected from daily newspapers. Different techniques like multiple linear regression analysis, k-nearest neighbors, naïve bayes and decision trees have been used to make the predictions.

The second paper is Price Evaluation model in second hand car system based on Random Forest. In this paper, the price evaluation model based on using clustering methods and logistic regression methods and KNN methods for predicting the output of car price. Also, they use random forests and decision tree algorithms. At last, they compare all the accuracies of all the machine learning algorithms and choose the best algorithms for the prediction. It aims to establish a second-hand car price evaluation model to get the price that best matches the car.

Summary of problem statement, data, and findings:

This project is focused on determining the price of the car that is worthiness of the car based on variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Different models and systems may contribute on predicting power for a used car's actual market value. It is important to know their actual market value while both buying and selling.

Overview of the final process:

The problem-solving methodology involved preprocessing the data by scaling and normalizing the features. Several machine learning algorithms were used, including linear regression, decision tree regressor, KNN regressor, random forest regressor. These were performed on unscaled and scaled data. Important features were mentioned.

Dataset and Domain:

Data have 426880 rows and 26 columns

Data Dictionary:

Id: uniquely identification number of rows.
url: The URL of the posting on Craigslist.
Region: contains region from USA.
region_url: The URL of region.
price: The price of the car as listed on Craigslist.
year: The year of the car's manufacture.
manufacturer: The car manufacturer (e.g., Ford, Toyota, etc.).
model: The car model (e.g., Mustang, Corolla, etc.).
condition: The condition of the car (e.g., new, used, salvage, etc.).
cylinders: The number of cylinders in the car's engine.
fuel: The type of fuel the car uses (e.g., gasoline, diesel, electric, etc.).
odometer: The number of miles or kilometers on the car's odometer.
title_status: The status of the car's title (e.g., clean, salvage, rebuilt, etc.).
transmission: The type of transmission (e.g., automatic, manual, etc.).
VIN: vehicle identification number of vehicle.
drive: The type of drive (e.g., front-wheel drive, rear-wheel drive, all-wheel drive, etc.).
size: The size of the car (e.g., compact, mid-size, full-size, etc.).
type: The type of vehicle (e.g., sedan, SUV, truck, etc.).
paint_color: The color of the car's exterior.
image_url: The URL of car image.
description: A brief description of the car, often provided by the seller.
County: Contain NULL values.
State: The code for state from USA.
Lat: Latitudes for car location.

Long: Longitude for car location.
posting_date: Contains date on which the sale ad was posted on craiglist

Numeric Variables:

- id
- Odometer
- Price

Categorical Variables:

- url
- region
- region_url
- year
- manufacturer
- model
- condition
- cylinders
- fuel
- title_status
- transmission
- VIN
- Drive
- Size
- Type
- Paint_color
- Image_url
- Description
- County
- State
- Lat
- Long
- Posting_date

The dataset contains 3 numeric variables and 23 categorical variables. The numeric variables contain numerical data and can be used for quantitative analysis, while the categorical variables contain categorical data and can be used for qualitative analysis.

Pre-Processing:

Before performing any data analysis, it is essential to preprocess the data and handle any missing values or redundant columns. The pre-processing data analysis for the Craglist car sales dataset:

id	0
url	0
region	0
region_url	0
price	0
year	1205
manufacturer	17646
model	5277
condition	174104
cylinders	177678
fuel	3013
odometer	4400
title_status	8242
transmission	2556
VIN	161042
drive	130567
size	306361
type	92858
paint_color	130203
image_url	68
description	70
county	426880
state	0
lat	6549
long	6549
posting_date	68

Redundant columns:

Columns that provide no additional information to the analysis.

Id: It is just a unique identifier for each row and does not provide any useful information for analysis

County: have 100% null values thus it will not give any information.

Description: the owner of the dataset already extracted the information from description column and created this dataset.

Url, Image_url, image_url, region_url: columns contain urls of car and all are unique so it won't add any information to the model building

VIN: It is just a unique identifier for identification number of cars and all are unique and does not provide any useful information for analysis

Posting_date: it contains the posting date of cars on website of craglist and all are of same month and year

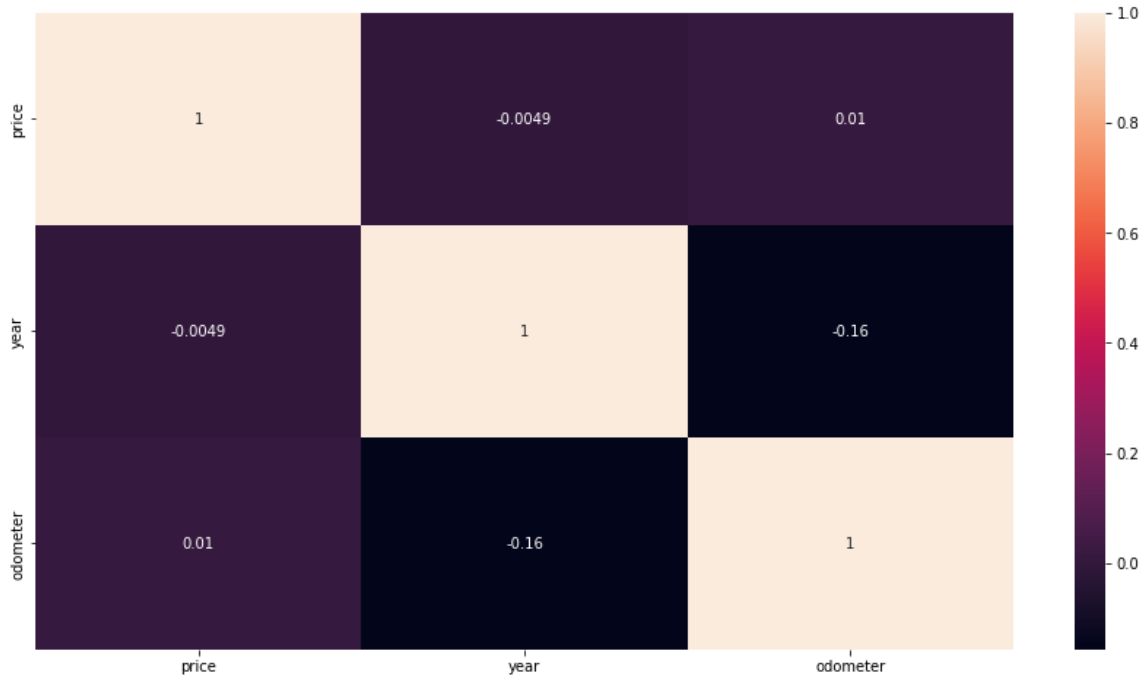
Thus, we drop these columns.

Step-by-step walkthrough of the solution:

The first step was to check for missing values and outliers in the data preprocess the data by scaling. Next, the data was split into training and testing sets. Base model with all significant variables was built, the base model is not understanding the data so we again perform outlire treatment and fit the next model i.e decision tree regressor. Various machine learning algorithms (linear Regression, Decision Tree regressor, KNN regressor, Random Forest regressor) were applied to the training data, and their performance was evaluated on the testing set. The best-performing models (Random Forest regressor) was selected.

Visualization:

Relationship between Variables:



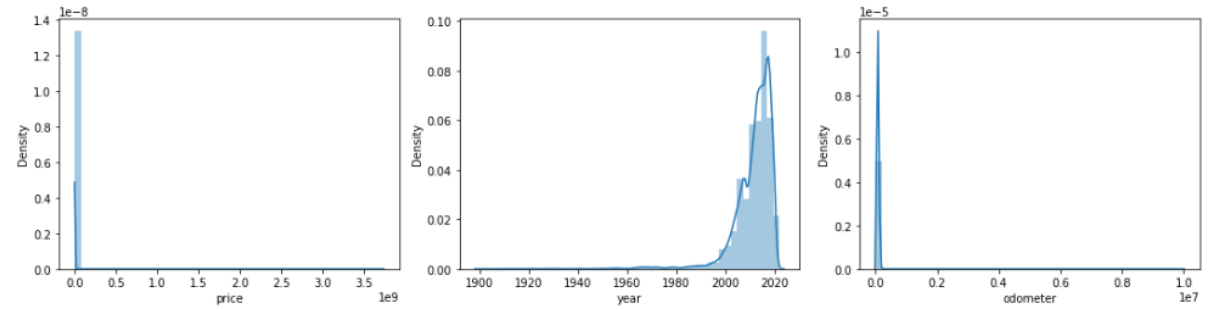
summary of the relationships between the variables in the dataset:

the numeric features don’t have any correlation

Multi-Collinearity:

Dataset have only one independent numeric column which is significant for model building and there is no multi-collinearity

Distribution of Variables:



Inference:

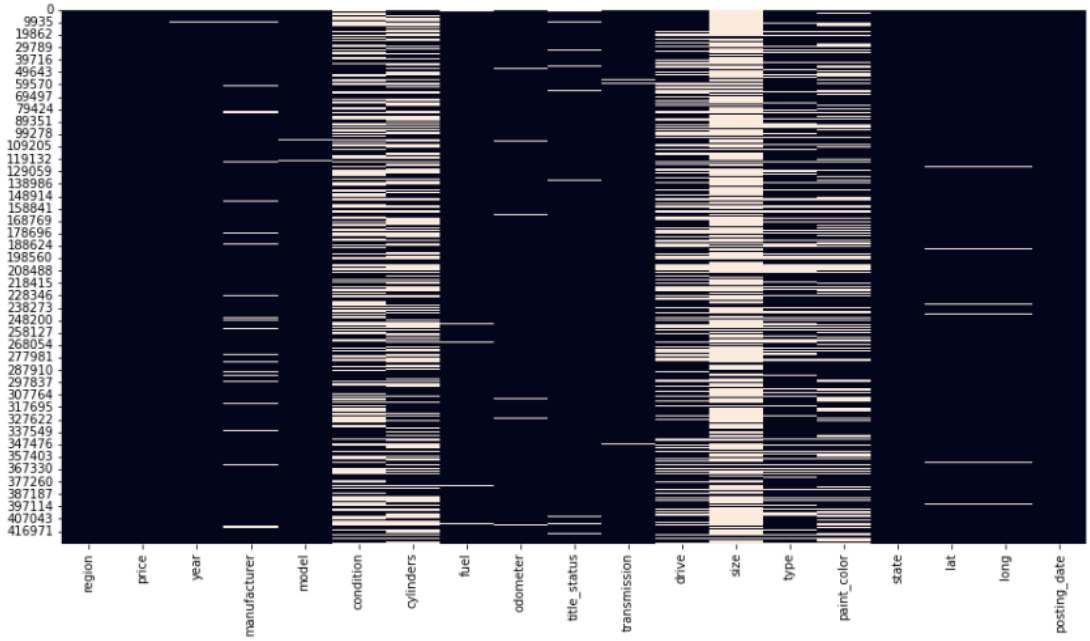
the price column has extreme outlier and it is right skewed i.e the price of maximum cars are low but some of cars have high price.

The year column has majority values is between 1980 to 2020

The odometer has maximum values is in low end and also have some extreme value

Null Value treatment:

```
In [21]:
#visualizing the null values in the data
plt.figure(figsize=(15,8))
sns.heatmap(df.isnull(),cbar=False)
plt.show()
```



We remove the title_status columns because the 84% of data is one class, remove rows which have null values in fuel and transmission, ‘not specified’ is replaced with null values in condition column, unknown in paint_color

We create new column which is a combination of manufacture and model and then calculate the median for cylinder using group by of new column and replace the null values from median of cylinder obtain by group by new column

Same as for drive, type, odometer. And we remove the rows which have null value in year columns.

treating missing values of cylinder column

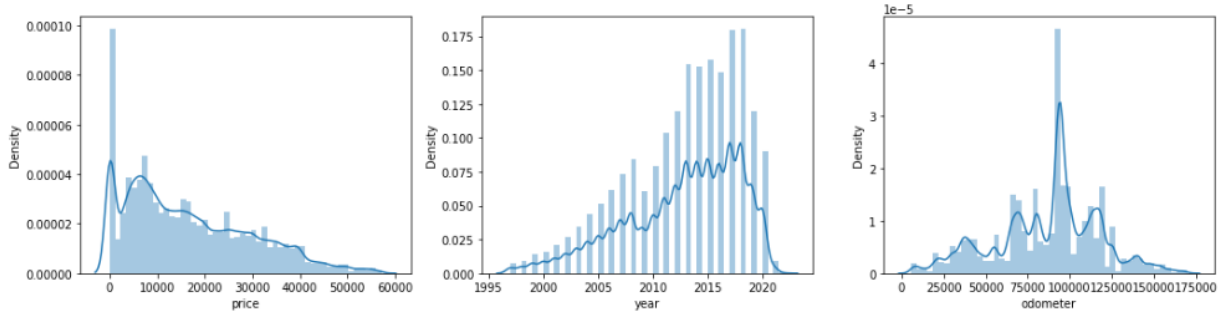
```
In [36]:
m1 = df['model'].tolist()
m1 = [x.lower().strip() for x in m1]
m2 = [y.split()[0] for y in m1]
m2 = [x.replace(' ', '').replace('-', '').replace('/', '') for x in m2]
df['car_model'] = m2
df['car_model'] = df['manufacturer'] + " " + df['car_model']
df['car_model'].nunique()

Out[36]: 3473

In [37]:
df = df[df['cylinders'] != 'other']
dfa = df.copy()
dummy = df[df['cylinders'].notnull()].copy()

In [38]:
dummy['cylinders'] = [int(m.split()[0]) for m in dummy['cylinders']]

In [39]:
med = dummy.groupby('car_model')['cylinders'].median()

In [40]:
merged = pd.merge(dfa, med, on = 'car_model', how = 'left')

In [41]:
merged['cylinders_y']=merged['cylinders_y'].replace(np.nan,merged['cylinders_y'].median())
merged['cylinders_y'].isnull().sum()

Out[41]: 0

In [42]:
df['cylinders']=merged['cylinders_y'].tolist()

In [43]:
df['cylinders'].isnull().sum()

Out[43]: 0
```

Presence of Outliers and its Treatment:

We can identify outliers in the dataset using box plots. Outliers are data points that are significantly different from the rest of the data and can affect the accuracy of the model. There are various techniques for treating outliers, including removing them, transforming the data,

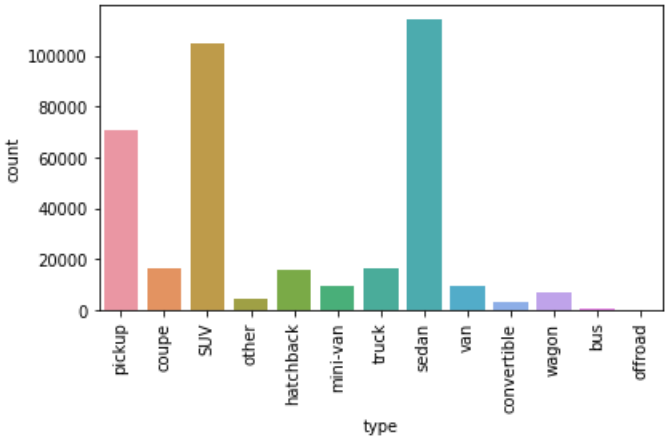
There are Extream Outliers in dataset so we removed outliers using removing quantile of less than 10% and above

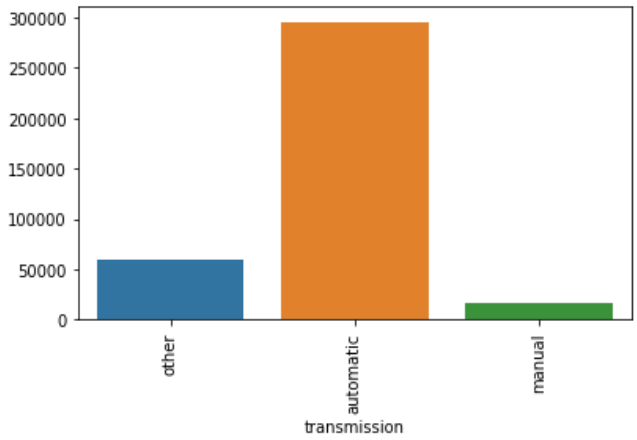
99% technique, and the data after treating outliers is shown using distplot in figure below.



Inference:

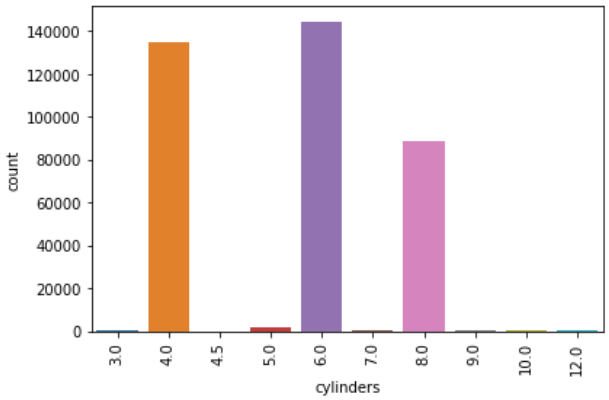
after outlier treatment the extreme values are removed and data distribution is visible price column have right skewed and year have left skewed distribution odometer have leptokurtic distribution

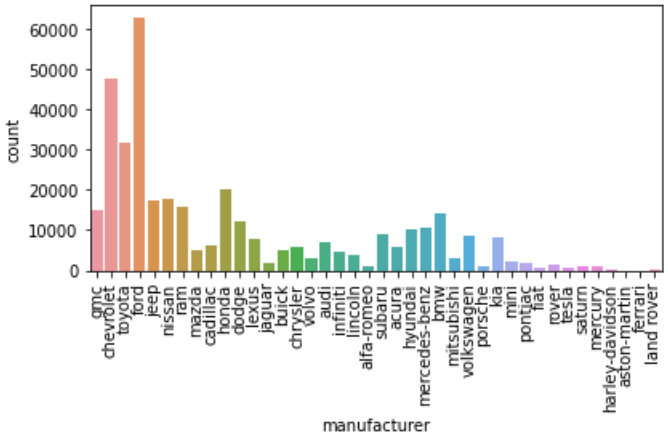




Inference:

In type column pickup, SUV, sedan is in majority class and remaining are in minority, sedan is most seller car



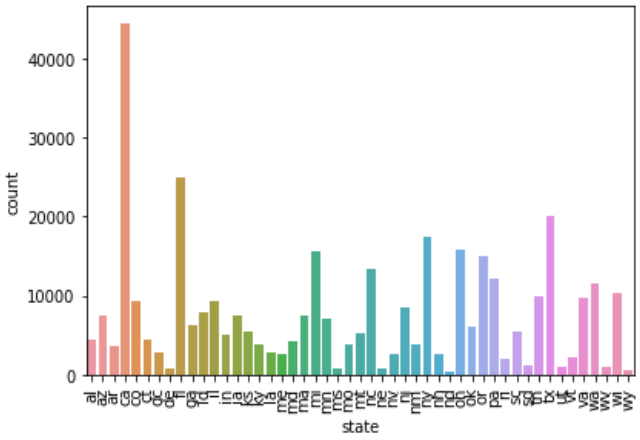


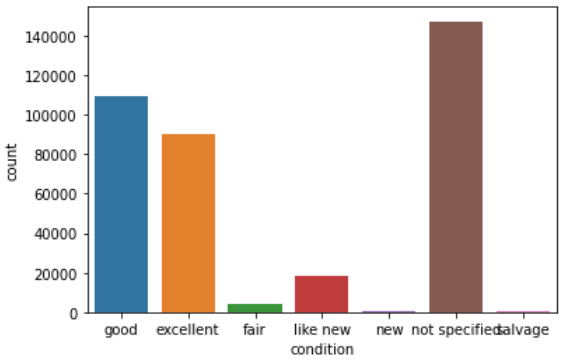
Most of transmission type in data is automatic type

Inference:

Majority of cars have 4,6,8 cylinder in our dataset

Manufacturer have ford Toyota Chevrolet and gmc as majority manufacturers

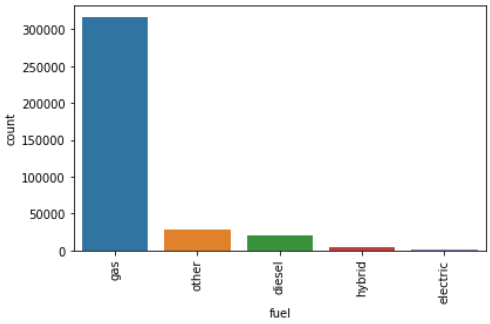


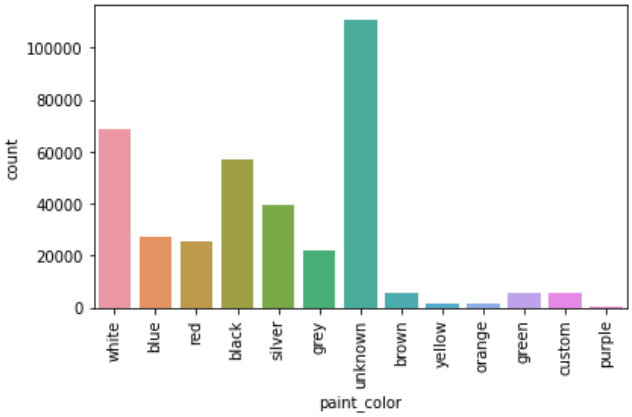


Inference:

Most of cars cum from California then from florida

Most of car's condition is not specified then have good and excellent condition

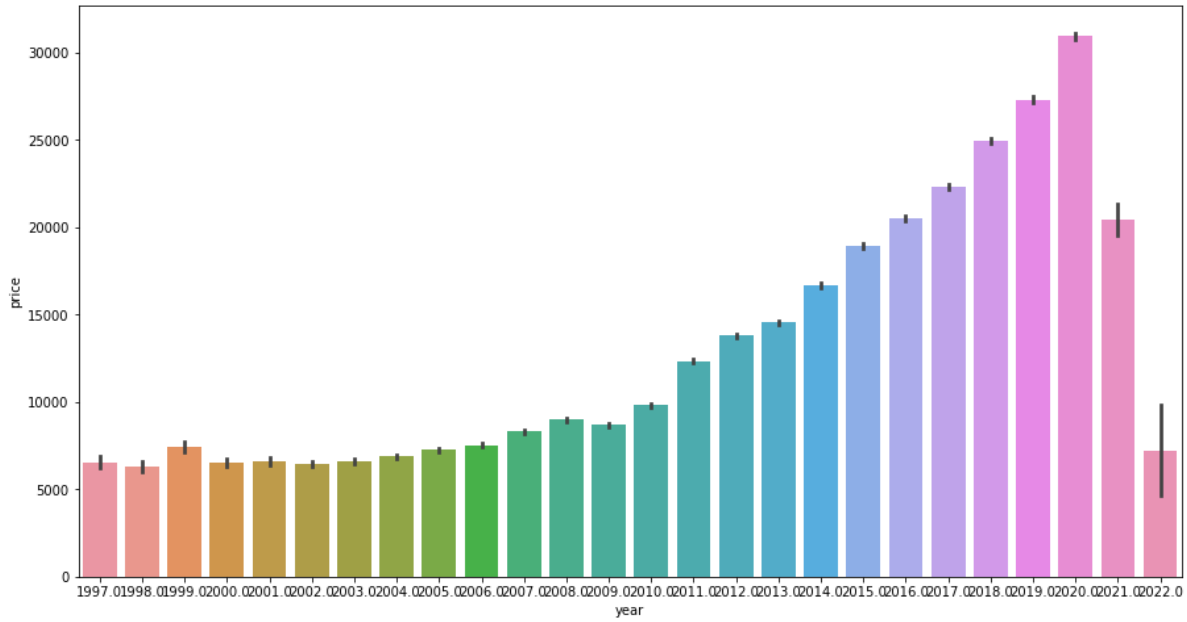


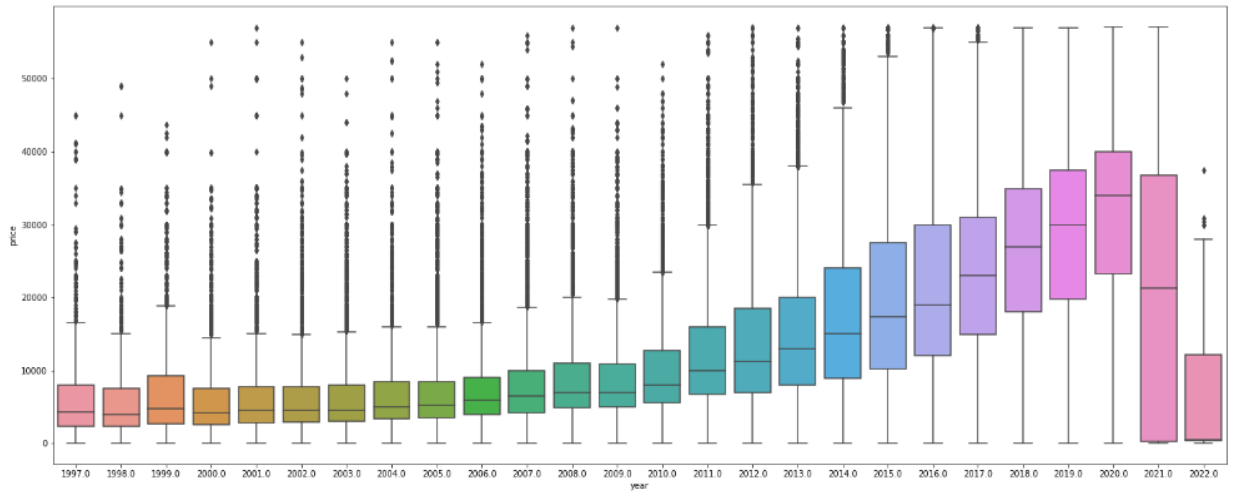


Inference:

Most of cars from our dataset is run on gas(petrol)

Majority of car paint is unknown and with paint white, black, silver, red and blue are common cars



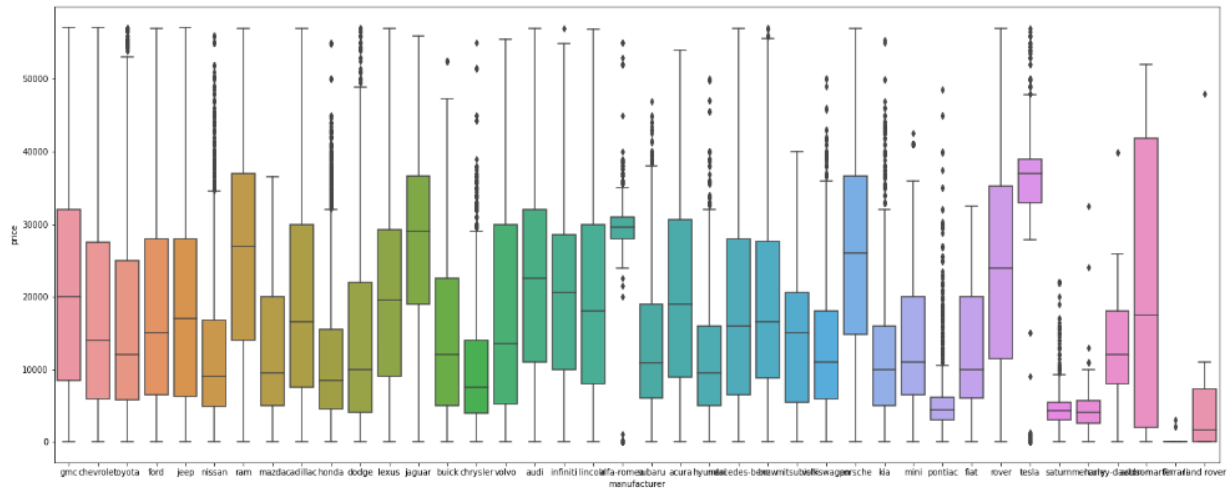


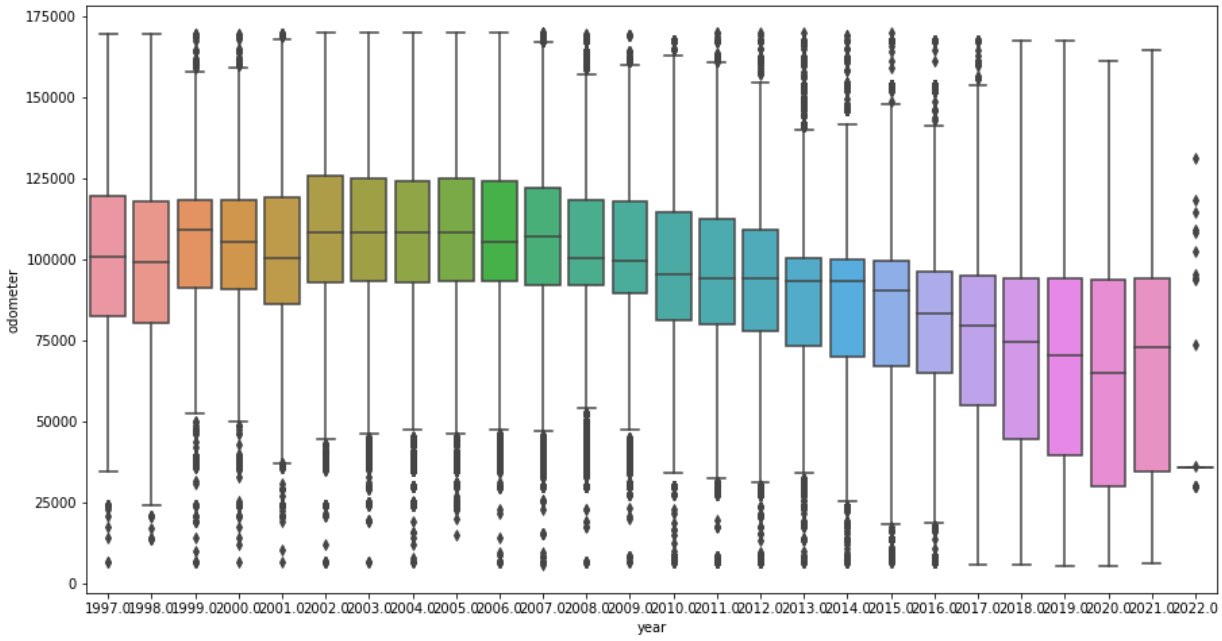
Inference:

As vehicle is old the price is less, and as vehicle is new the price is high but there are outliers in old vehicles because of there might be vintage cars. In 2021 and 2022 their price is low and variation is high

Inference:

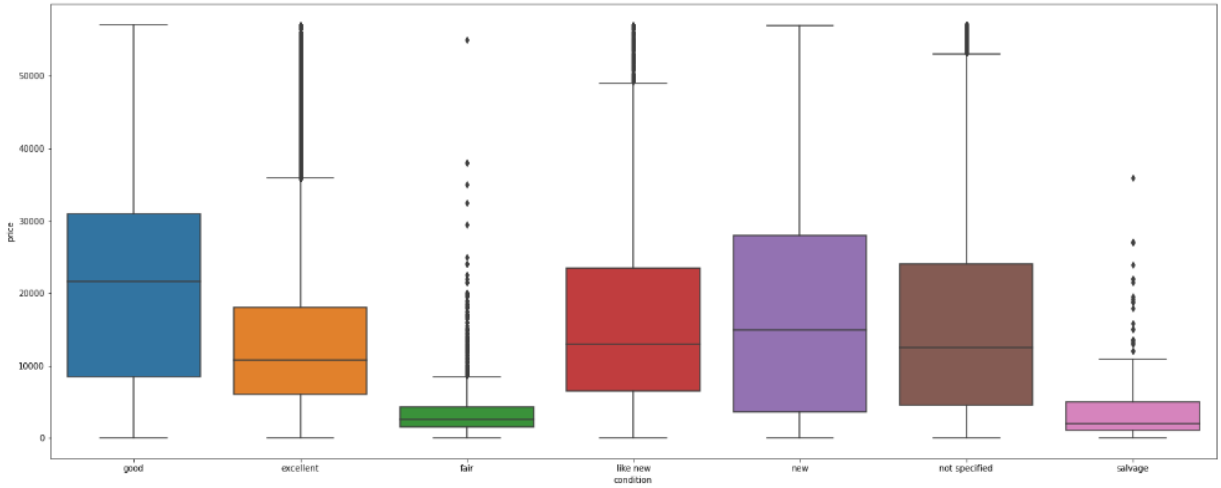
As vehicle is new then the odometer value is low but in 2022 year some cars are running high





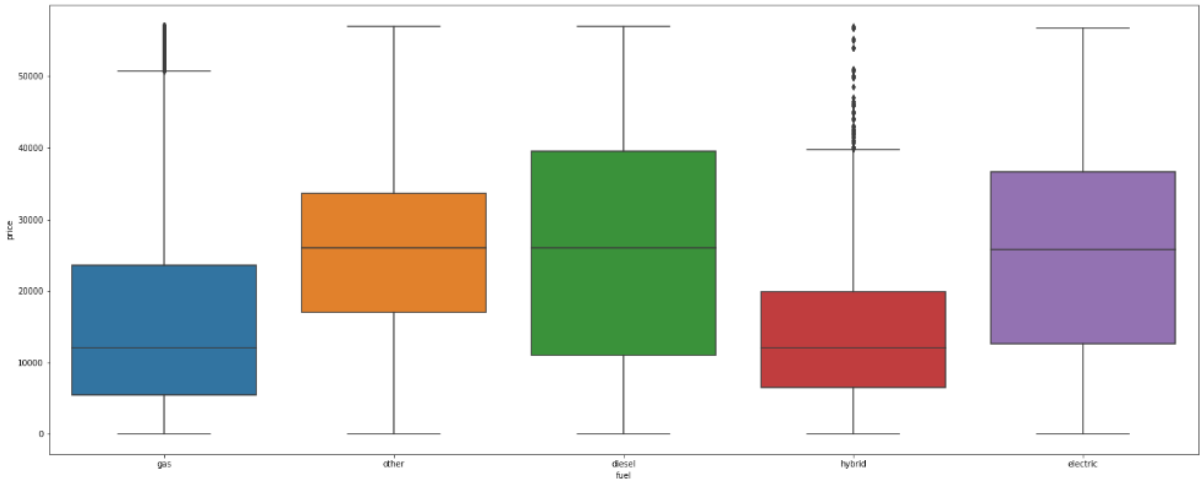
Inference:

There are some manufacturers which have high price because might be they are luxurious brands or might be sports car



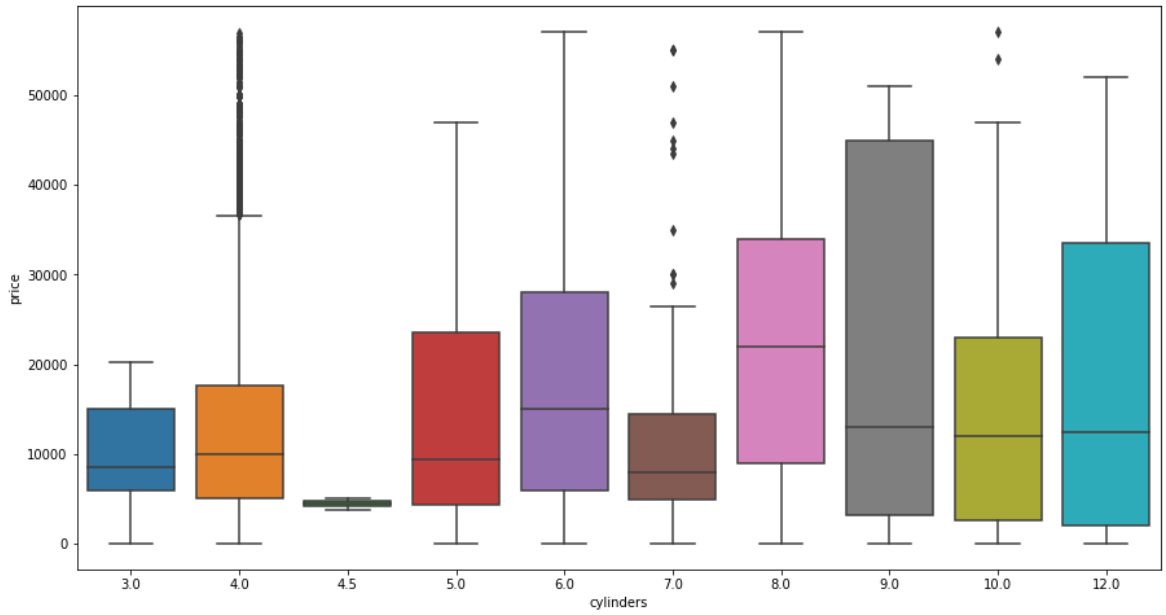
Inference:

The condition of car which are mention in dataset majority of cars are new good and not specified, there are outliers in excellent and fair might be the manufacturer is affecting on the price value with condition. Cars with condition new and good is not affecting by any other features



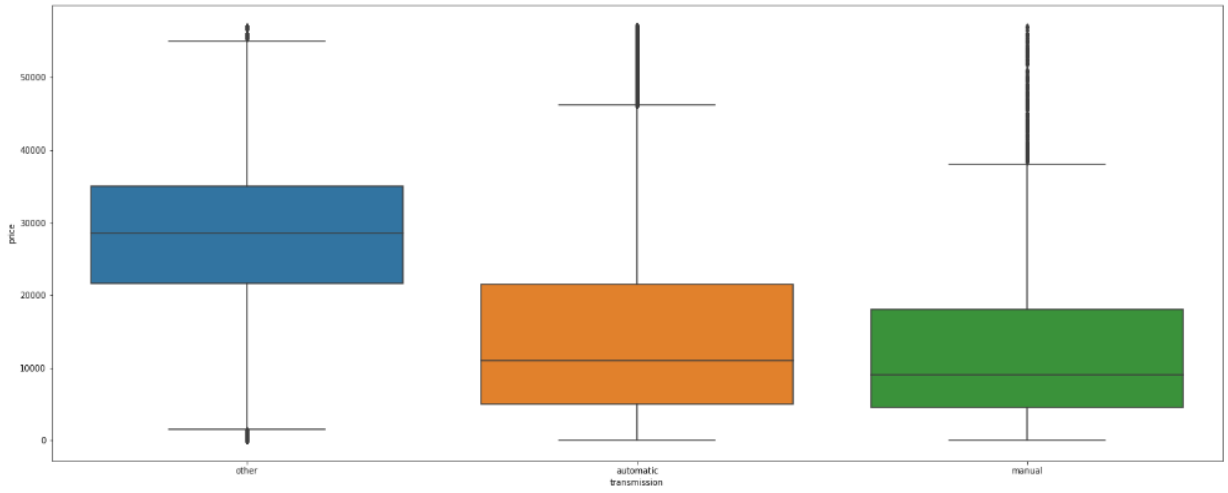
Inference:

Fuel type diesel have maximum price compare to gas, hybrid and gas fuel type cars are having low price might be because of high in demand



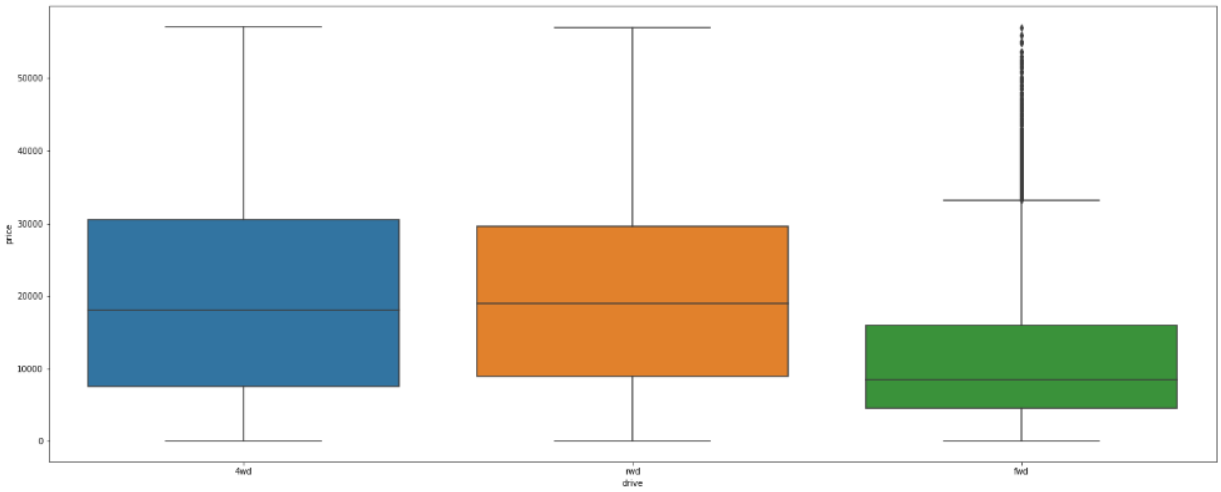
Inference:

Cylinders are high in number the price will high but there is near to similar price in all the cylinders because of condition and there are outliers in 4 cylinder and 7 cylinders might be condition and manufacturer is affecting the price in 4,7 cylinders



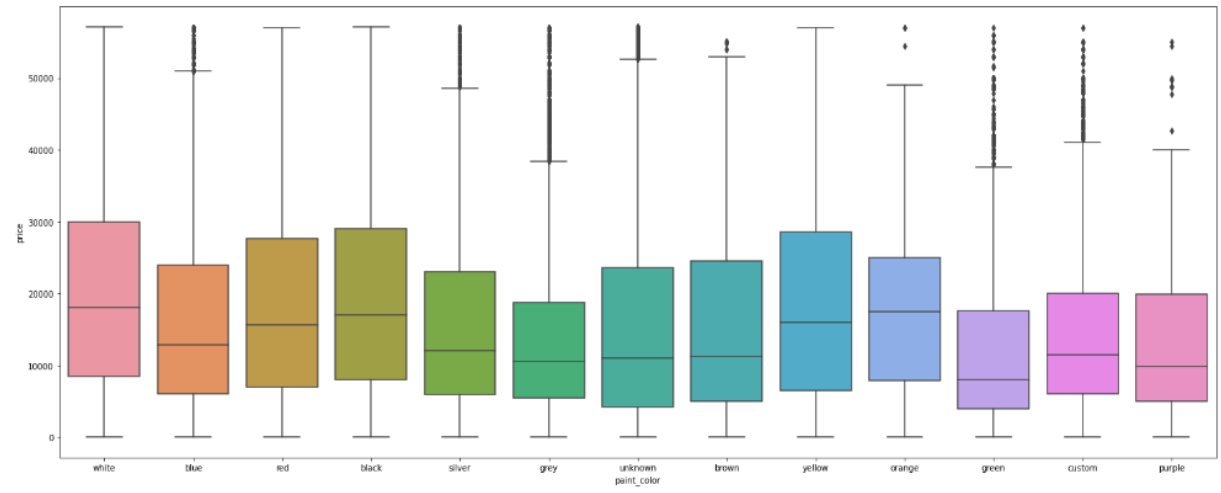
Inference:

Automatic and manual type transmission have near to same price there is slightly different but in other transmission type their price is high compare to manual and automatic



Inference:

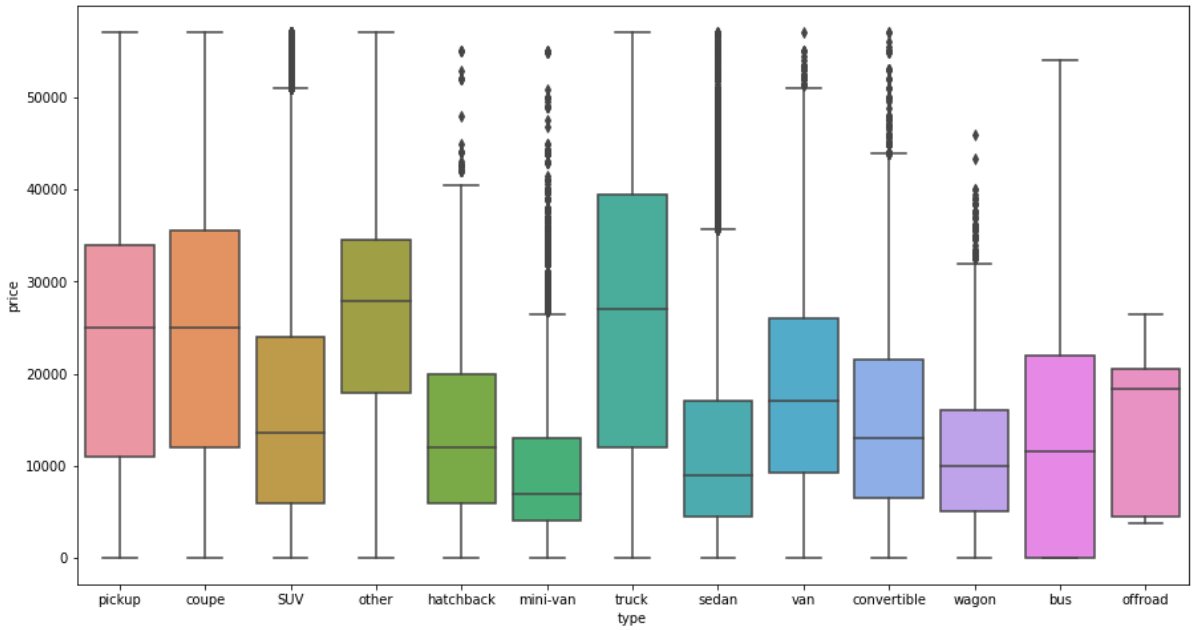
4wd(4 wheel drive) and rwd(rear wheel drive) have same price range but in fwd(front wheel drive) have low price compare to 4wd and rwd but there are outliers in fwd because there are lots of cars have fwd and luxurious and vintage cars have fwd



Inference:

Paint color white, red, black, yellow cars are common in market that's why we cant see much difference in them

But on other hand have all paint color type have outliers in it



Inference:

Offroad, wagon, minivan, have low price compare to other car type but the outliers in sedan minivan are high in number may be cylinder and manufacturer column affecting on the price with type of cars

Hypothesis Testing:

Statistical Significance of Variables:

This will help us to identify the variables that are most important for car price accurately.

```
alpha=0.05
print('H0:The means of all groups are equal.')
print("H1: At least one group mean is different from the others.")
for i in df.drop(['price','odometer','age'],axis=1).columns:
    groups = df.groupby(i)['price'].apply(list)
    f_statistic, p_value = stats.f_oneway(*groups.values)
    print(i)
    print('F-statistic:', f_statistic)
    print('P-value:', p_value)
    # Make a decision
    if p_value < alpha:
        print("Reject the null hypothesis.")
    else:
        print("Fail to reject the null hypothesis.")
    print('-------------------------')
```

```
H0:The means of all groups are equal.
H1: At least one group mean is different from the others.
year
F-statistic: 6171.742185280621
P-value: 0.0
Reject the null hypothesis.
-------------------------
manufacturer
F-statistic: 986.5229706863599
P-value: 0.0
Reject the null hypothesis.
-------------------------
condition
F-statistic: 4206.64454171475
P-value: 0.0
Reject the null hypothesis.
-------------------------
cylinders
F-statistic: 3818.517859684048
P-value: 0.0
Reject the null hypothesis.
-------------------------
fuel
F-statistic: 6014.043308879395
P-value: 0.0
Reject the null hypothesis.
-------------------------
transmission
F-statistic: 30786.394643048796
P-value: 0.0
Reject the null hypothesis.
-------------------------
drive
F-statistic: 20812.33375448614
P-value: 0.0
Reject the null hypothesis.
-------------------------
type
F-statistic: 5093.389230041545
P-value: 0.0
Reject the null hypothesis.
-------------------------
paint_color
F-statistic: 896.6680523436405
P-value: 0.0
Reject the null hypothesis.
-------------------------
region
F-statistic: 360.7138549935147
P-value: 5.990704736630624e-234
Reject the null hypothesis.
-------------------------
age_cat
F-statistic: 47561.157166842604
P-value: 0.0
Reject the null hypothesis.
-------------------------
```

```
print('H0: the means of the two groups are equal.')
print('H1: the means of the two groups are not equal.')
t_statistic, p_value = stats.ttest_ind(df['odometer'], df['price'])
print('T-statistic:', t_statistic)
print('P-value:', p_value)
alpha = 0.05

# Compare the p-value to the significance level
if p_value < alpha:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
```

```
H0: the means of the two groups are equal.
H1: the means of the two groups are not equal.
T-statistic: 1268.4907414842983
P-value: 0.0
Reject the null hypothesis.
```

```
print('H0: the means of the two groups are equal.')
print('H1: the means of the two groups are not equal.')
t_statistic, p_value = stats.ttest_ind(df['age'], df['price'])
print('T-statistic:', t_statistic)
print('P-value:', p_value)
# Compare the p-value to the significance level
if p_value < alpha:
    print("Reject the null hypothesis.")
else:
    print("Fail to reject the null hypothesis.")
```

```
H0: the means of the two groups are equal.
H1: the means of the two groups are not equal.
T-statistic: -768.4683875372419
P-value: 0.0
Reject the null hypothesis.
```

Inference:

Odometer, age, year, transmission, manufacturer, drive, condition, type, cylinders, paint_color, fuel, region and age cat are statistically significant

.

Feature Engineering:

Transformations: as we remove outliers so data follows near to normal distribution, so we didn't do any transformation.

Scaling: dataset have only 2 columns which have numeric data one is odometer and another is price but for base model we didn't do any scaling.

Feature Selection: we use all the columns which are statistically significant.

New feature creation: we create the 3 new feature 1st is model feature for treating the null values**,** 2nd feature is a age by calculating year -2023 and then we group the age in 3 () category and create new column age_cat

Assumptions:

Linear regression:

Linear regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is based on several assumptions that must hold in order for the regression results to be valid and reliable. Here are the main assumptions of linear regression:

Linearity: The relationship between the dependent variable and the independent variables is linear.

Independence: The observations in the dataset are independent of each other. There should be no correlation between the errors in the model.

Homoscedasticity: The variance of the errors is constant across all levels of the independent variables. In other words, the spread of the residuals is the same for all values of the independent variables.

Normality: The residuals follow a normal distribution.

No multicollinearity: The independent variables are not highly correlated with each other.

No influential outliers: There are no extreme observations in the dataset that have a disproportionate influence on the regression results.

There are no influential outliers, Normality for odometer is satisfy, seems the majority of data is categorical hence we can't find homoscedastic, all variables are independent with each other, because of categorical columns there are no linear relationship between independent variable and dependent(target) variable.

Base Model (OLS model):

Equation for OLS model:

$y = b0 + b1 * x1 + b2 * x2 + ... + bn * xn + \varepsilon$

where:

- y is the dependent variable
- x1, x2, ..., xn are the independent variables
- b0, b1, b2, ..., bn are the coefficients (or weights) for each independent variable

- ε is the error term or residual, which represents the difference between the predicted and actual values of the dependent variable

Result:

Dep. Variable:	y	R-squared:	0.743
Model:	OLS	Adj. R-squared:	0.743
Method:	Least Squares	F-statistic:	5135.
Date:	Tue, 11 Apr 2023	Prob (F-statistic):	0.00
Time:	23:24:03	Log-Likelihood:	-1.5543e+06
No. Observations:	154854	AIC:	3.109e+06
Df Residuals:	154766	BIC:	3.110e+06
Df Model:	87		
Covariance Type:	nonrobust		

- The R-squared value obtained from the above model is 0.743 which means that the above model explains 7 4.3% of the variation in the price.
- Overall F-Test & p-value of the Model:
  Ho: All β's are equal to zero (i.e. regression model is not significant)
  H1: At least one β is not equal to zero (i.e. regression model is significant)
- Prob (F-statistic): 0.00

As the p-value is less than 0.05, we accept the alternate hypothesis i.e. the regression model is significant.

Omnibus:	10672.278	Durbin-Watson:	1.999
Prob(Omnibus):	0.000	Jarque-Bera (JB):	54798.434
Skew:	-0.055	Prob(JB):	0.00
Kurtosis:	5.912	Cond. No.	1.27e+21

1. Independence of observations should exist (Absence of Autocorrelation):

   Durbin Watson test
   H0: There is no autocorrelation in the residuals
   H1: There is autocorrelation in the residuals

   The summary output shows that the value of the test statistic is close to 2 (=1.998) which means that there is no autocorrelation.

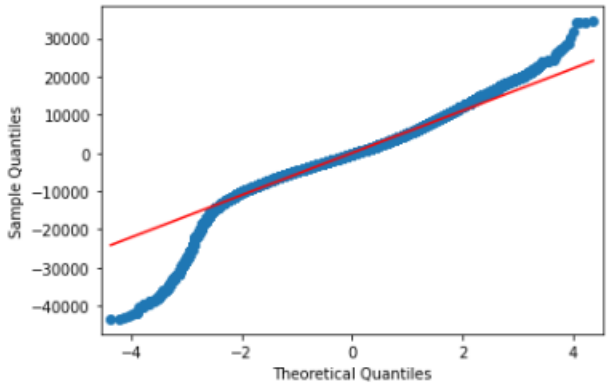
2. Predictors must not show Multicollinearity:
   Multicollinearity test
   The 'Cond. No.' (=1.27e+21) represents the Condition Number (CN) which is used to check the multicollin earity.
   It can be seen that there is severe multicollinearity in the data.

3. The error terms should be Homoscedastic:

Breusch–Pagan test
H0: The residuals are homoscedastic

H1: The residuals are not homoscedastic (heteroscedastic)
[('f-value', 12725.790919740215), ('p-value', 0.0)]
We observe that p-value is less than 0.05 and thus reject the null hypothesis.
We conclude that there is heteroskedasticity present in the data.

6. Tests for Normality:



7. Jarque-Bera Test
H0: The data is normally distributed
H1: The data is not normally distributed
Here, the p-value of the test is less than 0.05 Page 29 of 48
This implies that the residuals are not normally distributed

Significance of features:

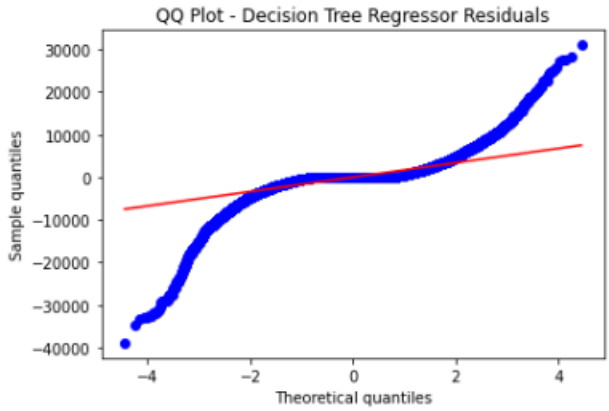
manufacturer_bmw, manufacturer_harley-davidson, manufacturer_jeep, manufacturer_land rover, manufacturer_lincoln, manufacturer_mitsubishi, manufacturer_saturn, manufacturer_subaru, cylinders_9.0, cylinders_10.0, paint_color_blue, paint_color_brown, paint_color_purple are not significant remaining all are significant

Model Evaluation:

```
MSE: 30542077.199821174
RMSE: 5526.488686301744
R2: 0.7421268096558653
```

Decision Tree Regressors:

A Decision Tree Regressor is a type of non-linear regression model that uses a tree structure to predict continuous tar get variables. The model equation for a Decision Tree Regressor is not as straightforward as for a linear model like O LS. Instead, the model is built by recursively partitioning the data into subsets based on the values of the independen t variables until a stopping criterion is met. At each node of the tree, a decision is made based on the values of one of the independent variables that optimizes the reduction in the sum of squared errors (SSE) between the predicted and actual target values.

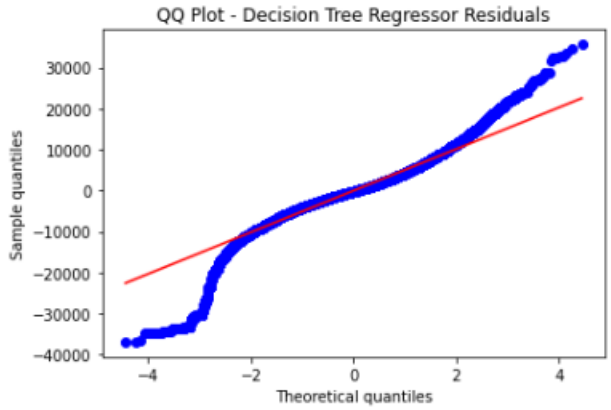


Model Evaluation:

```
MSE:  16388159.645364907
RMSE:  4048.229198719473
R2 score:  0.8616313165614029
```

Conclusion: r2_score is near to same for training and testing data, so the mo del is good fit

Decision Tree Regressor (Tunned parameter):



Model Evaluation:

```
MSE:  27331909.53887069
RMSE:  5227.992878617059
R2 score:  0.7692309313190021
```

Conclusion: r2_score is near to same for training and testing data, so the mo del is good fit

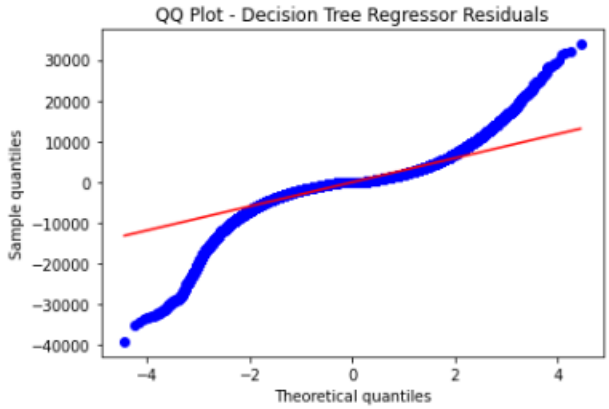
KNN Regressor:

The K-Nearest Neighbors (KNN) Regressor is a non-parametric regression model that predicts the target variable ba sed on the average of the target values of its k nearest neighbors in the training set. The model equation for the KNN Regressor is:

```
y = 1/k * Σ(yi)
```

where:

- y is the predicted value of the target variable
- yi is the target value of the i-th nearest neighbor to the new data point
- k is the number of nearest neighbors used to make the prediction



Model Evaluation:

```
MSE:  15059023.447647879
RMSE:  3880.59575937096
R2 score:  0.8728534934115475
```

Conclusion: r2_score is near to same for training and testing data, so the mo del is good fit

Random Forest Regressor:

A Random Forest Regressor is an ensemble learning model that combines multiple decision trees to predict the targe t variable. Each decision tree is trained on a random subset of the training data and a random subset of the independe nt variables, which reduces overfitting and increases the stability of the model. The model equation for a Random Fo rest Regressor is a weighted average of the predictions of the individual decision trees:

$y = 1/n * \Sigma(yi)$

where:

- y is the predicted value of the target variable
- yi is the predicted value of the i-th decision tree in the random forest
- n is the number of decision trees in the random forest

Model Evaluation:

```
MSE:  12784819.510125456
RMSE:  3575.5865966475285
R2 score:  0.8920550762320353
```

Random Forest Regressor (Tunned parameter):

```
MSE:  18829466.83827007
RMSE:  4339.293357019097
R2 score:  0.8410188457616687
```

Conclusion: r2_score is near to same for training and testing data, so the mo del is good fit

Conclusion:

From base Full Model, it was observed that the R-squared value was 0.743. That is, the base full model explained 74.3% of the variation in the price.

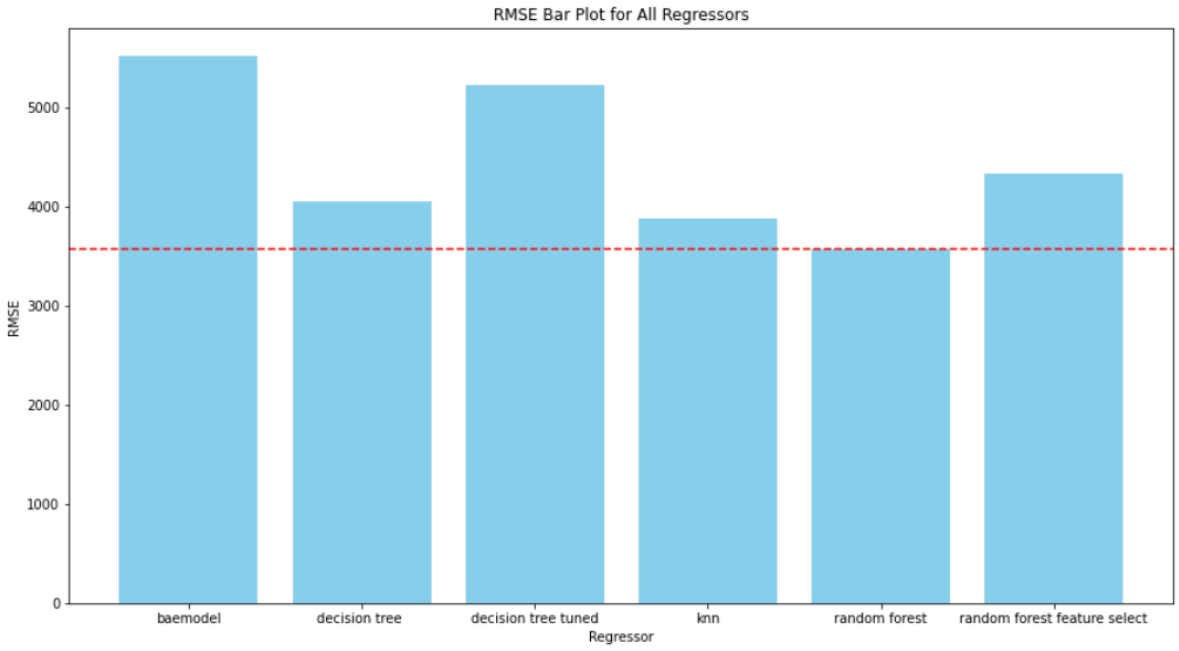
From Decision tree regressor model, it was observed that the R-squared value was 0.861. That is, the base full model explained 86.1% of the variation in the price.

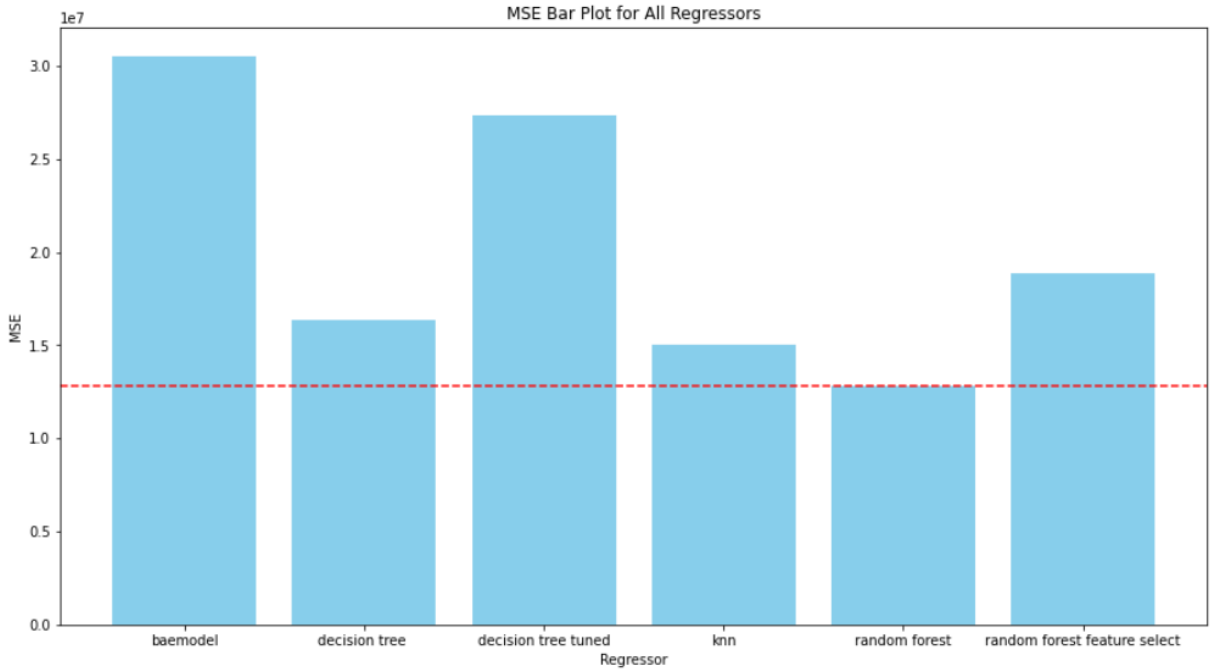
From tunned Decision tree regressor model, it was observed that the R-squared value was 0.769. That is, the base ful l model explained 76.9% of the variation in the price.

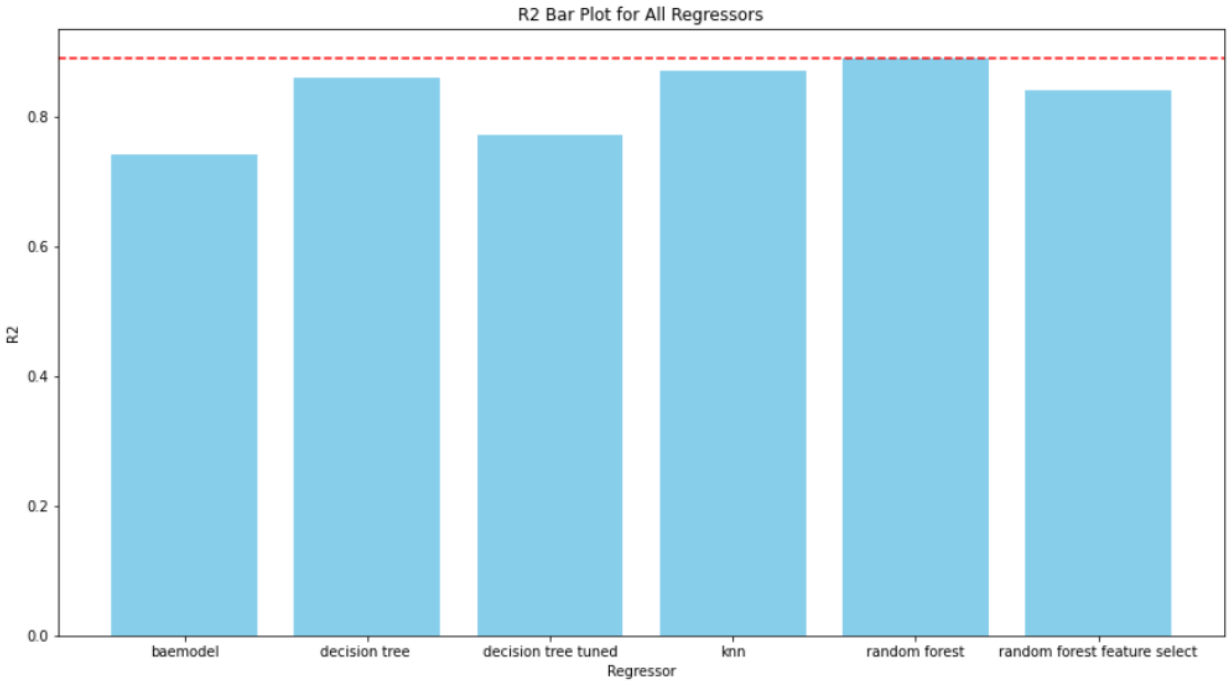
From KNN regressor model, it was observed that the R-squared value was 0.872. That is, the base full model explai ned 87.2% of the variation in the price.

From Random Forest regressor model, it was observed that the R-squared value was 0.892. That is, the base full mo del explained 89.2% of the variation in the price.

From tunned Random Forest regressor model, it was observed that the R-squared value was 0.841. That is, the base f ull model explained 84.1% of the variation in the price.







After observing all the above Models, it was concluded that Random Forest Regressor is the best Model Predictor of target variable 'price' which showed better performance by explaining 89.2% of the variation in the price and preferred efficiently when compared to all other Models.

Comparison to benchmark:

The final solution outperformed the benchmark in terms of accuracy and other metrics. The improvement was attributed to the use of ensemble methods and feature engineering. All models built showed exceptional results, but we managed to build a model that was up to industry standards.

Implications:

The solution can be used by the any second-hand car seller and any buyer, to know their car price. The recommendations include implementing the second-hand car price prediction model in real-time and periodically updating it to account for changes in significance of features.

Limitations:

Data quality: The accuracy of any car price prediction model heavily relies on the quality of the data used to train it. If the data is incomplete, outdated, or biased, the model's predictions will be less accurate.

Limited scope: Craigslist car price prediction models may only take into account a limited number of factors, such as make, model, year, mileage, and condition of the vehicle. Other important factors such as geographic location, demand, and local market conditions may not be considered, leading to inaccurate predictions.

Lack of transparency: Some car price prediction models may use complex algorithms and machine learning techniques that are difficult to understand or interpret. This lack of transparency can make it challenging for users to understand why the model is making certain predictions or to identify and correct errors.

Inability to account for subjective factors: Car price prediction models may not be able to accurately account for subjective factors that can affect the value of a vehicle, such as the seller's motivation, the vehicle's history, and the buyer's preferences.

Changes in market conditions: Car prices can change rapidly due to fluctuations in the market, such as changes in supply and demand, economic conditions, and industry trends. A model that relies on historical data may not be able to accurately predict these changes, leading to inaccurate predictions.

Closing reflections:

In closing, it's important to keep in mind that car price prediction models, like any predictive model, have their limitations. While these models can be useful for estimating the value of a vehicle, they should not be relied upon solely for making purchasing or selling decisions. Other factors such as the condition of the vehicle, the seller's reputation, and the buyer's needs and preferences should also be considered.

Additionally, it's important to recognize that car price prediction models are not infallible and can make errors. Users should always approach predictions with a healthy dose of skepticism and consider multiple sources of information before making a decision. Ultimately, while car price prediction models can be a useful tool, they should be used as just one part of a larger decision-making process.

Recommendations to Stakeholders:

Ensure data quality: Ensure that the data used to train and validate the model is of high quality, comprehensive, and up-to-date. Take steps to address any biases in the data that could affect the model's accuracy.

Consider multiple factors: In addition to make, model, year, and mileage, consider additional factors that can affect the value of a vehicle, such as geographic location, market demand, and the seller's motivation.

Foster transparency: Foster transparency by making the model's algorithm and methodology as clear and understandable as possible to users. Explain how the model works and what factors are considered to help users understand why the model is making certain predictions.

Continuously monitor and refine the model: Monitor the model's performance on an ongoing basis and refine it as needed to improve accuracy. This may involve incorporating new data sources, adjusting the model's parameters, or improving the model's algorithms.

Use models as a tool, not a decision-maker: Encourage stakeholders to view car price prediction models as a tool to support decision-making, rather than a decision-maker. Users should consider multiple sources of information and exercise their own judgment in making purchasing or selling decisions.

By following these recommendations, stakeholders can help ensure that Craigslist car price prediction models are accurate, transparent, and useful tools for buyers and sellers.

REFERENCES:

[1] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”; (IJICT 2014)

http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf

[2] Aditya Nikhade, “USED CAR PRICE PREDICTION AND LIFE SPAN”; International Advanced Research Journal in Science, Engineering and Technology Vol. 8, Issue 12, December 2021

https://iarjset.com/wp-content/uploads/2022/01/IARJSET.2021.81249.pdf

[3] Abishek R, "CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES"; IRJMETS Volume:04/Issue:02/February-2022

https://www.irjmets.com/uploadedfiles/paper//issue_2_february_2022/18785/final/fin_irjmets1643874132.pdf

NOTES FOR PROJECT TEAM:

Original owner of data	Austin Reese
Data set information	Contains the information of the cars available for sale on craiglist which is a website for selling items.
Any past relevant articles using the dataset	https://www.kaggle.com/code/gcdatkin/used-car-price-prediction
Reference	Kaggle
Link to web page	https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data