Used Cars Price Prediction

Abstract

Due to the unprecedented number of cars being purchased and sold, used car price prediction is a topic of high interest. Because of the affordability of used cars in developing countries, people tend more purchase used cars. A primary objective of this project is to estimate used car prices by using attributes that are highly correlated with a label (Price). To accomplish this, data mining technology has been employed. Null, redundant, and missing values were removed from the dataset during pre-processing. In this supervised learning study, three regressors (Random Forest Regressor, Linear Regression, and Bagging Regressor) have been trained, tested, and compared against a benchmark dataset. Among all the experiments, the Random Forest Regressor had the highest score at 95%, followed by 0.025 MSE, 0.0008 MAE, and 0.0378 RMSE respectively. In addition to Random Forest Regression, Bagging Regression performed well with an 88% score, followed by Linear Regression having an 85% mark. A train-test split of 80/20 with 40 random states was used in all experiments. The researchers of this project anticipate that in the near future, the most sophisticated algorithm is used for making predictions, and then the model will be integrated into a mobile app or web page for the general public to use.

AIM

The main aim of this project is to predict the price of a used car based on various features.

The solution is divided into the following sections:

- Data understanding and exploration.
- Data cleaning.
- Data preparation.
- Building model
- Conclusion.

Keywords

Car Price Prediction, supervised learning, linear regression, bagging regression, and classification.

Statement of problem

The research objective of this study is to predict used car prices using data mining techniques, by scraping data from websites that sell used cars and analyzing the different aspects and factors that lead to the actually used car price valuation. To enable consumers to know the actual worth of their car or desired car, by simply providing the program with a set of attributes from the desired car to predict the car price. The purpose of this study is to understand and evaluate used car prices and to develop a strategy that utilizes data mining techniques to predict used car prices.

Project goals

This project aims to deliver price prediction models to the public, to help guide individuals looking to buy or sell cars, and to give them a better insight into the automotive sector. Buying a used car from a dealer can be a frustrating and unsatisfying experience as some dealers are known to deploy deceitful sale tactics to close a deal. Therefore, to help consumers avoid falling victim to such tactics, this study hopes to equip consumers with the right tools to guide them in their shopping experience. Another goal of the project is to explore new methods to evaluate used car prices and compare their accuracies. Considering this is an interesting research topic in the research community, and in continuing their footsteps, we hope to achieve significant results using more advanced methods of previous work.

Methodology

The project deals with used cars. The project's methodology is as follows: Proposed Methodology

After data collection, the dataset was pre-processed to remove samples that have missing values, remove the non-numerical part from numerical attributes, convert categorical values into numerical (if needed), fix any discrepancies in the units, as well as removing attributes that don’t affect the price evaluations if needed to reduce the complexity of the model. Data Understanding and preparation are essential part of building a model as it gives insight into the data and what corrections or modifications shall be done before designing and executing the model, preliminary analysis of the data must be done to have a deeper understanding of the quality of the data, in terms of outliers and the skewness of the figures, descriptive Statistics of categorical and numerical variables was done for that to be achieved. As well as the ability to understand the main attributes that affect the results of the price. That was done through a correlation matrix for every attribute to understand the relations between the different factors.

- NULL cells conventions
- Missing values
- Encoding
- Normalization Pre-processing
- Logistic Regression
- Random Forest Regressor
- Accuracy
- MSE
- MAE
- RMSE

Afterward when the data is organized and transformed into a form that could be processed by the data mining technique. Different data mining models were designed to predict the prices and values of used cars. In this study, three models are proposed to be built using the Logistic Regression model technique, Random Forest Regressor, and Bagging Regressor. Firstly, the data was portioned into sections for training and another part for testing, portioning percentages can be tested with different ratios to analyze different results. All three models were evaluated on four evaluation matrices known as model score, Mean Square Error (MSE), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). From all, the Random Forest Regressor outperformed.

Machine learning

The goal of machine learning (ML) is to help a computer learn without being explicitly instructed to do so by means of mathematical models of data. Artificial intelligence (AI) is a subset of machine learning. Data is analyzed using algorithms to identify patterns, which are then used to create predictive models. Like humans, machine learning becomes more accurate with more data and experience. With machine learning, you can adapt to situations where data is constantly changing, the nature of the request or task is shifting, or coding a solution isn't feasible.

Machine Learning Categories Supervised and unsupervised learning are commonly used types while reinforcement is a sequential decision-maker technique. The main categories of supervised and unsupervised machine learning are:

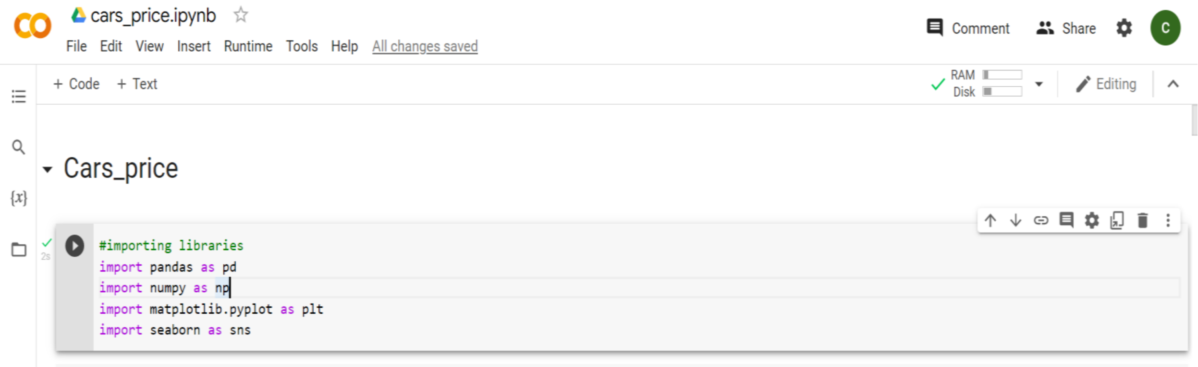
- Machine learning Supervised Learning
- Unsupervised Learning
- Semi-supervised Learning
- Reinforcement Learning

Supervised Learning Working and details of some famous supervised machine learning algorithms which are used in this project are Logistic Regression: Whenever the dependent variable is non-numerical (categorical) and the class should be predicted, not classified, the logistic regression algorithm needs to be abandoned. Machine learning technique Logistic Regression is commonly used to classify binary data. The function of Logistic Regression is to optimize results based on various datasets. To predict results, the default label class is always employed, but the results and probability are always calculated after all categorical values have been converted into numerical values and all data has been normalized. Logistic regression, also referred to as sigmoid regression, was designed by statisticians to explain the properties of the population increasing in the ecological study, growing fast, and maxing out on the capability to wear out the surroundings. With an S-shaped curve, any real-valued range can be mapped right into a number between zero and 1, but not precisely at the limit of 1. 1 1 + e $-x$ Equation 1 Sigmoid equation Where e is the base of the logarithms (Euler's wide variety or the EXP() characteristic on your spreadsheet) and price is the real numerical price which needs to be transformed. While the equation of regression in which intercept and slope are integrated is as follows: $y = mx + c$ Equation 2 Regression equation Below is a generalized equation for Multivariate regression model: y = β0 + β1.x1 + β2.x2 +….. + βn.xn There are few steps involved in generating the regression beginning with feature selection, normalizing features, select loss function and hypothesis, set hypothesis parameters, and minimizing the loss function, and finally testing the function of the data. Random Forest Regressor: Random Forest is already revealing that it creates a forest and then somehow randomizes it. It builds the forest through the ensemble of Decision Trees and most of the time trains it using a method called the Bagging Method. Since it uses the ensemble method, the result is improved. Decision tree and bagging classifier hypermeters are the same. Each feature in the tree can be made random simply by adding thresholds.

Dataset description

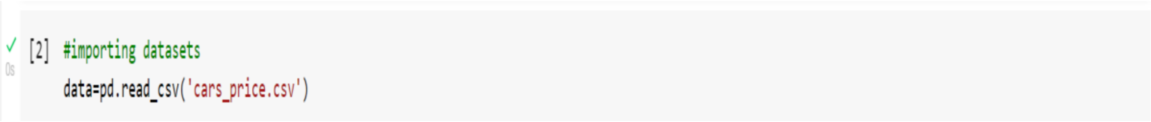
- symboling
- normalized-losses
- make
- fuel-type
- aspiration
- num-of-doors
- body-style
- drive-wheels
- engine-location
- wheel-base
- length
- width
- height
- curb-weight
- engine-type
- num-of-cylinders
- engine-size
- fuel-system
- bore
- stroke
- compression-ratio
- horsepower
- peak-rpm
- city-mpg
- highway-mpg

Importing Libraries



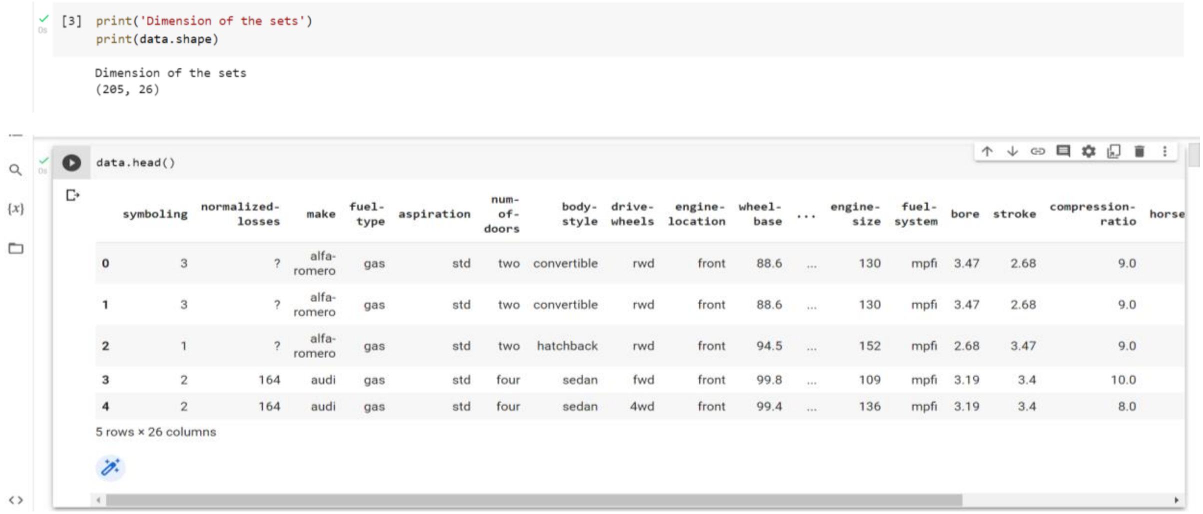
```
#importing libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

Loading Datasets



```
#importing datasets
data=pd.read_csv('cars_price.csv')
```

Dimensions of the sets



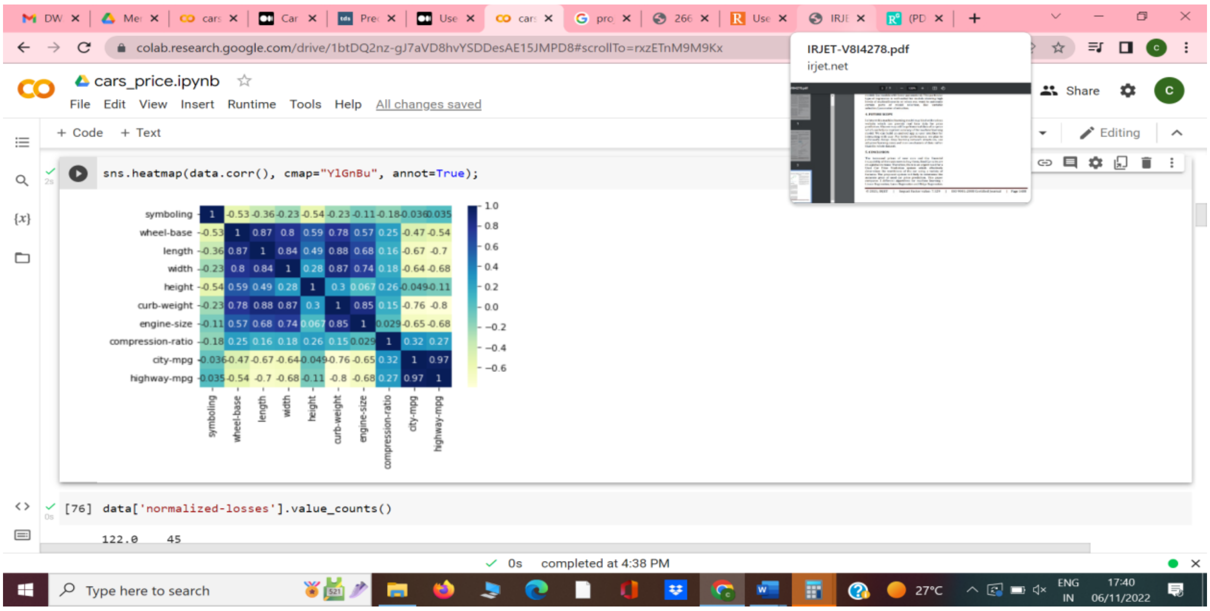
```
[3] print('Dimension of the sets')
    print(data.shape)

Dimension of the sets
(205, 26)
```

```
data.head()
```

	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	drive-wheels	engine-location	wheel-base	...	engine-size	fuel-system	bore	stroke	compression-ratio	horse
0	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	
1	3	?	alfa-romero	gas	std	two	convertible	rwd	front	88.6	...	130	mpfi	3.47	2.68	9.0	
2	1	?	alfa-romero	gas	std	two	hatchback	rwd	front	94.5	...	152	mpfi	2.68	3.47	9.0	
3	2	164	audi	gas	std	four	sedan	fwd	front	99.8	...	109	mpfi	3.19	3.4	10.0	
4	2	164	audi	gas	std	four	sedan	4wd	front	99.4	...	136	mpfi	3.19	3.4	8.0	

5 rows × 26 columns

Data Correlation

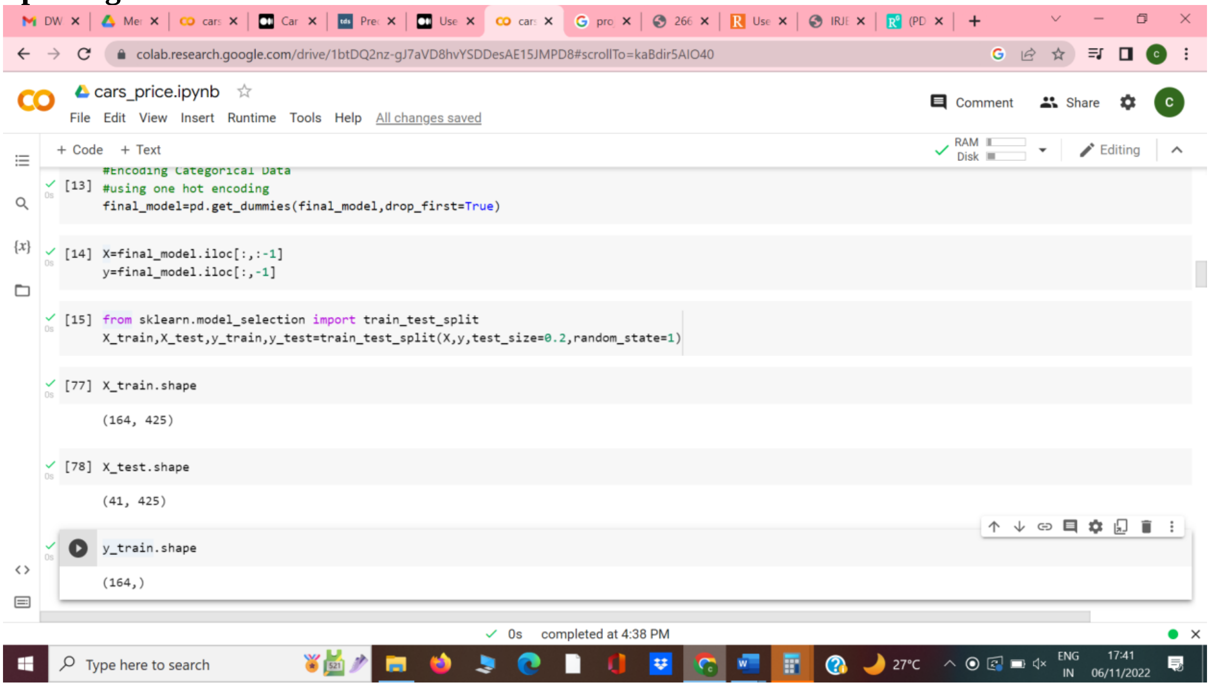


```
sns.heatmap(data.corr(), cmap="YlGnBu", annot=True);
```

	symboling	wheel-base	length	width	height	curb-weight	engine-size	compression-ratio	city-mpg	highway-mpg
symboling	1	-0.53	-0.36	-0.23	-0.54	-0.23	-0.11	-0.18	-0.036	0.035
wheel-base	-0.53	1	0.87	0.8	0.59	0.78	0.57	0.25	-0.47	-0.54
length	-0.36	0.87	1	0.84	0.49	0.88	0.68	0.16	-0.67	-0.7
width	-0.23	0.8	0.84	1	0.28	0.87	0.74	0.18	-0.64	-0.68
height	-0.54	0.59	0.49	0.28	1	0.3	0.067	0.26	-0.049	-0.11
curb-weight	-0.23	0.78	0.88	0.87	0.3	1	0.85	0.15	-0.76	-0.8
engine-size	-0.11	0.57	0.68	0.74	0.067	0.85	1	0.029	-0.65	-0.68
compression-ratio	-0.18	0.25	0.16	0.18	0.26	0.15	0.029	1	0.32	0.27
city-mpg	-0.036	-0.47	-0.67	-0.64	-0.049	-0.76	-0.65	0.32	1	0.97
highway-mpg	0.035	-0.54	-0.7	-0.68	-0.11	-0.8	-0.68	0.27	0.97	1

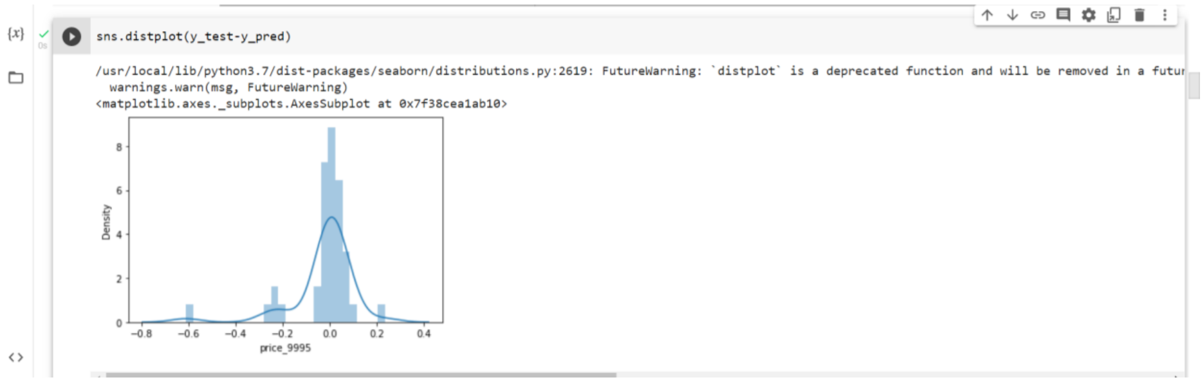
```
[76] data['normalized-losses'].value_counts()

122.0    45
```

Splitting Data into train and test


colab.research.google.com/drive/1btDQ2nz-gJ7aVD8hvYSDDesAE15JMPD8#scrollTo=kaBdirSAIO40
cars_price.ipynb
File Edit View Insert Runtime Tools Help All changes saved
Comment Share
+ Code + Text
RAM Disk
Editing
0s completed at 4:38 PM
Type here to search
27°C
ENG IN
17:41 06/11/2022

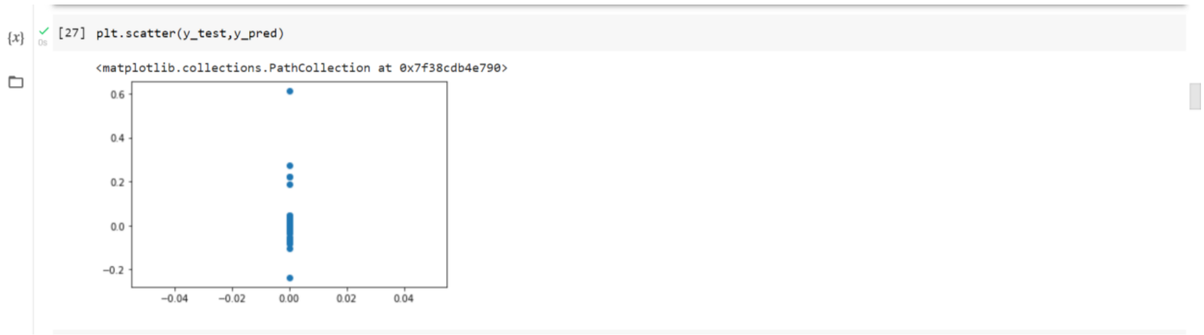
```
#Encoding Categorical Data
[13] #using one hot encoding
     final_model=pd.get_dummies(final_model,drop_first=True)

[14] X=final_model.iloc[:,:-1]
     y=final_model.iloc[:,-1]

[15] from sklearn.model_selection import train_test_split
     X_train,X_test,y_train,y_test=train_test_split(X,y,test_size=0.2,random_state=1)

[77] X_train.shape

     (164, 425)

[78] X_test.shape

     (41, 425)

     y_train.shape

     (164,)
```

Plotting



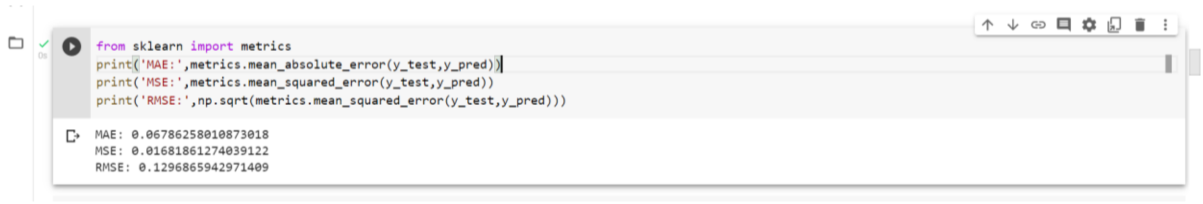
```
sns.distplot(y_test-y_pred)

/usr/local/lib/python3.7/dist-packages/seaborn/distributions.py:2619: FutureWarning: `distplot` is a deprecated function and will be removed in a futur
  warnings.warn(msg, FutureWarning)
<matplotlib.axes._subplots.AxesSubplot at 0x7f38cea1ab10>
```

Scatter plot



```
[27] plt.scatter(y_test,y_pred)

     <matplotlib.collections.PathCollection at 0x7f38cdb4e790>
```

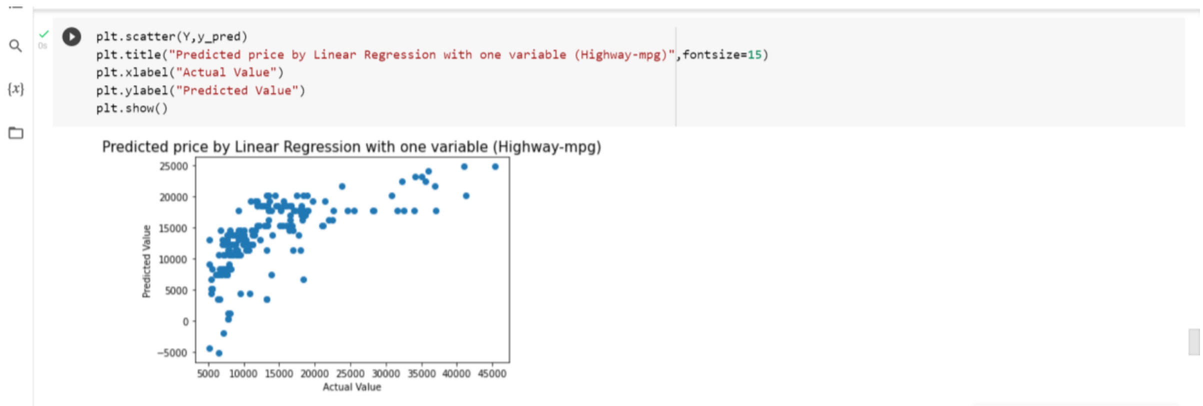
MSE, MAE, R2 Score



```
from sklearn import metrics
print('MAE:',metrics.mean_absolute_error(y_test,y_pred))
print('MSE:',metrics.mean_squared_error(y_test,y_pred))
print('RMSE:',np.sqrt(metrics.mean_squared_error(y_test,y_pred)))
```

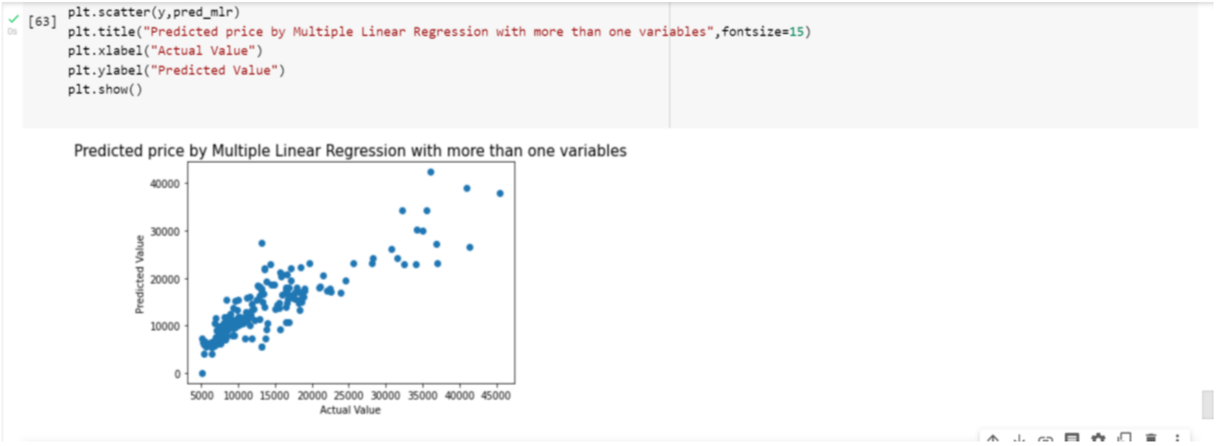
```
MAE: 0.06786258010873018
MSE: 0.01681861274039122
RMSE: 0.1296865942971409
```

Predicted Price by Linear Regression



```
plt.scatter(Y,y_pred)
plt.title("Predicted price by Linear Regression with one variable (Highway-mpg)",fontsize=15)
plt.xlabel("Actual Value")
plt.ylabel("Predicted Value")
plt.show()
```

Predicted Value by Multiple Linear Regression



```
[63] plt.scatter(y,pred_mlr)
     plt.title("Predicted price by Multiple Linear Regression with more than one variables",fontsize=15)
     plt.xlabel("Actual Value")
     plt.ylabel("Predicted Value")
     plt.show()
```

Conclusion

Using data mining and machine learning approaches, this project proposed a used car price prediction. An efficient machine learning model is built by training, testing, and evaluating three machine learning regressors named Random Forest Regressor, Linear Regression, and Bagging Regressor. As a result of pre-processing and transformation, Random Forest Regressor came out on top with 95% accuracy.