Used Car Price Prediction

Submitted By,
Raj Sharma

Acknowledgement:

I wish to express my sincere thanks to the following companies, without whom I would have not got opportunity to work on this project; Data Trained Institute and Flip Robo Technology.

Introduction:

Business Problem Facing:

With the covid 19 impact in the market, we have seen lot of changes in the car market. Now some cars are in demand hence making them costly and some are not in demand hence cheaper. One of our clients works with small traders, who sell used cars. With the change in market due to covid 19 impact, our client is facing problems with their previous car price valuation machine learning models. So, they are looking for new machine learning models from new data. We have to make car price valuation model.

Based on the business requirements of the Client, data is scrapped from the well-known e-commerce websites such as cars 24, OLX and cardekho.com. Based on the Data collected, we will be predicting the prices of used cars. We will be building various Machine Learning models. In the end, we will see how all the machine learning models performs. And based on which we will sort the best machine learning model and hyperparameter tune the same to get the improved performance.

Concept Background of the Domain Problem:

To understand the business problem, there are certain factors that will influence the automotive industries in the future. Some of them include digital technologies, changing customer preferences, electrical vehicles, intelligent ability, and technical advancements. Technologies such as artificial intelligence, machine learning, cloud computing, and internet of things will also play an important role in developing new business models. Apart from that, they enable customers to ensure a better mobility experience. In other words, technologies may impact automotive industry units significantly that will change the markets. The introduction of electrical cars and hybrid vehicles may transform the automobile industries in coming years.

Review of Literature:

As per the requirement of our client, data is scraped from different used cars selling merchants websites, and so based on the data collected I have tried analysing based on what factors the used car price is decided? What is the relationship between cost of the used cars and other factors like Fuel type, Brand and Model, year the car is purchased and Number of owners before selling? And so based on all the above consideration I have developed a model that will predict the price of the used cars.

Motivation for Problem Undertaken:

This problem is taken based on the requirement of the client and also, with a curiosity to know how the used cars markets are at the time of pandemic.

Analytical Problem Framing:

Mathematical/Analytical Modelling of the Problem:

Web Scraping using selenium is used to scrap the data from carsdekho website. Different types cars are scrapped from different locations. Scraped data contains 7436 rows and 7 columns which is stored in excel file.

First, the analysis is started with importing the data, this dataset contains no Null values.

Once data is cleaned outliers and skewness are checked, if present they are removed, then Data Pre-processing, Standard Scaler is used to standardize the data and VIF is checked for multicollinearity in dataset.

Once this is all done then data is ready for modelling. As the target variable contains continuous data, Regression is used. 4 regressors – Decision Tree Regressor, Support Vector Regressor, KNeighbors Regressor, Linear Regression, 4 ensemble -Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor,3 metrics – r2_score, mean_squared_error, mean_absolute_error and 3 regularization – Lasso, Ridge, ElasticNet techinques are used in this project to build Regression model. From this the best model is identified and using Cross Validation technique is checked for overfitting and underfitting and Hypertuning is done to increase accuracy. Then, Finally the best model is saved.

Data Source and their Format:

The data is scrapped using selenium webscraping which is stored in xlsx format.

Data contains 7436 entries each having 7 variables.

Here the dataset is divided into independent and target variable.

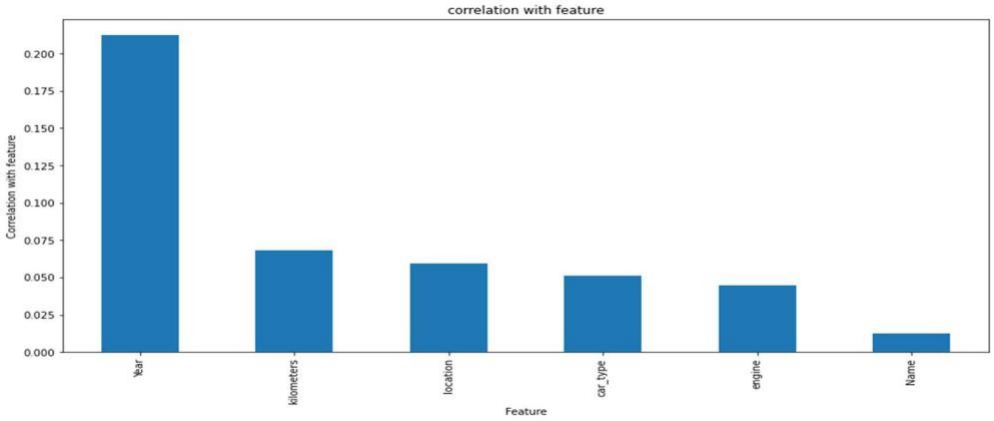
Data Pre-Processing:

In Machine Learning, data pre-processing refers to the process of cleaning and organising raw data in order to make it appropriate for creating and training Machine Learning models. In other words, anytime data is received from various sources, it is collected in raw format, which makes analysis impossible. Data pre-processing is a crucial stage in Machine Learning since the quality of data and the relevant information that can be gleaned from it has a direct impact on our model's capacity to learn; consequently, we must pre-process our data before feeding it into our model. As a result, it is the first and most important stage in developing a machine learning model.

Some of the techniques used in this project are listed below:

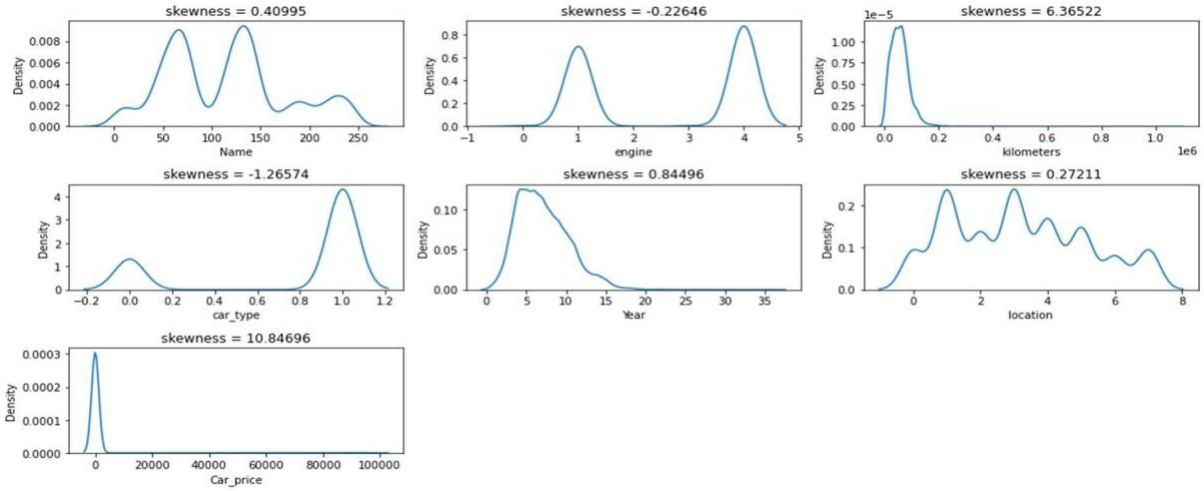
1. **Started with web scraping using selenium from carsdekho website. Different types of cars is scrapped using different locations. Then the scraped data is stored in xlsx format.**
2. **Then the required libraries are imported and then the dataset which is in xlsx format.**
3. **Dataset contains 7436 rows and 7 variables. Then the information of the columns is observed carefully.**
4. **Few columns like 'Unnamed: 0' is dropped as it does not contribute much in model building.**

5. **Then Null values is checked, as there are no null values, procedded with further analysis process.**
6. **Numeric data and Categorical data are then separated so that visualization is done with distplot for numeric data and countplot for categorical data.**
7. **Then all the categorical data is converted to numeric by using LabelEncoder so that analysis can be made in better way.**
8. **Data Description is made followed by correlation where positive and negative correlated data is checked.**

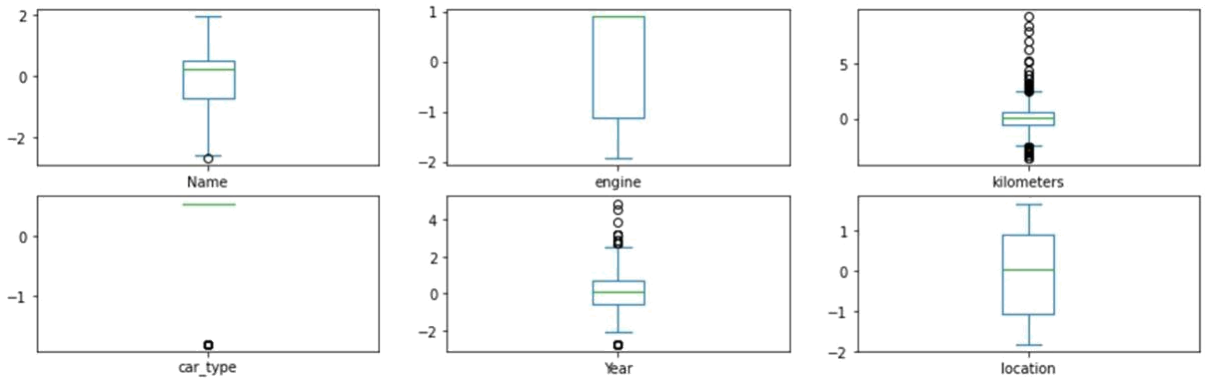


'Year' is highly correlated with Price which means if the age of car is less the price of car selling is also increases.
'Kilometers' driven is less then price also increases. Which is also positively correlated with price column.
'Location' also contribute highly in deciding the price od used car, followed by 'car_type', 'engine', 'name'

9. **Outliers and Skewness is checked and removed in order to avoid bias while model building.**



Skewness is removed by using power_transform method.



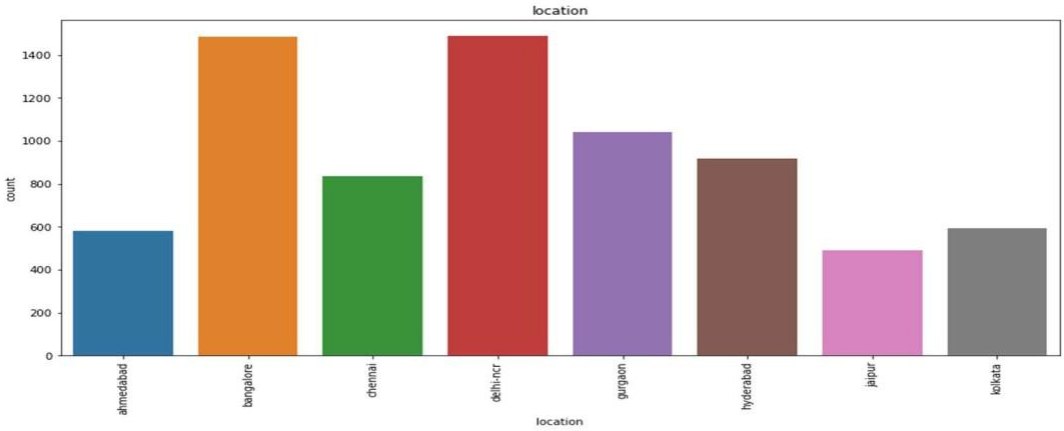
Outliers are removed by using zscore method.

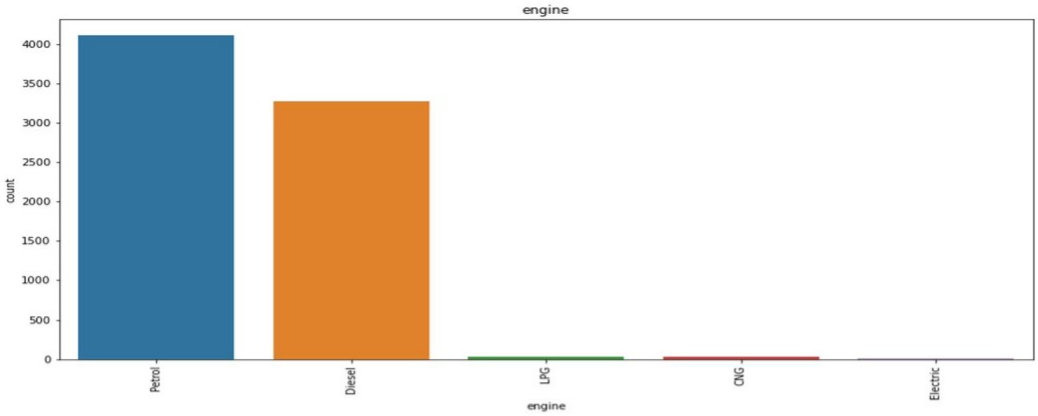
10. **Standard Scaler is used for data standardization and VIF is used to check Multicolinearity is checked to find if any data variable is correlated with each other and it is removed.**
11. **Once all these processes are done then data is ready for model building where various Machine Learning models is used to check the accuracy of data.**

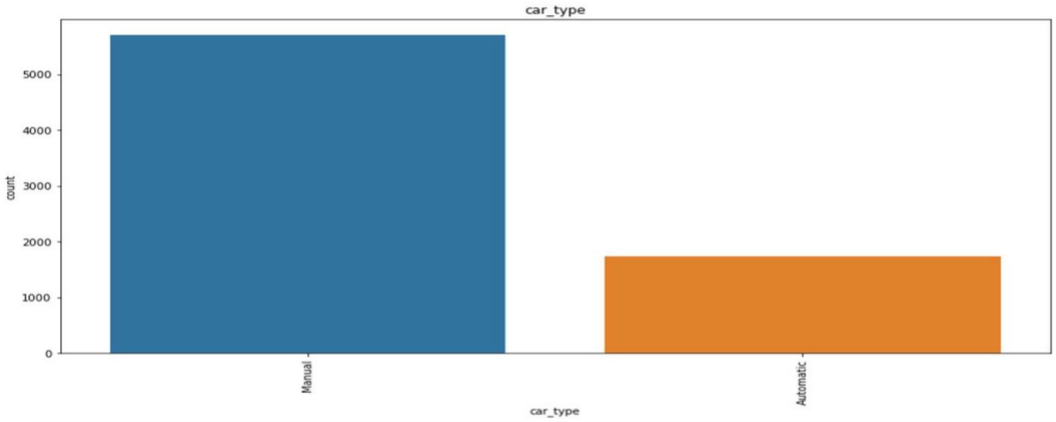
Data Input Logic Output Relations:

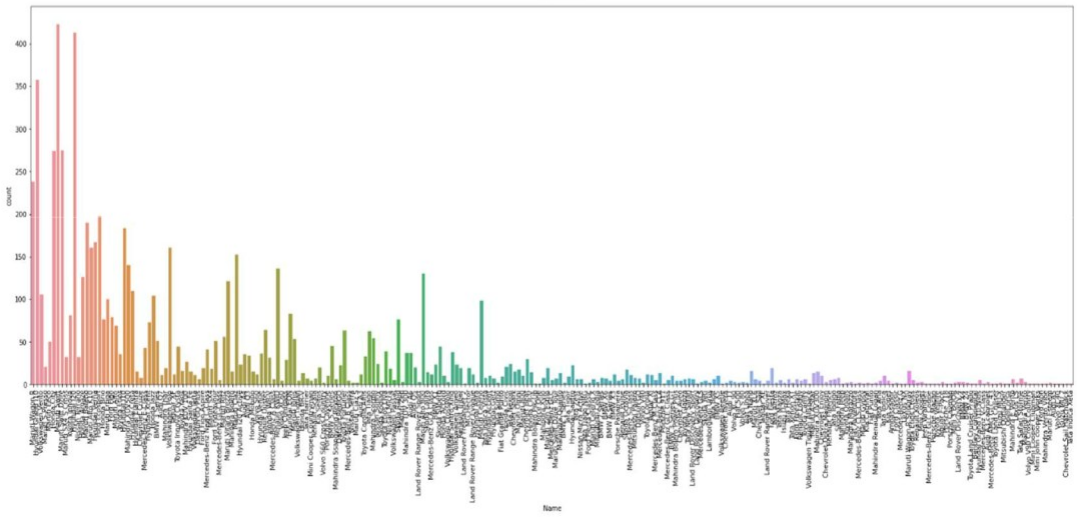
Data visualization is used to find the relation between input and the output variable.

Data Visualization:
Visualization of Categorical Variables:





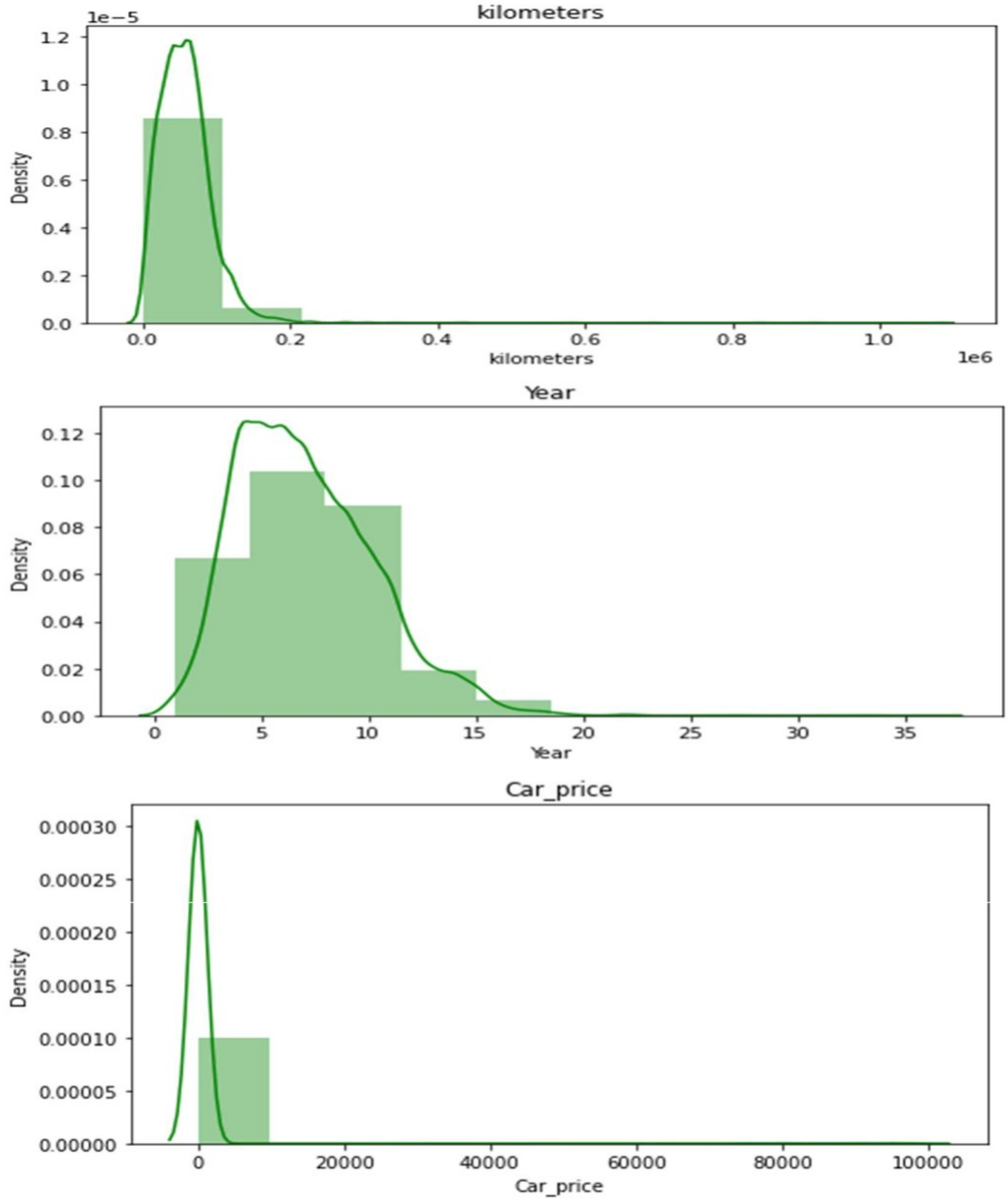




Key Observations:

1. **Petrol engine cars are more sold for high price compared to other engine type, followed by diesel engine cars and the lowest running type is electric.**
2. **Manual cars are sold in large amount compared to automatic cars.**
3. **In our dataset, Bangalore and Delhi-NCR locations sees hig sell of used cars followed by Gurgaon, Jaipur sees less amount of used cars selling.**
4. **Maruti swift followed by Hyundai i20 are sold more followed by Maruti WagonR. Least sold brands are Minicoper, Cheverolet**

Visualization of Numerical columns:



Key Observations:

Here we can see ‘Kilometers’, ‘Year’, ‘Car_price’ columns has skewness in it.

From the visualization above we can clearly understand that the used car price factors are decided by the factors such as brand, location, model, year made, number of owners used the car before, fuel type of the car. From that we can clearly say that the used car price depending on the Brand that is the manufacturer and model it varies. The manufacturer like Land Rover, Benz, BMW cars are costliest used car in the market comparatively to other cars, the low kilometres driven and also if the manufacturing year is lesser on these brands those card sells in much higher rates or closest to the buying new car rates. The Diesel variant and Automatic shift variants are also costliest user car variants in the used car market

Tools Used:

1. Python 3.8
2. Numpy
3. Pandas
4. Matplotlib
5. Seaborn
6. Data Science
7. Machine Learning

Model/s Development and Evaluation:

Identification of possible problem solving approach:

In this project, both Statistical and Analytical methods are used, in which Data Pre-processing, Exploratory Data Analysis is used after ensuring that data is cleaned.Here Target column is Sale Price, as it is continuous data Regression algorithm is used. This project consists of 4 regressors – Decision Tree Regressor, Support Vector Regressor, KNeighbors Regressor, Linear Regression, 4 ensemble -Random Forest Regressor, AdaBoost Regressor, Gradient Boosting Regressor, 3 metrics – r2_score, mean_squared_error, mean_absolute_error and 3 regularization – Lasso, Ridge, ElasticNet techniques. Out of which Gradient Boosting Regressor gave high accuracy which is also chosen as best model in which there is least difference between r2 score and cv. Then with this Hypertuning is done in which accuracy is slightly reduced in order to correct the over fitting of the model. Once all this is done, best model is saved using joblib.Then by using this saved model the output is determined which predicted the Used car price CarPrice.

Test of Identified Approaches:

The approaches followed in this project are

1. Find the best random state of a model.
2. Use all the other models to find accuracy score, mean squared error, mean absolute error and r2 score.
3. Then find Cross Validation of all models to find the best accuracy, which is the least difference between r2 score and cv score.
4. With the best model accuracy, we need to do hyper tuning using GridSearchCV.
5. Then finally saving the best model and by keeping this we need to predict the CarPrice of test dataset.

Run and evaluate of selected model:

```
score = []
mean_squared_err = []
mean_absolute_err = []
r2 = []

for m in model:
    m.fit(x_train,y_train)
    m.score(x_test,y_test)
    predm = m.predict(x_test)
    print("Accuracy Score of ",m," is ",m.score(x_train,y_train))
    score.append(m.score(x_train,y_train))

    print("Mean Squared Error is ",mean_squared_error(y_test,predm))
    mean_squared_err.append(mean_squared_error(y_test,predm))
    print("Mean Absolute Error is ",mean_absolute_error(y_test,predm))
    mean_absolute_err.append(mean_absolute_error(y_test,predm))
    print("R2 Score is ",r2_score(y_test,predm))
    r2.append(r2_score(y_test,predm))
    print("\n\n")
```

```
Accuracy Score of  LinearRegression()  is  0.038776559410836464
Mean Squared error :  31640962.31956184
Mean Absolute Error :  1800.7541801487782
R2 Score :  0.019010985633283628
```

Linear Regression gives 3.8% accuracy and r2_score 1.90

```
Accuracy Score of  Lasso()  is  0.03877644758881993
Mean Squared error :  31639031.283813372
Mean Absolute Error :  1799.6709257083407
R2 Score :  0.019070855015145716
```

Lasso gives 3.8% accuracy and r2_score 1.90

```
Accuracy Score of  Ridge()  is  0.03877655699433524
Mean Squared error :  31640609.095377423
Mean Absolute Error :  1800.5388196286117
R2 Score :  0.019021936913494963
```

Ridge give 3.8% accuracy score and r2_score 1.9

```
Accuracy Score of  ElasticNet()  is  0.034212281875037664
Mean Squared error :  31401820.925836664
Mean Absolute Error :  1508.3772847202868
R2 Score :  0.026425269616034197
```

ElasticNet gives 3.4% accuracy score and r2_score 2.6

```
Accuracy Score of  KNeighborsRegressor()  is  0.4283337529319927
Mean Squared error :  30769948.381106973
Mean Absolute Error :  729.3346122814881
R2 Score :  0.04601569858590604
```

KNeighborsRegressor gives 4.2% accuracy and r2 score 4.6

```
Accuracy Score of  DecisionTreeRegressor()  is  0.9999999999413345
Mean Squared error :  31089496.242412414
Mean Absolute Error :  443.44483191393994
R2 Score :  0.03610851123998848
```

DecisionTreeRegressor gives 99.9% accuracy

```
Accuracy Score of  DecisionTreeRegressor()  is  0.9999999999413345
Mean Squared error :  37442408.8310265
Mean Absolute Error :  507.67449499477067
R2 Score :  -0.16085570861275134
```

SVR gives 1.07% accuracy and 0.602 r2_score

```
Accuracy Score of  SVR()  is  -0.010171619543121846
Mean Squared error :  32448425.46284205
Mean Absolute Error :  438.5181520481758
R2 Score :  -0.006023413291245561
```

RandomForestRegressor gives 89% accuracy and r2_score 0.36

```
Accuracy Score of  RandomForestRegressor()  is  0.896032926540629
Mean Squared error :  20496266.710772656
Mean Absolute Error :  723.4724589713408
R2 Score :  0.3645385283883287
```

AdaBoostRegressor gives 15% accuracy score and r2_score 0.18

```
Accuracy Score of  AdaBoostRegressor()  is  0.157265131039938
Mean Squared error :  26217450.581283838
Mean Absolute Error :  966.5709760079675
R2 Score :  0.18716027833827653
```

Gradient Boosting Regressor gives 54% accuracy and 0.088 r2_score

```
Accuracy Score of  GradientBoostingRegressor()  is  0.5474009726897979
Mean Squared error :  29392571.306489266
Mean Absolute Error :  1084.4580450965389
R2 Score :  0.08871957608476355
```

Cross Validation:

```
Score of  LinearRegression()  is  [-0.0385512   0.02383937 -0.039
40274  0.00828827  0.02062571]
Mean score is :  -0.005040118403806582
Standard Deviation is :  0.02819286698683195
```

CV Score of LinearRegressor is 0.00504

```
Score of  Lasso()  is  [-0.03825487  0.02388898 -0.03930388  0.00
833286  0.02053204]
Mean score is :  -0.004960974744678093
Standard Deviation is :  0.028095757549029595
```

CV Score of Lasso is 0.0049

```
Score of  Ridge()  is  [-0.03850965  0.02384009 -0.03938193  0.00
829423  0.02061962]
Mean score is :  -0.005027528393300807
Standard Deviation is :  0.028177516840745095
```

CV Score of Ridge is 0.00502

```
Score of  ElasticNet()  is  [ 0.0079623   0.02366574 -0.01200191
0.01481686  0.00947543]
Mean score is :  0.008783684509380829
Standard Deviation is :  0.011756438352992742
```

CV Score of ElasticNet is 0.00878

```
Score of  KNeighborsRegressor()  is  [-0.00321379 -0.06356115 -0.
04189242  0.1353953   0.15783778]
Mean score is :  0.03691314518556772
Standard Deviation is :  0.09190981834504193
```

CV Score of KNeighborsRegressor is 0.036

```
Score of  DecisionTreeRegressor()  is  [-0.298922   -1.68586933 -1.29714581  0.00563464 -0.05964471]
Mean score is :  -0.667189442089905
Standard Deviation is :  0.6916614055900185
```

CV Score of DecisionTreeRegressor

```
Score of  SVR()  is  [-0.00513791 -0.00736611 -0.00281011 -0.0053
7089 -0.02410116]
Mean score is :  -0.00895723571509901
Standard Deviation is :  0.007708518977628053
```

CV Score of SVR is 0.00895

```
Score of  RandomForestRegressor()  is  [ 0.06838574 -0.35618378  0.20252381  0.25952524  0.11797151]
Mean score is :  0.058444503502894096
Standard Deviation is :  0.21760110833589905
```

CV Score of RandomForestRegressor is 0.0254

```
Score of  AdaBoostRegressor()  is  [ 0.10338739 -1.36437299 -2.10850446  0.14533775  0.01674925]
Mean score is :  -0.6414806134692295
Standard Deviation is :  0.9254091397796964
```

CV Score of AdaBoostRegressor is 0.4250

```
Score of  GradientBoostingRegressor()  is  [ 0.03432907 -0.21422802  0.14093764  0.10925434  0.15016937]
Mean score is :  0.0440924791418319
Standard Deviation is :  0.1354268544710798
```

```
Score of  ExtraTreesRegressor()  is  [ 0.02224922 -0.1520503  -1.87469563  0.40781993  0.10976555]
Mean score is :  -0.297382244022634
Standard Deviation is :  0.8092320555052255
```

CV Score of GradientBoostingRegressor is
CV Score of ExtraTreeRegressor as above
By this we can conclude that DecisionTreeRegressor is having accuracy score of 83%. which can be sent for Hypertuning.

HyperTuning:

```
from sklearn.model_selection import GridSearchCV
```

```
dt = DecisionTreeRegressor(random_state=42)
```

```
# Create the parameter grid based on the results of random search
params = {
    'max_depth': [2, 3, 5, 10, 20],
    'min_samples_leaf': [5, 10, 20, 50, 100],

}
```

```
grid_search = GridSearchCV(estimator=dt,
                           param_grid=params,
                           cv=7, verbose=1)
```

```
grid_search.fit(X,y)
```

```
Fitting 7 folds for each of 25 candidates, totalling 175 fits
```

```
GridSearchCV(cv=7, estimator=DecisionTreeRegressor(random_state=42),
             param_grid={'max_depth': [2, 3, 5, 10, 20],
                         'min_samples_leaf': [5, 10, 20, 50, 100]},
             verbose=1)
```

In a Jupyter environment, please rerun this cell to show the HTML representation or trust the notebook.
On GitHub, the HTML representation is unable to render, please try loading this page with nbviewer.org.

```
pred = grid_search.best_estimator_
```

After Hypertuning we have 91% accuracy score which is improved.

Saving the best model:

The best model is saved using pickle.

Saving best model

```
import pickle
file = open('used_car_price_prediction.pkl','wb')
pickle.dump(pred,file)
```

Key Metrics for success in solving problem under consideration:

Some of key metrics for the evaluation of this project are

1. **R2_score**
2. **Mean_squared_error**
3. **Mean_absolute_error**
4. **Cross_validation**
5. **Hypertuning**

R2 Score:

R-squared is a statistical measure that represents the goodness of fit of a regression model. The ideal value for r-square is 1. Closer the r2_square value to 1 better the model fits.

Mean Squared Error:

The average of the squares of the errors — that is, the average squared difference between the estimated values and what is estimated is measured by the MSE of an estimator (of a process for estimating an unobserved variable). MSE is a risk function that represents the squared error loss's anticipated value. The Root Mean Squared Error is abbreviated as RMSE.

Mean Absolute Error:

MAE is a statistic that assesses the average magnitude of mistakes in a set of forecasts without taking into account their direction. It's the average of the absolute differences between forecast and actual observation over the test sample, where all individual deviations are given equal weight.

Cross Validation:

Cross-validation aids in determining the model's overfitting and underfitting. The model is constructed to run on several subsets of the dataset in cross validation, resulting in numerous measurements of the model. If we fold the data five times, it will be separated into five parts, each representing 20% of the whole dataset. During the Cross-validation, the first part (20%) of the 5 parts will be left out as a holdout set for validation, while the rest of the data will be utilised for training. We'll acquire the initial estimate of the dataset's model quality this way.

Further rounds are produced in the same way for the second 20% of the dataset, which is kept as a holdout set while the remaining four portions are utilised for training data during the process. We'll acquire the second estimate of the dataset's model

quality this way. During the cross-validation procedure, these stages are repeated to obtain the remaining estimate of model quality.

Hypertuning:

Hyperparameters in Machine learning are those parameters that are explicitly defined by the user to control the learning process. These hyperparameters are used to improve the learning of the model, and their values are set before starting the learning process of the model.

Interpretation of the Results:

From the visualization above we can clearly understand that the used car price factors are decided by the factors such as brand, location, model, year made, number of owners used the car before, fuel type of the car. From that we can clearly say that the used car price depending on the Brand that is the manufacturer and model it varies. The manufacturer like Land Rover, Benz, BMW cars are costliest used car in the market comparatively to other cars, the low kilometres driven and also if the manufacturing year is lesser on these brands those card sells in much higher rates or closest to the buying new car rates. The Diesel variant and Automatic shift variants are also costliest user car variants in the used car market.

CONCLUSION:

Key Findings:

The manufacturer like Land Rover, Benz, BMW cars are costliest used car in the market comparatively to other cars, the low kilometres driven and also if the manufacturing year is lesser on these brands those card sells in much higher rates or closest to the buying new car rates. The Diesel variant and Automatic shift variants are also costliest user car variants in the used car market.

Learning outcomes of study in Data Science:

The above research will help our client to study about the latest used car market and with the help of the model built he can easily predict the price ranges of the cars, and also will helps him to understand based on what factors the Car Price is decided.

Limitations of this work and scope of Future work:

The limitation of the study is that in the volatile changing market we have taken the data, to be more precise we have taken the data at the time of pandemic, so when the pandemic ends the market correction might happen slowly. So based on that again the deciding factors of the used car prize might change and we have shortlisted and taken these data from the important cities across India, if the seller is from the different city our model might fail to predict the accurate prize of that used car.