**Car Price Prediction**

**1. Introduction**

**2. Loading Data and Explanation of Features**

**3. Exploratory Data Analysis (EDA)**

1. **Feature Engineering**
2. **Feature Selection**

**6. Applying Regression Models**

**7. Experimental Results**

**8. CONCLUSION**

1. **Introduction :**

Car price prediction is an important task in the automobile industry. Machine learning algorithms can be used to predict the price of a car based on various features such as make, model, year of manufacture, mileage, and more.

This project aims to solve the problem of predicting the price of a used car, using Sklearn's supervised machine learning techniques.

It is clearly a regression problem and predictions are carried out on dataset of used car sales. Several regression techniques have been studied, including Linear Regression, Decision Trees and Random forests of decision trees. Their performances were compared in order to determine which one works best with out dataset.

1. **Loading Data and Explanation of Features :**

The datasets consist of several independent variables include:

* Name
* Year
* km\_driven
* fuel
* seller\_type
* transmission
* owner
* mileage
* engine
* max\_power
* torque
* seats

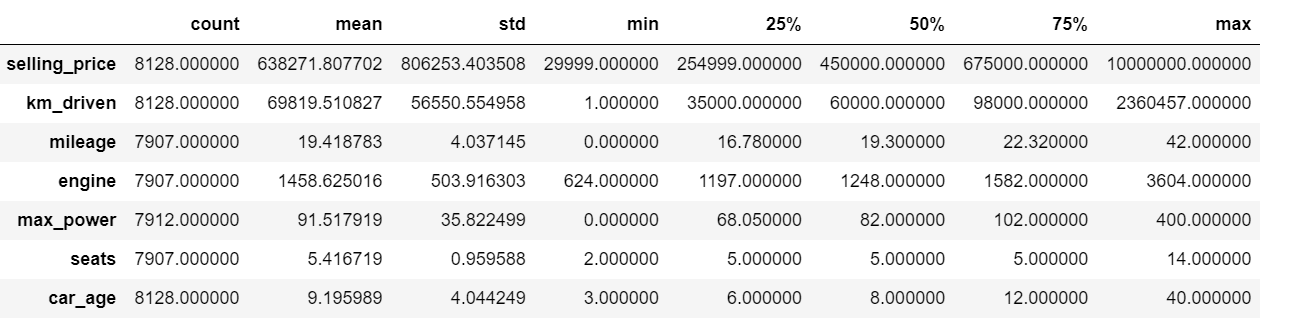
Dependent variable include

* selling\_price

1. **Exploratory Data Analysis (EDA):**

**Following steps are performed to do exploratory data analysis on the data**

1. **looking describe dataset**
2. **make dtypes of some variable 'category'**
3. **create 'car\_brand\_name' feature from 'name' feature**
4. **extract value of engine and mileage variable**
5. **extract value of 'max\_power' variable**
6. **create 'car\_age' feature from 'year' column**
7. **drop the features of 'name','year' and 'torque'**
8. **describe price value**
   1. **Avg price= 638271**
   2. **Min price= 29999**
   3. **Max price= 10000000**
9. description of numeric variable
   1. minimum selling price is 29999 USD and maximum price is 10000000 USD and average selling price is 638271 USD.
   2. The driving distance of the least driven car is 1 km, the most driven car's driving distance is 2360457 km, and average driving distance is 69819.
   3. The no. of seats of cars change from 2 seats to 14 seats
   4. Minimum mileage is 0, maximum mileage is 42, average mileage is 19.4
   5. Engine volume changes from 624 to 3604, average is 1458.

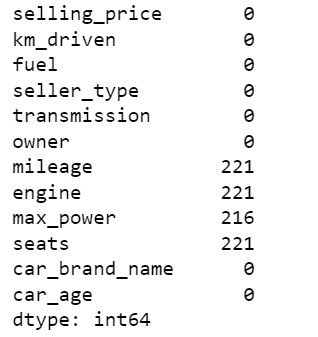


1. **About categoric variable**
2. **Fuel kind with highest frequency :- Diesel (frequency=4402)**
3. **seller type with highest frequency:- Individual(frequency=6766)**
4. **Transmission type with highest frequency:- Manual ( frequency= 7078)**
5. **owner type with highest frequency :- First Owner (frequency= 5289)**
6. **car brand name with highest frequency : Maruti (frequency= 2448)**

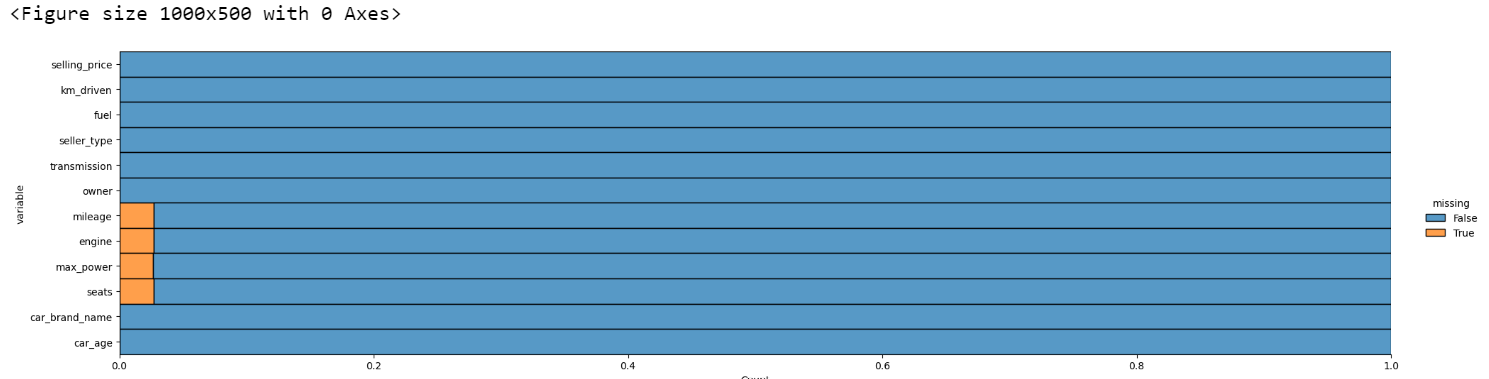
1. **Feature Engineering:**

The next step is to select and engineer the relevant features that can help in predicting the price of a car. Some of the important features that can be used in the model include make, model, year, mileage, fuel type, engine size, transmission type, and other features that can affect the value of a car.

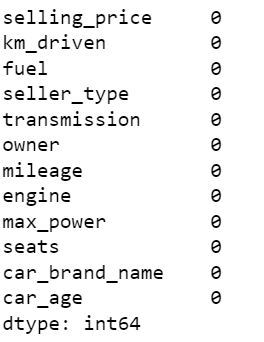
1. checking missing value and fill them



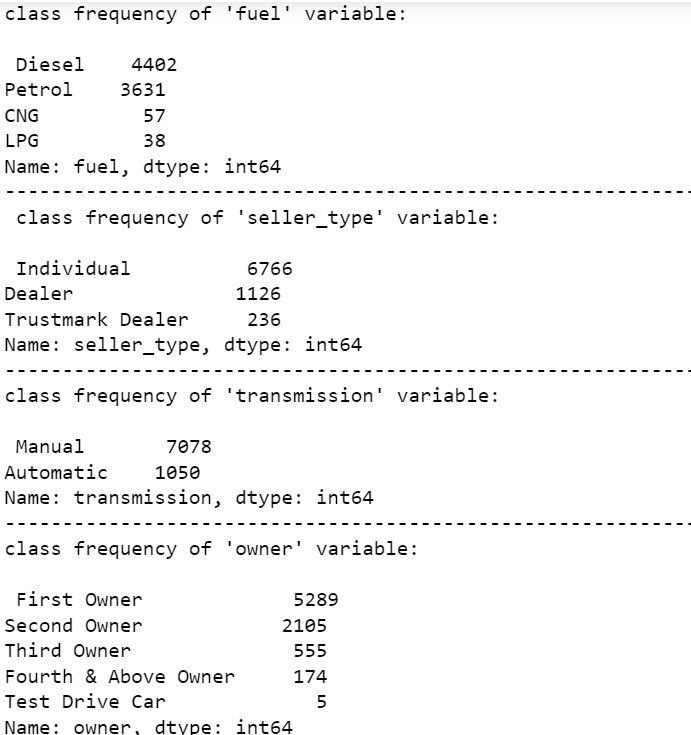
1. visualize missing values with seaborn (distplot)



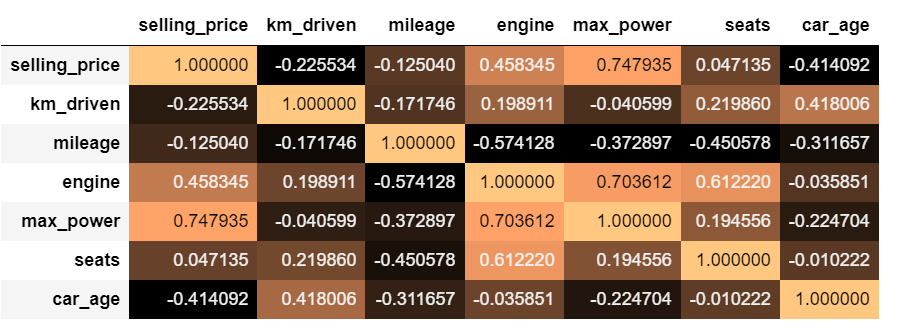
1. fill null values with median (numeric) and frequent values('categoric')
2. fill numeric data with median / KNN imputer
3. Replace missing values in each categorical column with the most frequent value
   1. we filled null values with medians of numeric variable and the most frequent values of categorical variable with the codes above
4. check the null value again



1. get frequency of some variable



1. check correlation between the variable of dataset



Some correlation detected.

As seen there is high correlation between some of the variables:

1. Between ' mileage ' and 'engine' variables is -0.57

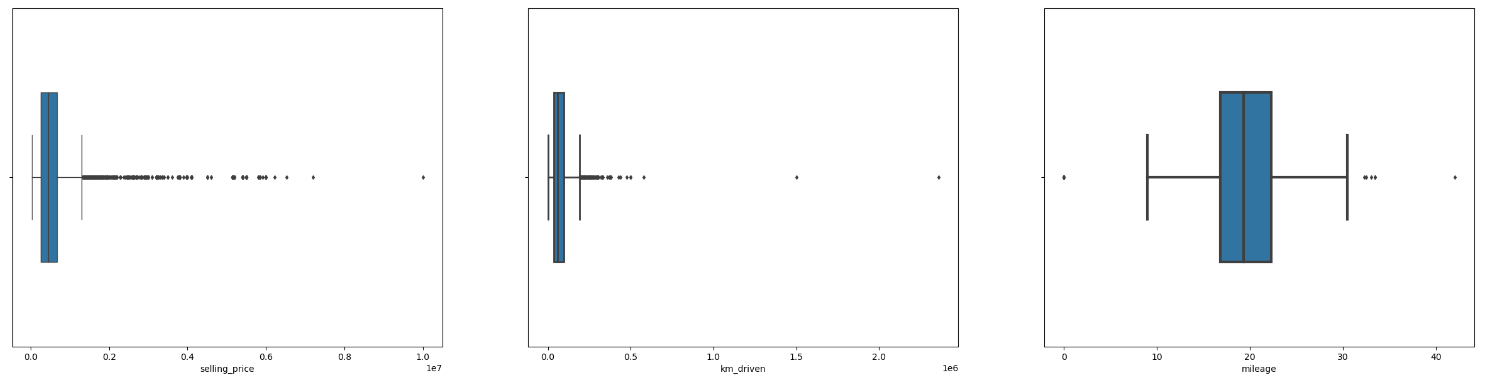
2. Between ' max\_power' and 'engine' variable is 0.70

3. Between ' seats' and 'engine' variable is 0.61

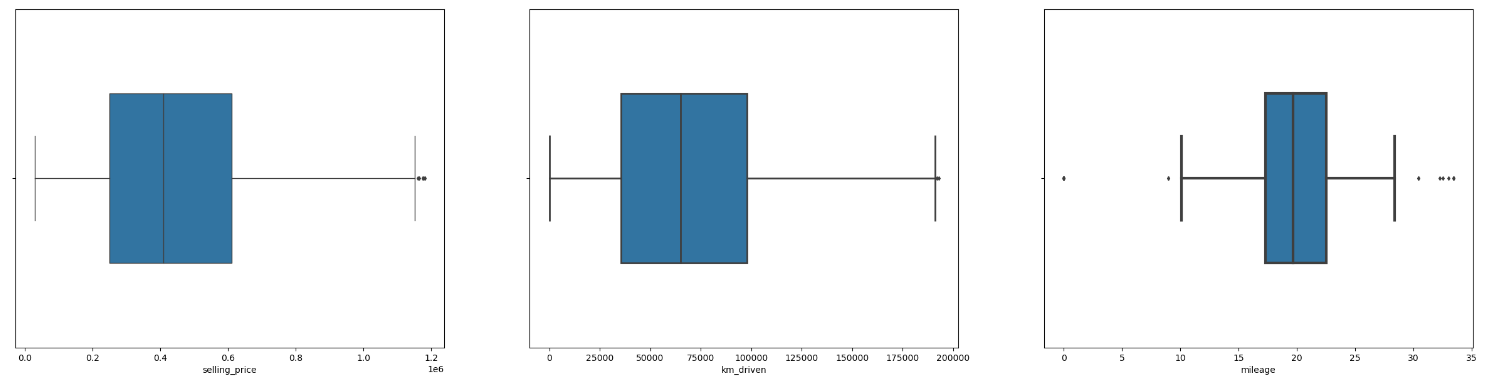
4. Between 'max\_power' and 'selling\_price' variables is 0.74

There is middle level of correlation between other variables too.

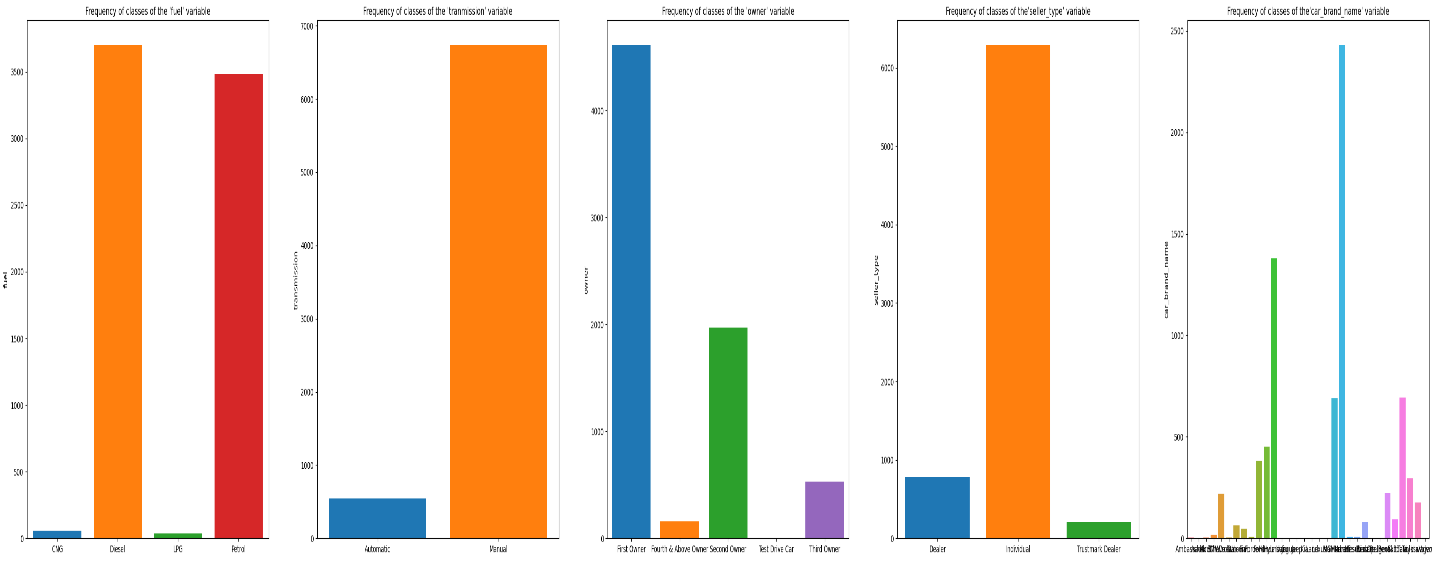
1. Checking Outliers



1. Removing Outliers from selling\_price , km\_driven



1. visualizes frequency of class categoric\_columns



1. convert categorical variable into numerical variable using Label Encoding
2. **Feature Selection :**

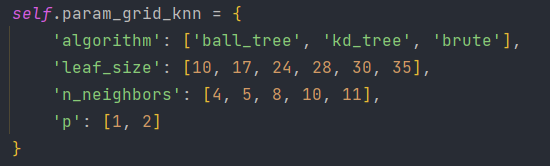
Feature selection is an important step in building a car price prediction model using machine learning. The goal is to identify the features that are most relevant to predicting car price while avoiding overfitting or including unnecessary features in the model.

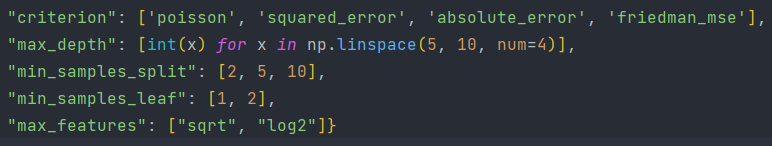
Here are some steps to perform feature selection.

1. Identified Highly Correlated Features: If two or more features are highly correlated with each other, they may not both be needed in the model. Used correlation matrices or scatter plots to identify features that are highly correlated.
2. Used Feature Selection Techniques: There are several feature selection techniques available, such as Recursive Feature Elimination (RFE), Univariate Feature Selection, and Principal Component Analysis (PCA). These techniques used to identify the most important features to include in the model.
3. **Applying Regression Models**

The problem statement was solved using four machine learning algorithms namely Linear Regression, K-Nearest Neighbor (KNN), Decision Tree, and Random Forest. Each model was hyper tuned using RandomisedSearchCV algorithm to obtain the best parameters for each algorithm. The model performance was evaluated using mean absolute error, mean squared error and r-squared value. The best performing model was selected based on r-squared value, which ensured our model was free from overfitting and a generalized model is obtained.

The parameters that were tuned for KNN algorithm, leaf\_size, n\_neighbors, and p values, where algorithm is the algorithm used to compute nearest neighbors, leaf\_size is the leaf size passed to ballTree or kdTree algorithms, n\_neighbors is the no of neighbors to be used, and p represents power parameter for Minkowski metric.



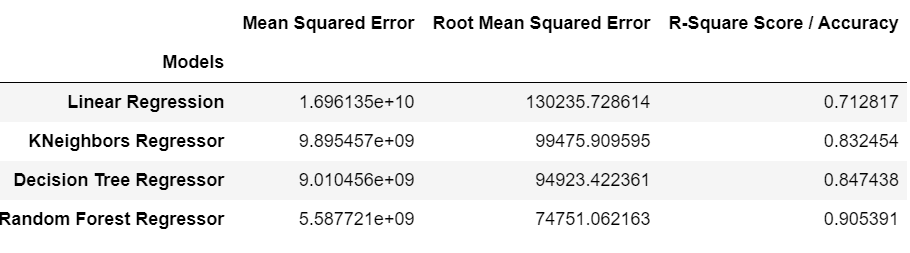
Similarly, the parameters that were tuned for decision tree and random forest algorithm, criterion, max\_depth, min\_samples\_split, min\_samples\_leaf, max\_features, where criterion is the function to measure the quality of split, max\_depth is maximum depth of the tree, min\_samples\_split is the minimum no of samples required to split an internal node, min\_samples\_leaf is the minimum no of samples required to be at leaf node, and max\_features represents is the no of features to be looked for best split.

The model should be trained on the training set and evaluated on the testing set.

The performance of the model should be evaluated using metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-Squared (R2).

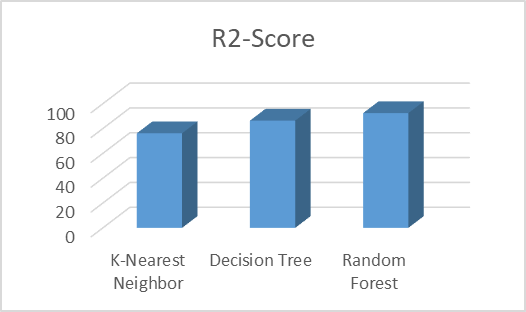
1. **Experimental Results**

The table below shows experimental results obtained for each model. Evaluated the performance of the model using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or R-squared.

****

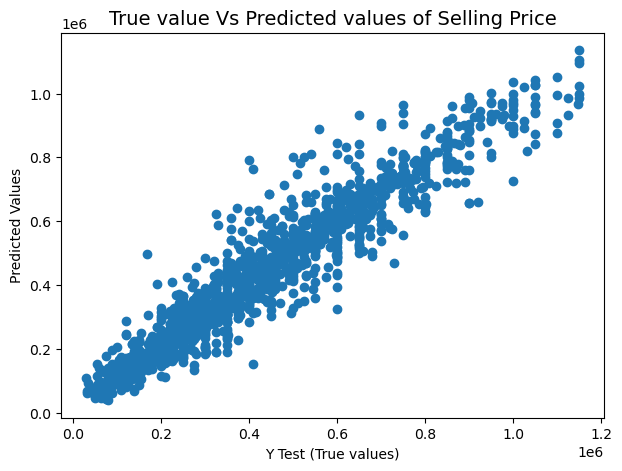
1. **CONCLUSION**

In conclusion, building a car price prediction model using machine learning involves collecting and preprocessing data, building a model, and deploying the model for predicting car prices. The accuracy of the model can be improved by using more features, tuning hyperparameters, and using more advanced machine learning algorithms.

****

Here **Random Forest algorithm gives highest R2 score/ accuracy as 90.56 %.**

The actual vs predicted values are



The residuals of RF is

