Car Price Prediction

Dhainik Suthar

Car Price Prediction

Predict car price using 12 predictors

Column names:

Column name	Column Description
name	Car name
year	Car model year
Selling_price	Car Price (output variable)
Km_driven	Car driven distance in km
fuel	Petrol/diesel/LPG
Seller_type	Individual/Dealer
transmission	Manual/Automatic
owner	First/second
mileage	Mileage In kmpl
engine	Engine in CC
Max_power	Power in bhp
torque	torque
seats	seats

Predict selling_price using 12 predictors

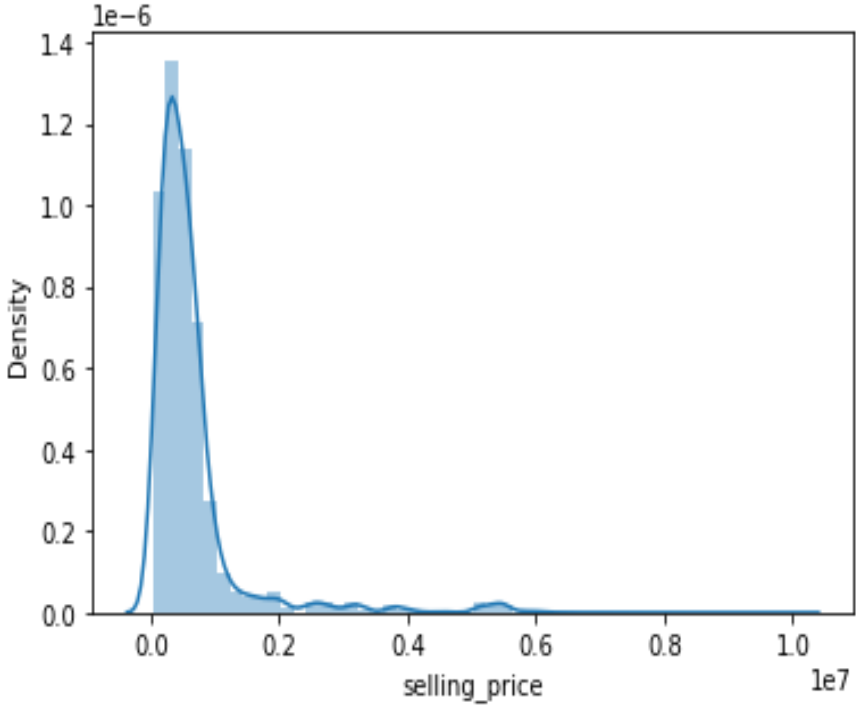
Dataset size : 8128*13

7 continuous variables, 5 character variable

Data Modelling approach

- Check for missing values:
- Feel missing value
- Data type conversions (String to Integer)
- Remove outlier using graphs
- Convert into normal distribution
- Built machine learning models on the data
- Important predictors

Distribution of target variable



Selling_price skew = 4.19

Correlation among predictors

Model Performance Metrics

Mean squared error:

a measure of how close a fitted line is to data points

adjusted r2 score:

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.

Mean absolute error:

Adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model.

Model 1: Linear Regression

Linear Regression show linear relationship between input and output variables.

Mean squared error : 0.0947

Adjusted r2 score : 0.7498

Mean absolute error : 0.2281

Model - 2 : Decision Tree Regressor

A decision tree is a graphical representation of all possible solutions to a decision based on certain conditions.

mean squared error : 0.089

Adjusted r2 score : 0.7537

Mean absolute error : 0.188

Model - 3 : Random Forest Regressor

-Random forest fits multiple decision trees and averages them

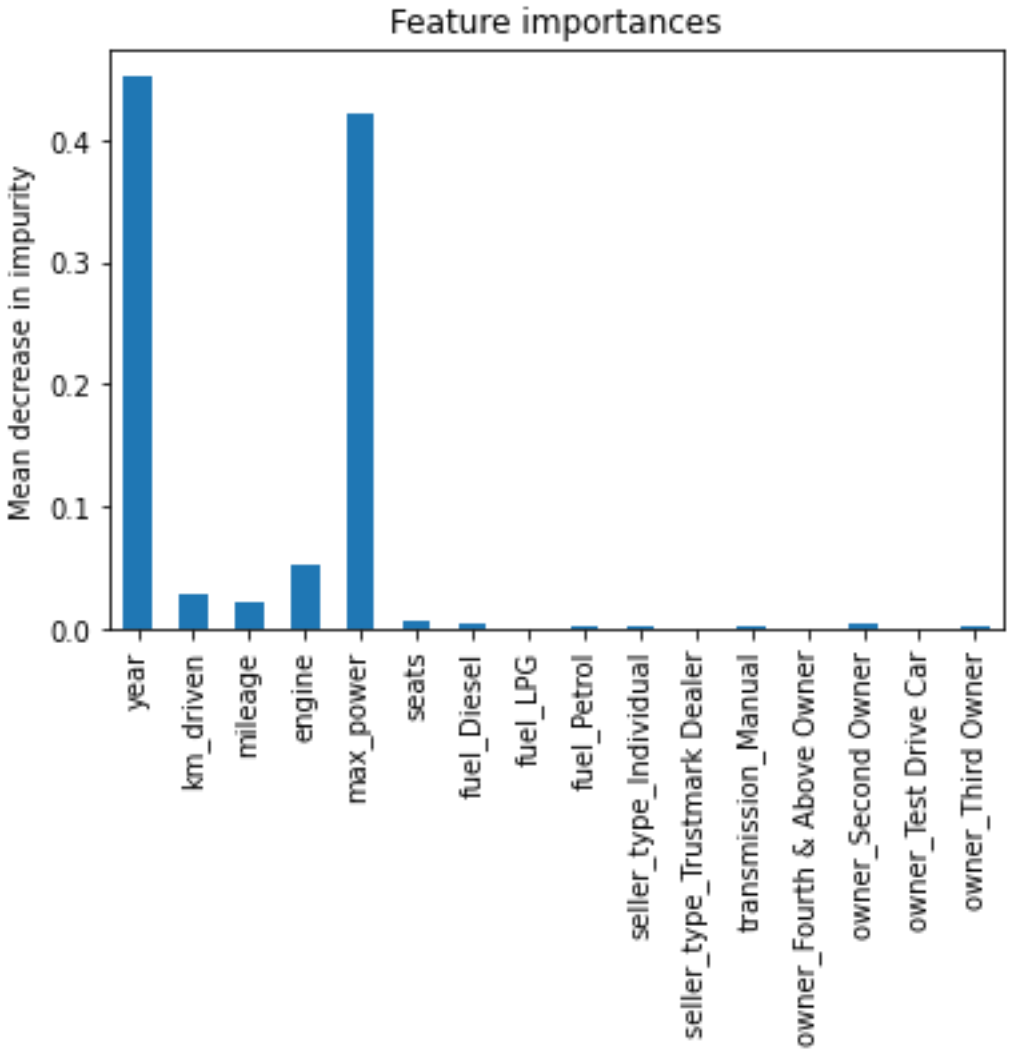
-This reduces the tendency to overfit but also adds complexity

mean squared error : 0.050

Adjusted r2 score : 0.8569

Mean absolute error : 0.151

Feature Importance in random forest



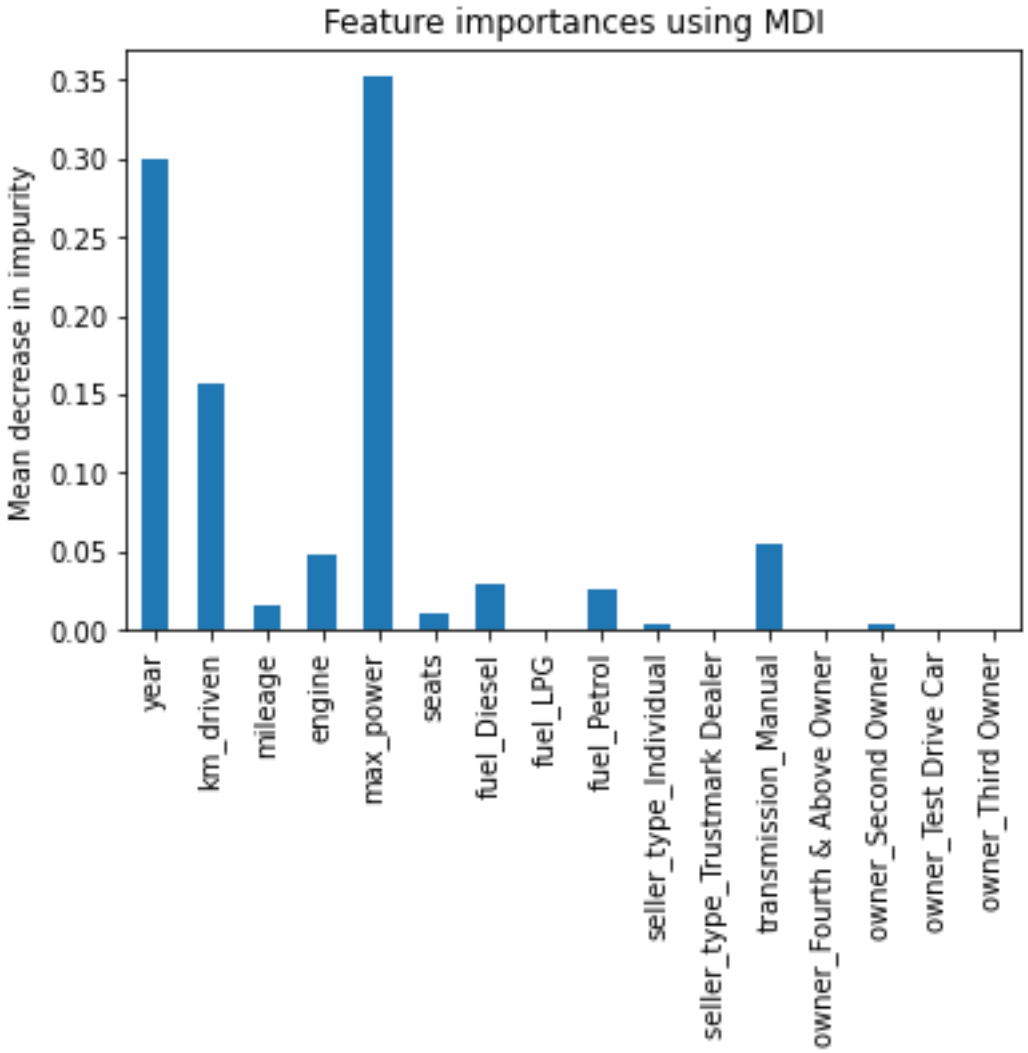
For random forest year and max_power is most important features.

Model - 4 : AdaBoostRegressor

mean squared error: 0.108

Adjusted r2 score: 0.706

Mean absolute error: 0.2589



Model – 5 : GradientBoostingRegressor

mean squared error: 0.057165754522063766

Adjusted r2 score: 0.8384259103542547

Mean absolute error: 0.17473602468416563

XGBRegressor

mean squared error: 0.050311619579111844

Adjusted r2 score: 0.8570559384763965

Mean absolute error: 0.14975477934081596

Results

Model	MSE	Ad. R2 score	MAE
Linear Reg.	0.0919	0.7469	0.2281
Decision Tree	0.0893	0.7537	0.1884
Random Forest	0.0503	0.8569	0.1510
Ada Boost	0.1082	0.7060	0.2589
Gradient Boosting	0.0571	0.8384	0.1747
XGB Regressor	0.0503	0.8570	0.1497

Overall, XGB regressor is the best performing model with:

MSE: 0.0503

Ad. R2 score: 0.8570

MAE: 0.1497