
Vehicle price prediction
- FNU Vishal

01

Problem Statement

Why and Where we need this?

Problem Statement

Determining the appropriate price of the vehicle play a crucial role in vehicle brokerage. But it is not done correctly at all times which leads to losses and decreased sales rate.

Our objective is to identify the features which acts as a deciding factor of the vehicle price. To build a regression model which helps to predict the correct price of the vehicle.

Our task includes cleaning and analysing the data, building a number of machine learning models, Training and Testing of those models and identification of best suitable model.

02

Data Cleaning

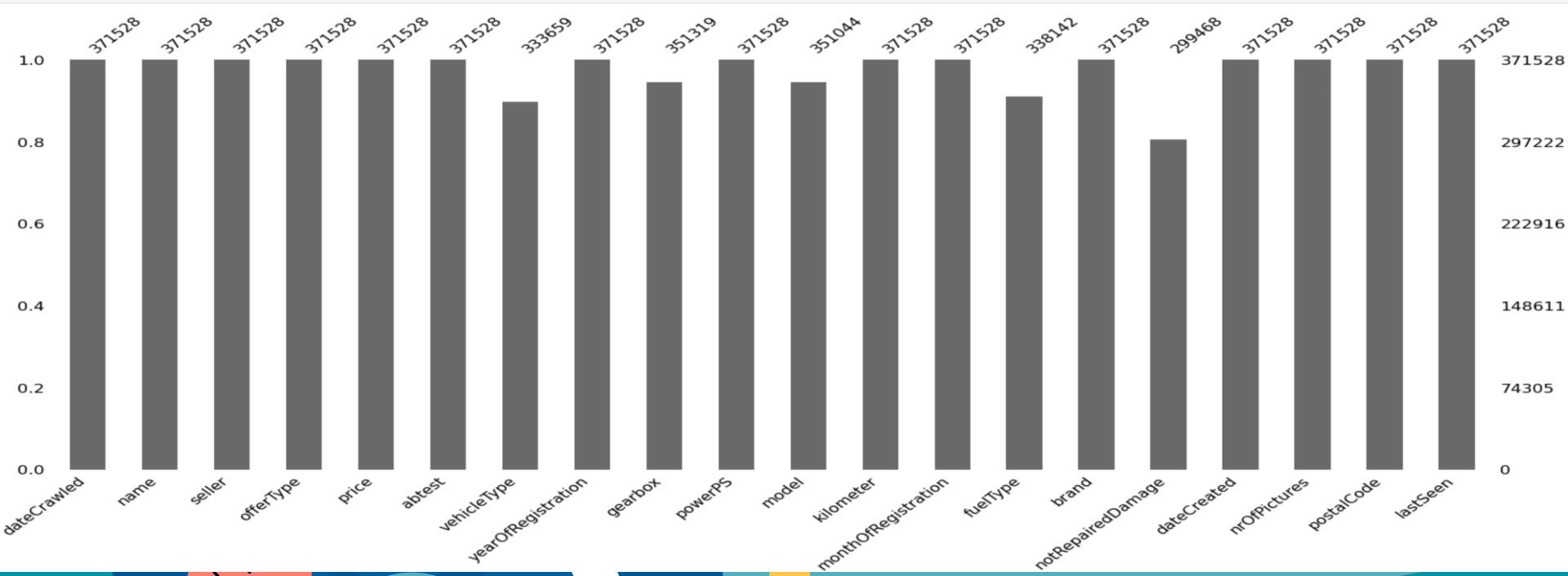
To make the data consistent

Data Cleaning

The dataset contains 371528 rows and 20 columns

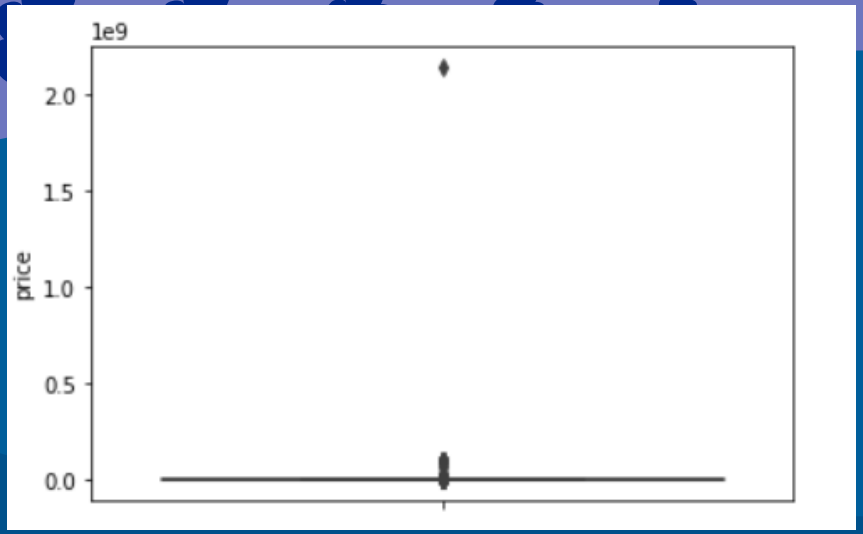
- We have started by generating a list of null values in each feature after the initial import into the DataFrame.
- After that, we have checked all the datatypes of all columns and used .isnull() to give a rough sense of the distribution for each non-numeric feature. It helps us to understand if the data is nominal or ordinal.
- Then we have moved towards several visualizations of the data to check the outliers and treatment of missing values. All actions are given below
  - Dropped rows where the price is more than 40000 or less than 10.
  - No missing value in the price or target column so no missing value treatment is applied.
  - Dropped some not important columns such as name, monthOfRegistration, and lastSeen.
- Based on that we have dropped several rows and columns that are not at all important. The final share of the data is (356769, 20) at the end of cleaning process.

Checking null with msno()

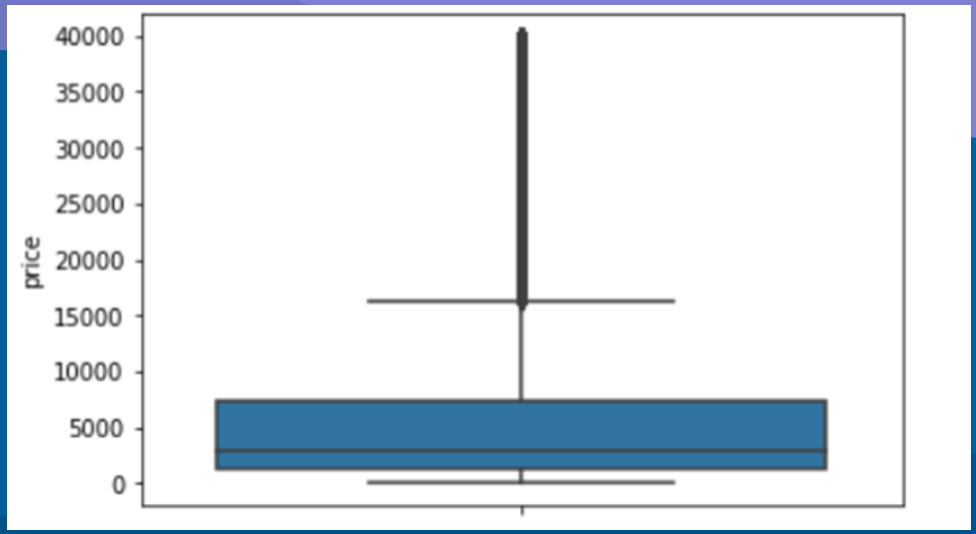


Removing outliers

Before

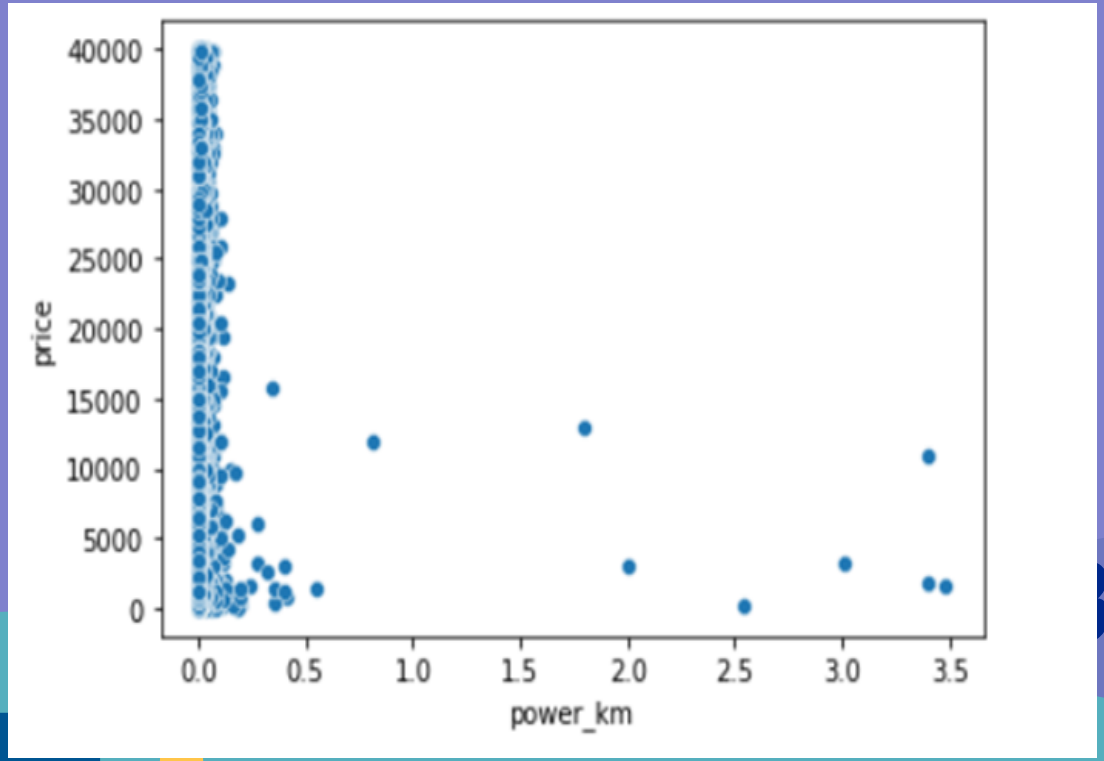


After



Removing outliers

- Removed powerPS and kilometer
- And keep power_km
- power_km = powerPS/ kilometer



03

Exploratory Data Analysis

Story telling by the DATA

Exploratory Data Analysis

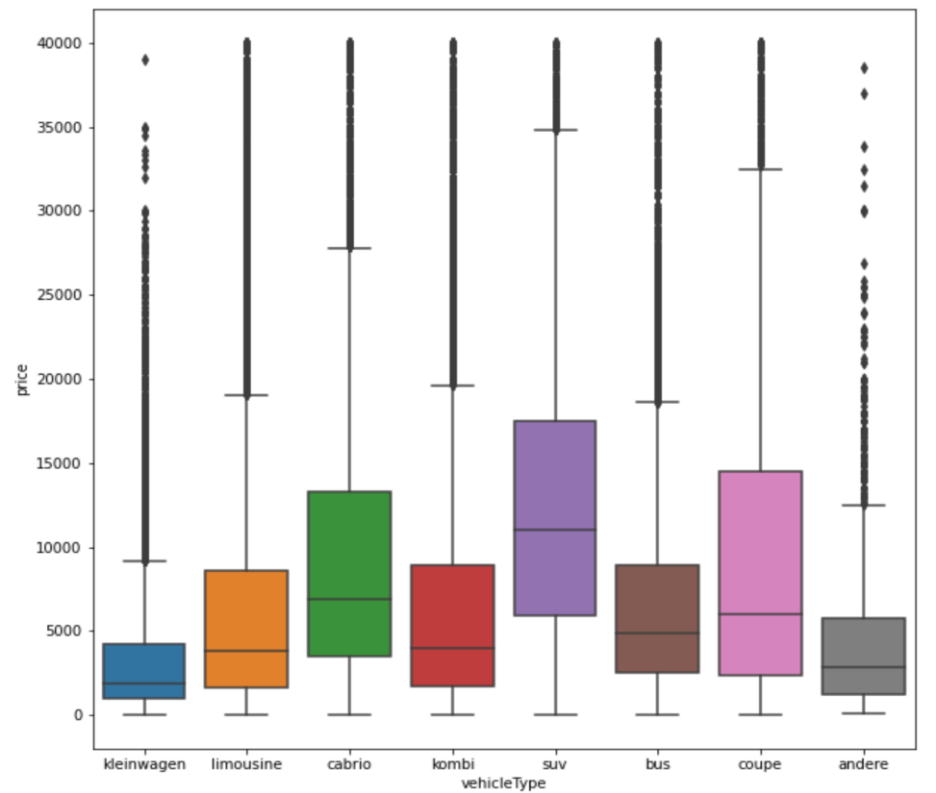
- The vehicle type SUV has the highest average price of 12115.19
- The vehicle type kleinwagen has the lowest average price of 2796.41
- The vehicle model Transporter has the highest average price of 9348.80
- The vehicle model Twingo has the lowest average price of 9348.80
- The vehicles sold by private sellers has the highest average price of 5470.40
- The kilometer feature has indirect relation with average price of the vehicle
- Year 2000 has the Most number of vehicle registrations - 22927
- The vehicles with automatic gearbox are valued more.
- The vehicles from 1940's and 1950's are valued more.
- The model c_klasse is valued more in SUV vehicle type.

04

Visualizations

Adding colors and shapes to that story

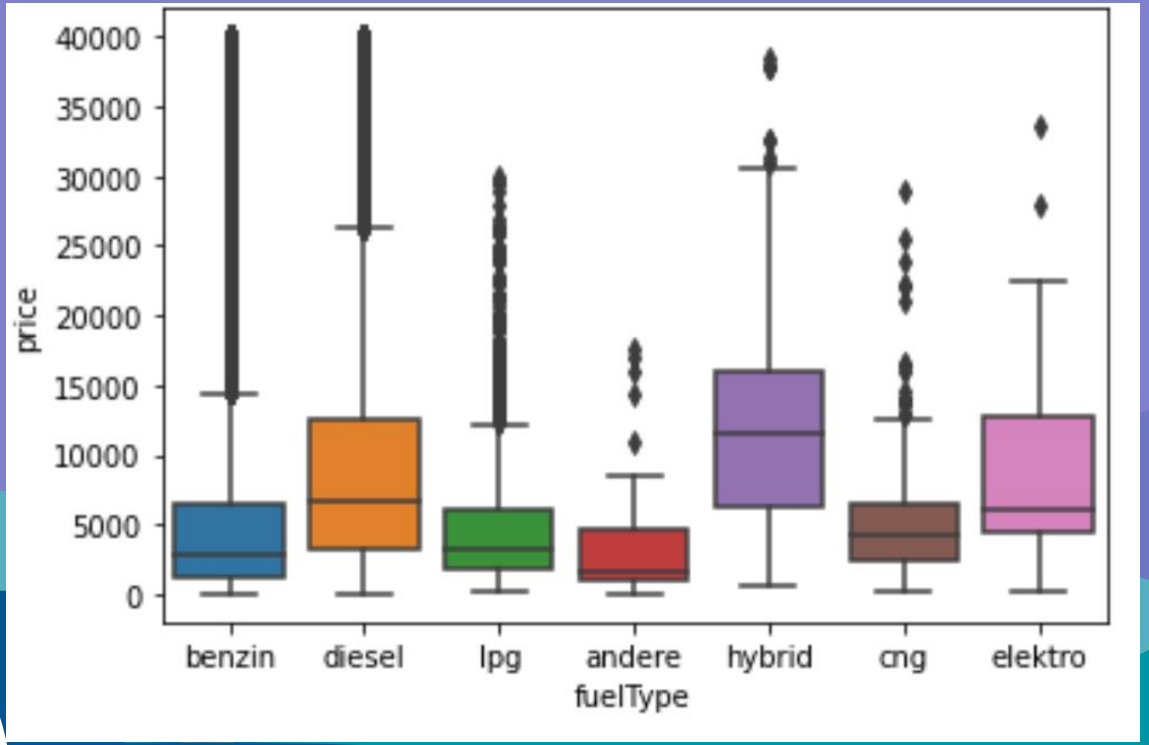
vehicleType and price

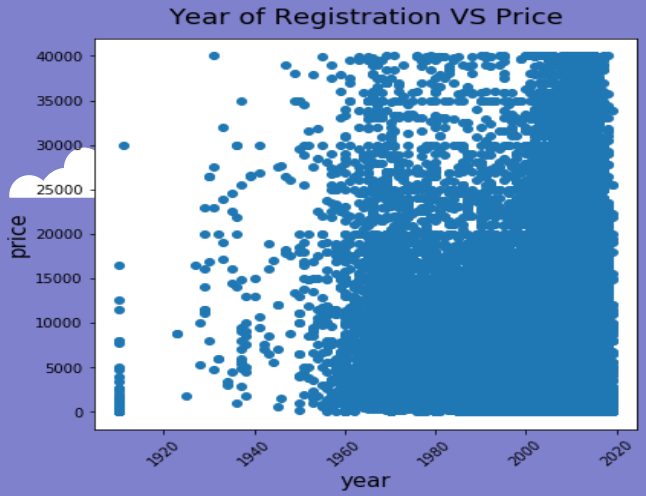


We have 8 different types of vehicles in our data set and most the cars who has higher prices are copupe and Suvs.

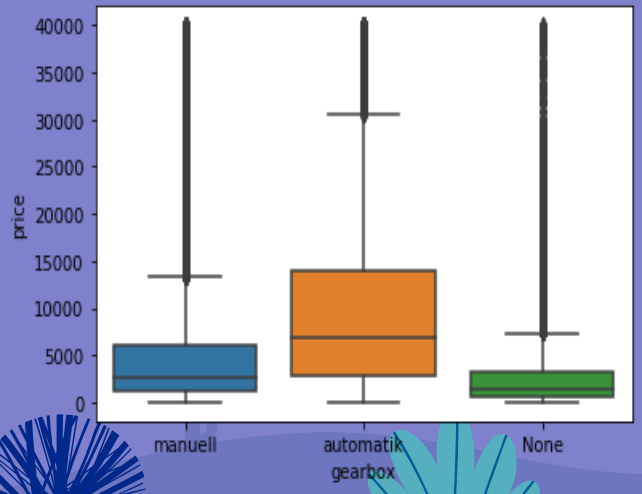
Relationship between Price and Fuel type

Hybrids cars are most commonly used and popular against other fueltype cars. Electric and diesel comes after hybrids cars. Leastused car is ‘andre’ fueltype cars.

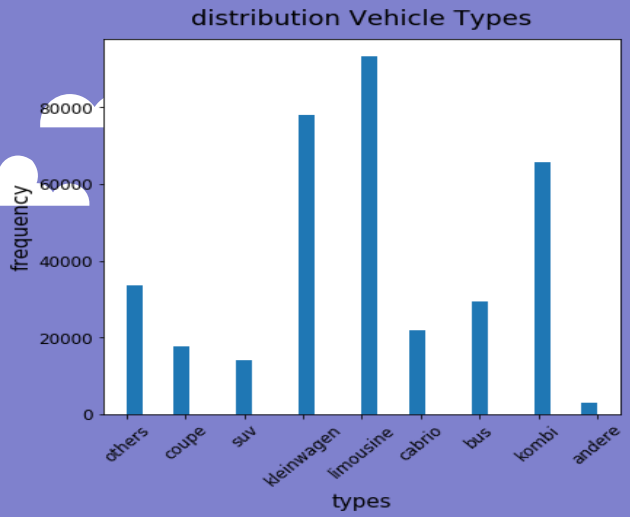




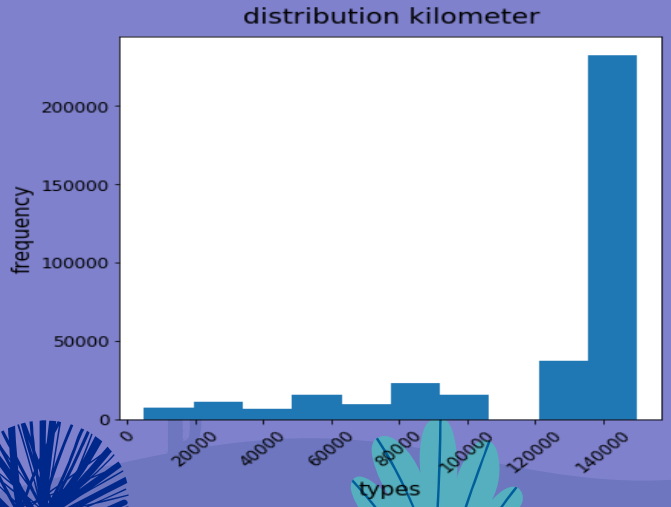
The scatter plot shows the relation between price and year of registration. From the plot it is observed that recently registered vehicles has higher price mostly



The Box plot gives the statistical information about the data. From the plot we can observe the median, symmetry of data and to know the outliers.



The Histogram shows the distribution of vehicle types. From the plot it is observed that limousine type is more common.



The Histogram shows the distribution of kilometer. From the plot it is observed that most of the car have more than 120,000 kms of run.

05 Modeling

Various regression models are built, trained and tested

Data Preprocessing

- The features like vehicleType, gearbox, yearOfRegistration, model and kilometer are selected to predict the price of the vehicle.
- Since vehicleType, gearbox and model are categorical data it is dummified using .get_dummies()
- Then the data is splitted into training and testing set in the ratio of 7:3

01 Linear Regression

Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set

Performance:

RMSE

- Test : 4463.06
- Train : 4474.81

Score

- Train : 0.5621
- Test : 0.5582

02

Ridge Regression

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization.

Performance:

RMSE

- Test : 4478.06
- Train : 4463.85

Score

- Train : 0.5621
- Test : 0.5582

03

LASSO Regression

Lasso regression is a regularization technique. It is used over regression methods for a more accurate prediction. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable.

Performance:

RMSE

- Test : 4465.95
- Train : 4480.50

Score

- Train : 0.5617
- Test : 0.5578

04

Decision Tree

Decision Tree regression model consists of an ensemble of decision trees. An aggregation is performed over the ensemble of trees to find a Gaussian distribution closest to the combined distribution for all trees in the model.

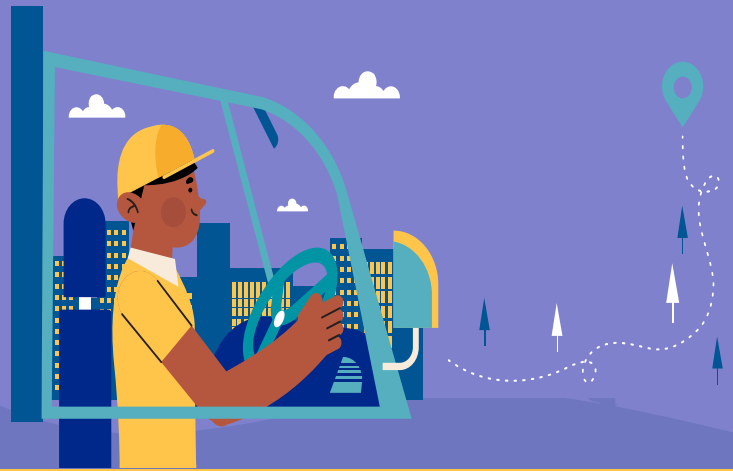
Performance:

RMSE

- Test : 3425.86
- Train : 2766.53

Score

- Train : 0.8283
- Test : 0.7398



05 Bagged Decision Trees

Bootstrap Aggregation is a general procedure that can be used to reduce the variance for those algorithm that have high variance. An algorithm that has high variance are decision trees, like classification and regression trees

Performance:

RMSE

- Test : 3296.04
- Train : 2820.88

Score

- Train : 0.8215
- Test : 0.7591

06 Random Forest Regression

Random forest is a type of supervised learning algorithm that uses bagging method,The algorithm operates by constructing a multitude of decision trees at training time and outputting the mean/mode of prediction of the individual trees.

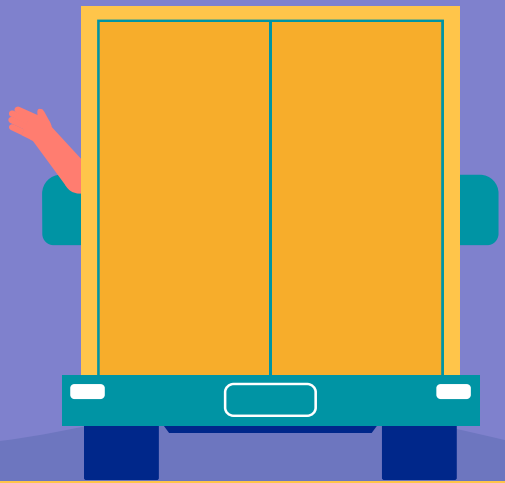
Performance:

RMSE

- Test : 2759.14
- Train : 3202.83

Score

- Test : 0.80845
- Train : 0.74250



07

AdaBoost Regressor

An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

Performance:

RMSE

- Test : 4958.67
- Train : 4955.36

Score

- Test : 0.38133
- Train : 0.38360

08 Gradient Boosting Regression

Linear Regression is the process of finding a line that best fits the data points available on the plot, so that we can use it to predict output values for inputs that are not present in the data set

Performance:

RMSE

- Test : 3390.19
- Train : 3389.07

Score

- Test : 0.71081
- Train : 0.71168

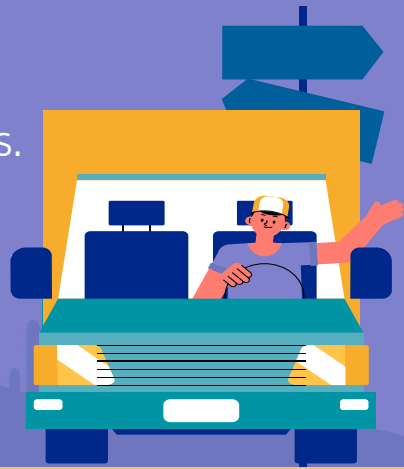
Performance Summary

Result

Model	Testing RMSE	R2 Training	R2 Testing
Linear Regression	4463.84	0.5621	0.5582
Ridge Regression	4463.85	0.5621	0.5582
Lasso Regression	4465.95	0.5617	0.5578
k-Nearest Neighbors	3682.11	0.6978	0.6596
Decision Tree	3425.86	0.8283	0.7398
Bagging Decision Trees	3296.04	0.8215	0.7591
Random Forest	3277.08	0.8243	0.7619
AdaBoost	4592.47	0.5325	0.5324
Gradient Boosting	3441.97	0.7403	0.7373

Conclusion And Recommendations

- The Random Forest regressor is found to be the best suitable model to predict the price of the vehicle.
- It has Less RMSE and more accuracy than all other models comparatively.
- As per the model of the car, it is predicting the best match of its price.
- If the kilometer is high, then the price will be low,
- If the gearbox is automatic, the vehicle price will between 10,000 to 40,000.
- more recent the yearOfRegistration is , higher the price.
- The SUV type vehicles are valued more at almost all conditions.



Thank you