**Auto Price Prediction**

**Understand the data set:**

It is an auto price data set contains 26 features(including target) and based on this feature’s attribute values variation we have to predict the price.

Out of 26 there are 11 Numerical features and 15 objective type features are there.

**Basic Checks:**

* Data contains 26 columns including the target variable.
* Target is ‘price’ in the dataset
* Data is very small in size with 201 datapoints
* normalized-losses, bore, stroke, horsepower and peak\_rpm features have few missing attributes classified as '?’.

**EDA:**

**Univariate Analysis**:

* 90% of vehicles are Gas type and only 10% vehicles are diesel type
* 82% vehicles are Std\_aspirated engines and only 18% are turbo aspirated engines.
* toyoa is the largest car making brand
* almost all vehicles are frount engine only with 99% share
* most of the cars are medium safe rated with 0 & 1 symboling score.
* almost all vehicles are frount engine only with 99% of total share
* 47% of vehicles having sedan body\_style
* 4 door and 2 door vehicles, almost have equl share
* 59% vehicles are front wheel drive , 37% are rear wheel drive type and very less number of cars are four wheel drive
* 72% vehicles are 'Overhead\_camshaft' engine vehicles
* 4 cylinder vehicles take more share with 78% and then 6 cylinder vehicels have 12%
* multi point fuel injection(mpfi) vehcles are more with 46% and then 2 barrel with 32%

**Bi-Variate Analysis:**

* length & width of the vehicles have somewhat positive relation with price
* if curb\_weight raises price also increasing
* engine\_size have direct postive relation with price( if engine\_size increases price increases)
* if city\_mpg and highway\_mpg reducing corresponding vehicle price increasing
* though diesel vehicels are 10% , the average cost of diesel vehicles are more than gas vehicles, so cost of diesel vehicles are higher than gas vehicels
* Price of Turbo\_aspiration engine vehicleas are higher than std aspiration\_engine vehilcles
* hardtop & convertable body\_style vehicle price is high and hatchback type vehicle price lesser
* cost of rear wheel drive vehicle are almost double the price of front wheel dive vehicles
* cost of rear engine vehicles are double than the price of frount engine vehicles
* ohc engine vehicles are cheaper thar other engine type vehicels
* if number of cylinders are more price is also high
* top-3 highest priced cars are 'jaguar',mercedes-benz' and 'porche' and lowest priced car is 'chevrolet

**Multi-Variate Analysis:**

* city\_mpg & highway\_mpg are highly positive correlated (97%)
* wheel base have 87% positive correlation with length and 81% positive correlation with width and 78% +ve correlation with curb\_weight of vehicle
* wheel base is having 50% negetive correlation with the milageper\_gallon(mpg)

##### *length:*

* length and width are 85% +ve correlated and length and curb\_weight is having 88% +ve correlation
* length is almost 70% -ve correlated with milageper\_gallon(mpg)

##### *width:*

* width and curb weight also highly correlated (86% +ve)
* city\_mpg and highway\_mpg is 65% positively correlated with width

##### *curb\_weight:*

* curb\_weight and engine size 84% positively correlated
* curb\_weight is 75% +vely correlated with city\_mpg and highway\_mpg

**Data Pre-processing:**

* Treated(imputed) wrongly classified attributes in some of the features
* Used count of frequency and ordinal encoding techniques to encode the categorical objective type features
* Not treated outliers as those points are essential data points and performance getting dropped if I treat it.

**Feature Selection:**

* I have dropped few features having high correlation with other independent features, and low correlation with target variable.
  + "fuel\_type" and "city\_mpg"

**Model Selection:**

* As data contains many continuous features, I have done data scaling using Minmax Scaler .
* And all the models performing good with scaled data only

**Model Summary:**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Models** | | | | | | | | |
| **Metrics** | **Linear Regressor** | **KNR** | **Decision tree** | **Random Forest** | **Bagging LR** | **Ada Boost** | **GBR** | **XGBR** | **HR** |
| **R2 score** | **0.87** | **0.76** | **0.94** | **0.94** | **0.85** | **0.93** | **0.95** | **0.96** | **0.7** |
| **Adjusted R2 score** | **0.85** | **0.73** | **0.93** | **0.93** | **0.83** | **0.92** | **0.94** | **0.96** | **0.66** |
| **Mean squared error** | **16087165** | **28995407** | **7666880** | **7465907** | **18116305** | **9172570** | **6054429** | **4762545** | **37123875** |
| **Mean absolute error** | **2678.49** | **3415.35** | **1919.76** | **1644.81** | **2856.21** | **2304.36** | **1529.83** | **1396.69** | **3680.29** |
| **Root Mean squared error** | **4010.88** | **5384.74** | **2768.91** | **2732.38** | **4256.33** | **3028.63** | **2460.57** | **2182.33** | **6092.94** |
| **Train accuracy** | **0.90** | **0.63** | **0.99** | **1.0** | **0.89** | **0.95** | **0.99** | **0.99** | **0.84** |
| **Test Accuracy** | **0.87** | **0.43** | **0.94** | **0.94** | **0.85** | **0.93** | **0.95** | **0.96** | **0.70** |
|  |  |  |  |  |  |  |  |  |  |

**Result:** XGB regressor performing better than all other models

**Challenges Faced:**

* Found Missing Attributes in ‘Age’ Independent feature , which are denoted as ‘?’, and the feature is of ‘objective’ type, so I have replaced ‘?’ using mean/median/mode of the feature.
* ‘Make’ feature have 22 categories , so used different encoding techniques to encode these categories