Cardekho Team Level Documentation

Note - Consider opening this file in google doc for better visibility

# Data Extraction and Concatenation:

The csv is in unstructured format to make it semi structured format I have done a three steps :

## Step 1 - Extracting Column Wise Data:

* Reading File and converting json-string into a Dictionary Format:
* Extracting the data inside the dictionary and putting it as a separate columns:
* Finally concatenating all the dataframe

## Step 2 - Extracting the column 18 to 52 (Top and Data Dictionary):

* JSON-like strings into dictionaries
* Converting the Json-string into a Dictionary Cell wise:
* Handling and extracting the data inside the dictionary based on its structure:
* Converting the Data into the Data Frame:

## Step 3 - Extracting only the required data from the required column:

* Load the file and mentioning the column name which we want to extract:
* Function to Split List Values into Separate Columns:
* Applying the Function to Each Row for the Specified Columns:
* Concatenating the Structured File :

# Data Cleaning and Filtering only the Required Column:

## Step 1 - Deciding what are the column is necessary:

* Dropping unnecessary columns based on the Domain Knowledge:

There are about 8300 rows and nearly 129 columns cleaning all these columns is unnecessary and based on the use case.

* Dropping the Duplicate columns:

## Step 2 - Renaming the Columns for better understanding :

* This has around 23 columns and some columns are not abbreviated and some are not given a meaning so I have changed those kinds of columns in a meaningful way.

## Step 3 - Standardising Data Formats:

* Removing the Unnecessary units and other stuff in the required columns.
* Conversion or Replace in Price columns

1. Understanding of the Price Column
2. Split the values and its unit for better handling
3. Conversion of value into a correct range

## Step 4 - Type Conversion:

* After structured out what are the column need to be convert I have started the conversion process

## Step - 5 Handling the Missing value :

* Handling categorical Missing value :

1. Here I have used 2 techniques to fill the missing values.

* Handling the Numerical Missing data:

## Step - 6 Handling Outliers:

* Issue In Filling outlier
* Capping the outliers

# EDA - Exploratory data analysis.

## Step 1 - Visual Analysis

* Let see which car the cardekho have the most Top 15 car Count wise Based on Model
* Which variant does the cardekho have the most
* Which city have the more number of cars in cardekho
* Which fuel types cars does cardekho have in different cities
* Lets compare the same for the Transmission Type
* which year model cars do we have the most
* which type of safety systems we have in our car the most:
* which type of interior systems we have in our car the most:
* which Model perform exceptionally well
* How car age is important for the Price
* How car Previous owner is important to the Price
* Effect of Engine on price
* What type of Body type vehicle we do have the most
* Spread of Insurance Validity in cars in different cities

## Step -2 Hypothesis Testing for getting the idea for feature Importance

* kruskal - Test

1. City Vs price:
2. Fuel Type Vs price:
3. Manufactured By Vs The Price

* mann whitney u - Test

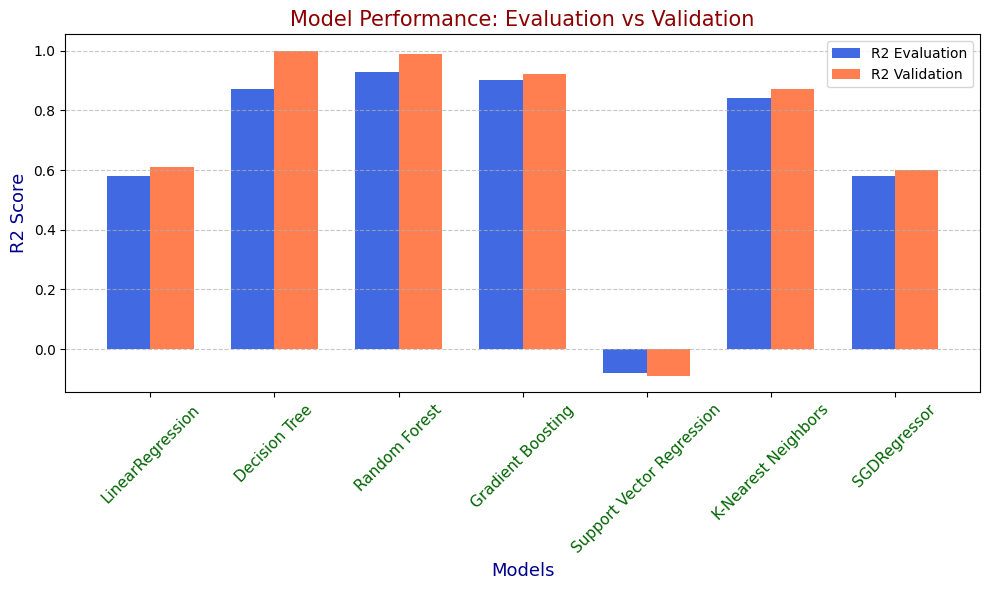
1. Transmission Type Vs Price

# ML - Machine Learning Algorithm and Developing the prediction model.

* Encoding for the Categorical data containing columns.
* Correlation of the Feature and target.
* Deciding and taking the Features and Targets
* Splitting the Features and targets for Training and Testing.
* Scaling or Normalization of the features
* Train 7 Model with this dataset to compare which model is performing well

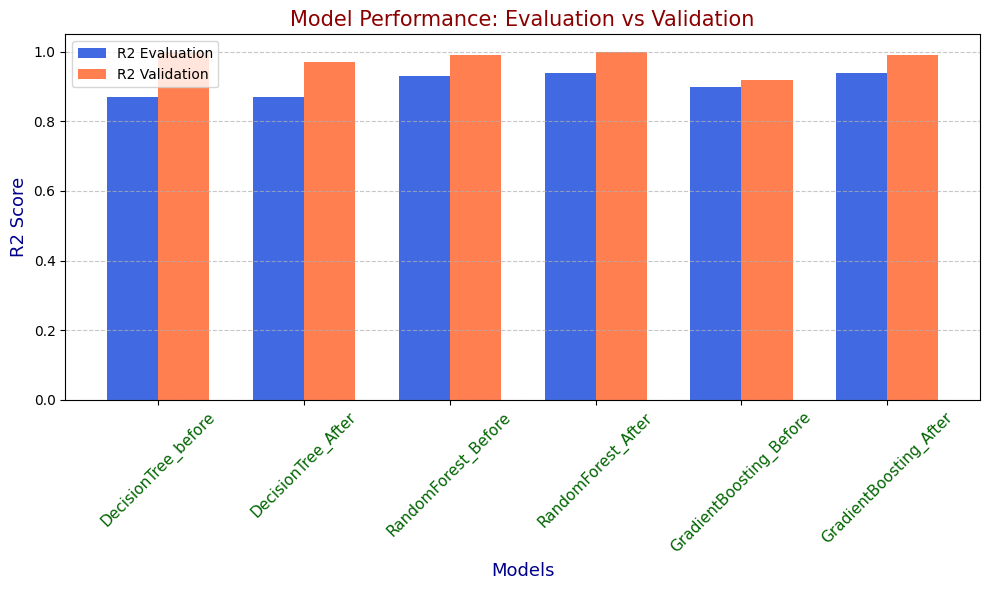
1. Linear Regression
2. Decision Tree
3. Random Forest
4. Gradient Boosting
5. Support Vector Regression (SVR)
6. K-Nearest Neighbors (KNN)
7. SGDRegressor

* Checking the R2 score of all the Trained Model



* Model Optimization - Using Grid search CV

1. Decision Tree
2. Random Forest Model Optimization
3. Finally The Gradient Boosting



Chosen model **Gradient Boosting \_ Before** model even though it has 90% accuracy it was not overfit or underfit so I have chosen this for a Deployment.

# App Deployment:

* Pages

1. Predict Price

Select Box Features - City, Body type, Transmission Type , Fuel Type and Manufactured By

Slider Feature - Kilometers Driven , Seats, Mileage,Engine CC,Previous owner, car age.

1. User Guide Page

Complete instruction of how to use the app.

Demo Video

That is the complete process of how I completed the project and steps I have done to complete the project.For detailed understanding of the project code and everything please refer to the Detailed documentation.

**Thank You**