LEARN AND BUILD

MINI PROJECT

NAME: R. Venkata Tagore Reddy **PROFESSION:** Data Science Intern

PROBLEM STATEMENT:

A used car dealership specializes in selling cars from various brands. They would like to know if the mileage of these cars is a good predictor of their sale prices, and if the slopes and intercepts differ when comparing mileage and price for different brands of cars. What other factors might play a role and how in deciding the price that a customer might be willing to pay. As a data expert, the company relies on your expert analysis and recommendations to increase their profitability by setting the right pricing for their car sales business, such that it delights the customers and gains the company positive feedback/reviews so that their traction in the market increases and they can become one of the key players.

DATASET SOURCE:

https://drive.google.com/file/d/1Li-1EvAmOW8qznQTQKHy3nueOYSf855V/view?usp=sharing

1.INTRODUCTION:

The used car dealership specializing in selling cars from various brands is seeking to understand the relationship between the mileage of cars and their sale prices. They want to determine if mileage is a reliable predictor of sale prices and whether the slopes and intercepts differ when comparing mileage and price for different car brands. By gaining insights into these relationships, the dealership aims to optimize their pricing strategy, enhance customer satisfaction, and establish themselves as key players in the market.

The dataset used for analysis consists of information on the mileage and sale prices of used cars. The key variables in the dataset are:

1. **Mileage:** This variable represents the recorded mileage of each car in miles. It indicates the distance the car has traveled and can serve as a proxy for its overall usage and condition.
2. **Sale Price:** This variable represents the price at which each car was sold. It reflects the market value of the car and is influenced by various factors, including mileage.

To initiate the analysis, exploratory data analysis (EDA) is conducted to identify patterns and relationships between mileage and sale prices for each car brand. Additionally, other interesting variables and their relationships will be explored to gain insights into factors that may impact the price customers are willing to pay.

During EDA, summary statistics and visualizations will be utilized to examine the distribution of the response variable (sale price) and its relationship with the mileage. Box plots, scatter plots, and correlation analysis will be employed to understand the associations between mileage and sale prices across different car brands. Furthermore, other relevant variables, such as the age of the car, fuel efficiency, or horsepower, may be explored to identify potential additional factors that might play a role in determining the price a customer is willing to pay.

The analysis will focus on uncovering meaningful insights to inform the dealership's pricing strategy and help them achieve their objectives of increasing profitability, customer satisfaction, and market traction.

2.REGRESSION ANALYSIS:

To compute the price for vehicles, this platform may compute linear regression model that defines a set of input variables. However, it does not give details as what features can be used for specific type of vehicles for such prediction.

We can observe that the distribution of prices shows a high positive skewness to the left (skew > 1). A kurtosis value of 17 is very high, meaning that there is a profusion of outliers in the dataset.

i.RELATIONSHIP OF PRICE WITH OTHER PARAMETERS:

The most important feature related to target “Price” is Power, Engine, Year, Seats, Kilometers_Driven, Mileage respectively which is found by analyzing the correlation between and sorted in descending order.

```
Find most important features relative to Price-target
Price              1.000000
Power              0.755995
Engine             0.687712
Year               0.503999
Seats              0.163818
Kilometers_Driven -0.047187
Mileage           -0.265965
Name: Price, dtype: float64
```

When we compare the Price with Fuel_Type, the vehicles which run via “Diesel” cost **high** than the vehicles run with “Petrol”.

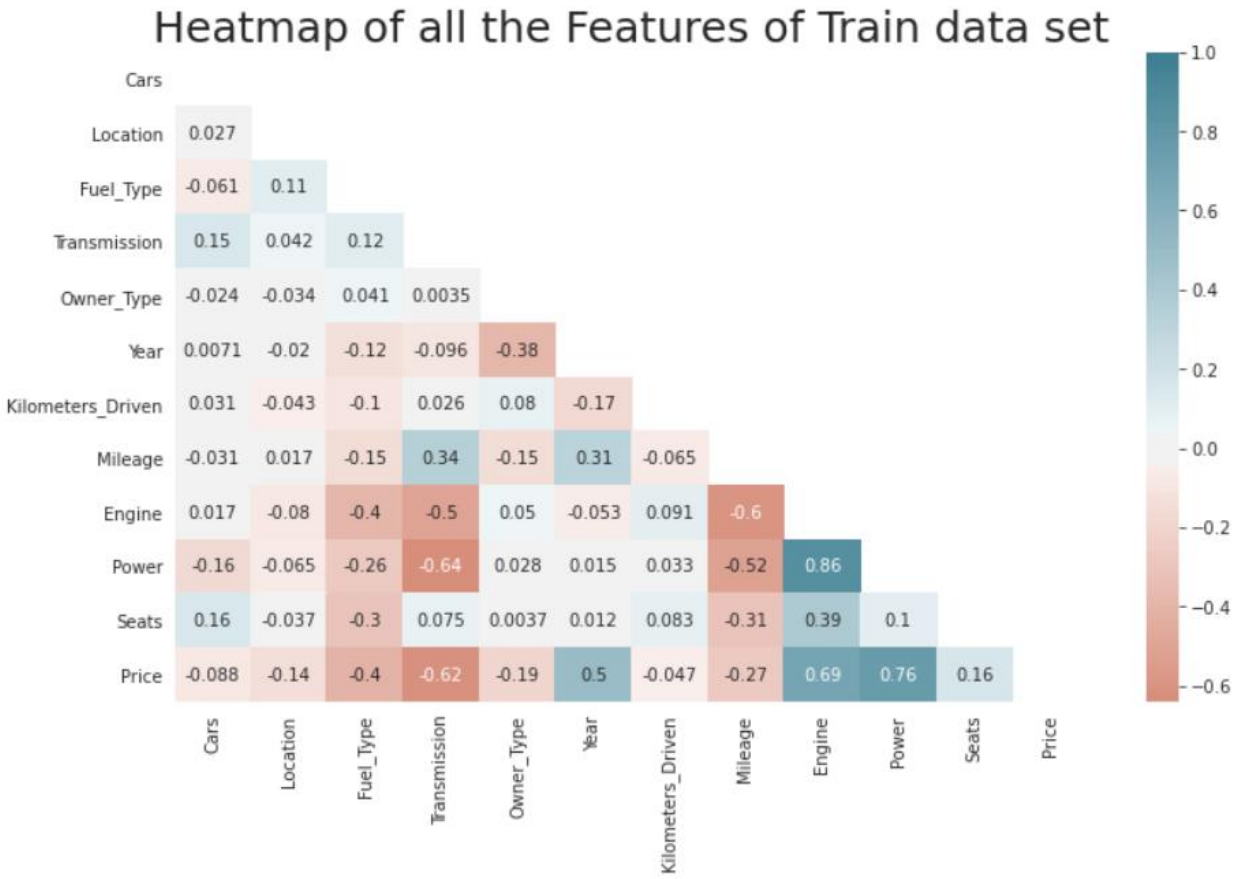
As the time goes on the “Price” of the cars which run automatically increases and manually gets decreasing respectively. Simply, "New" cars are more expensive than "Old" cars, and Automatic cars are more costly. From “Price-Year-Owner_Type” relation we can find that, the THIRD OWNER' CARS are sometime more **expensive** than the second-hand.

ii. Training and Testing

We split our dataset into training, testing data with a 70:30 split ratio. The splitting was done by picking at random which results in a balance between the training data and testing data amongst the whole dataset. This is done to avoid overfitting and enhance generalization. Finaly, we selected 11 characters in the dataset to train the model. The characters are: ['Cars', 'Location', 'Year', 'Kilometers_Driven', 'Fuel_Type', 'Transmission', 'Owner_Type', 'Mileage', 'Engine', 'Power', 'Seats','Price'].

We create different functions to calculate deviations, important features and graphical illustration.

Firstly, take a quick look at the correlation matrix:



iii. Models comparison

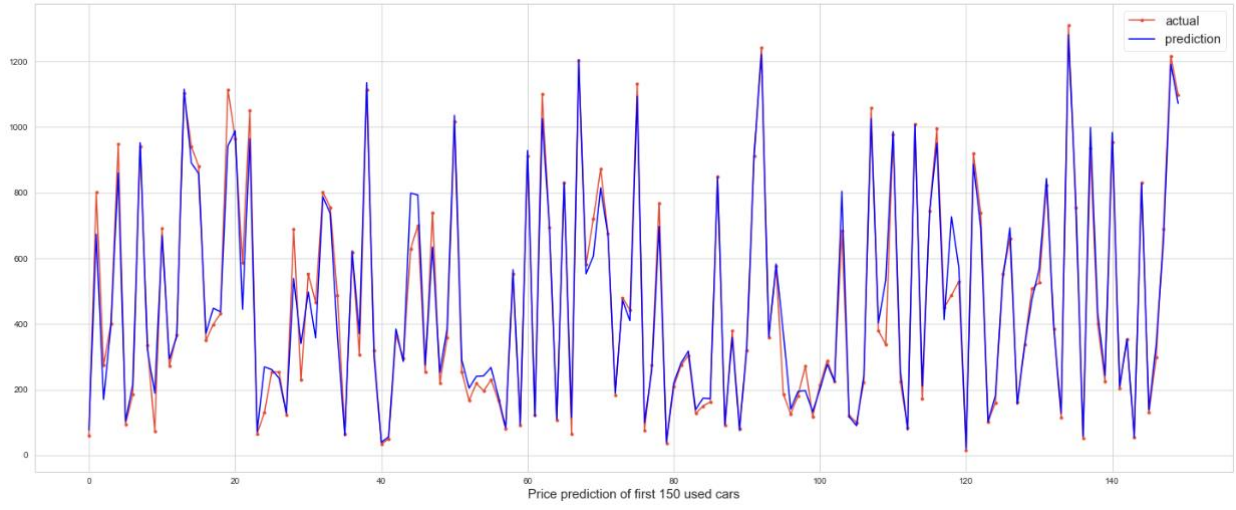
The model score is the coefficient of determination R2 of the prediction. In total, we have examinated 9 models to train/predict the used cars price.

The below table shows the values of “Root Mean Squared Error”, “Accuracy on Training Set”, “Accuracy on Texting Set” of 6 models. They are: “DecisionTreeRegressor”, “XGBRegressor”, “RandomForestRegressor”, “MLPRegressor”, “AdaBoostRegressor”, “ExtraTreesRegressor”.

	model	Root Mean Squared Error	Accuracy on Traing set	Accuracy on Testing set
3	MLPRegressor	196.206905	0.718621	0.679405
4	AdaBoostRegressor	149.916672	0.830629	0.812833
0	DecisionTreeRegressor	110.409521	0.999993	0.898482
2	RandomForestRegressor	84.120027	0.991532	0.941071
5	ExtraTreesRegressor	80.388945	0.999993	0.946183
1	XGBRegressor	74.815814	0.994635	0.953386

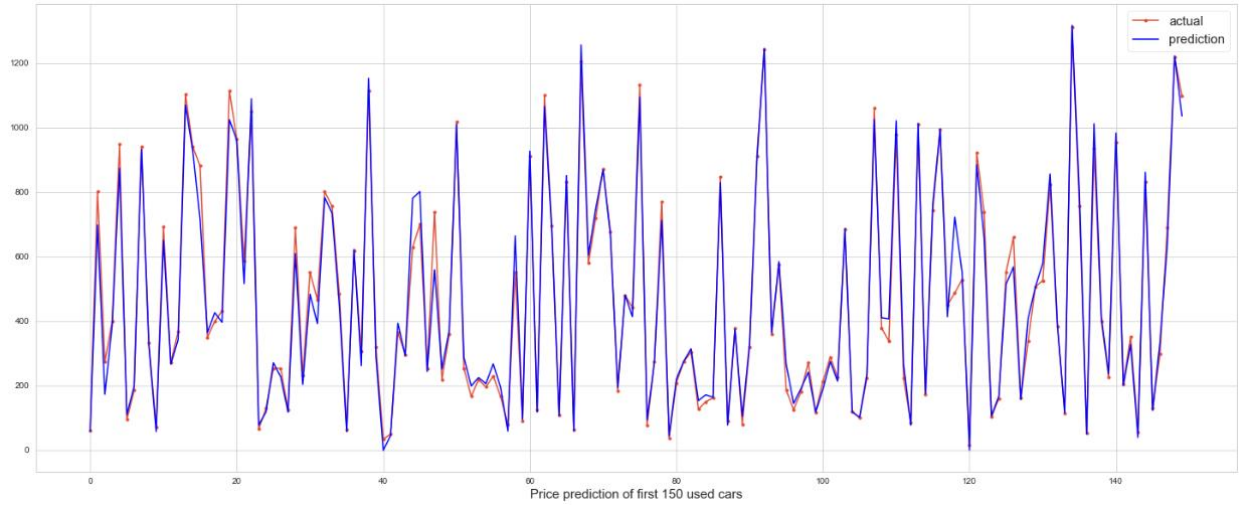
The Error Table of “RandomForestRegressor”:

```
                Error Table
Mean Absolute Error      :  56.00352060070422
Mean Squared  Error      :  6240.918616942188
Root Mean Squared  Error :  78.99948491567643
Accuracy on Traing set   :  0.9926342237172002
Accuracy on Testing set  :  0.948027067521619
```



The Error Table of “GradientBoostingRegressor”:

```
              Error Table
Mean Absolute Error      :  47.05285207381301
Mean Squared  Error      :  4775.906080512007
Root Mean Squared  Error :  69.10793066292759
Accuracy on Traing set   :  0.9947381034525249
Accuracy on Testing set  :  0.9602273544199561
```



3.DISCUSSION:

After careful consideration of all Regression models with their R-squared, mean squared error (MSE), or root mean squared error (RMSE), we consider “**GradientBoostingRegressor**” gives the BEST accuracy of **96.2%** for test data and **99.1%** for train-data. The model is said to be better fit to any prediction if it has “**LOW**” R-squared, mean squared error (MSE), or root mean squared error (RMSE). All regressors which are trained and analyzied above has “**HIGH**” R-squared, mean squared error (MSE), or root mean squared error (RMSE) than “**GradientBoostingRegressor**”. So, “**GradientBoostingRegressor”** model is best fit for our used-car price prediction.

4.LIMITATIONS:

While conducting the analysis and regression modeling, there are several limitations that should be acknowledged:

i. **Data Quality:** Ensure accurate and representative data for reliable analysis.
ii. **Missing Data:** Handle missing values appropriately to avoid biased estimates.
iii. **Model Assumptions:** Validate assumptions for the reliability of the regression model.

iv. **Causality and Confounding Variables:** Consider additional factors that influence prices beyond mileage to avoid biased estimates.
v. **Generalizability:** Interpret findings within the context of the data and consider variations across different markets or customer segments.

If given the opportunity to redo the project or continue working on it, the following steps could be taken:

i. **Data Enrichment:** Include additional variables to improve predictive power and accuracy.
ii. **Market Segmentation:** Analyze data separately for different customer segments or regions for targeted pricing strategies.
iii. **Advanced Modeling Techniques:** Explore advanced techniques to capture complex patterns in the data.
iv. **Continuous Data Collection:** Continuously update data analysis to stay informed about market trends.

5.CONCLUSION:

The best accuracy of 96.2% for test data and 99.1% for train-data is obtained by CarDataSet_Train & CarDataSet_Test datasets. Being a sophisticated model, GradientBoostingRegressor gives the BEST accuracy in comparison to all prior works using these datasets.

6.ADDITIONAL WORK:

Keeping the current model as a baseline, we intend to use some advanced techniques algorithms to predict car prices as our future work. We intend to develop a fully automatic, interactive system that contains a repository of used-cars with their prices. This enables a user to know the price of a similar car using a recommendation engine, which we would work in the future.