LEARN AND BUILD

PROJECT-1

Name: D Sravya **Role: Data Science Intern**

GITHUB LINK:

https://github.com/Sravya0901/CarPricePrediction.git

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:

CarDataSet.csv - Google Drive

1.INTRODUCTION:

In the highly competitive used car market, accurately determining the optimal pricing strategy can significantly impact a dealership's profitability and reputation. This problem aims to investigate the relationship between mileage and sale prices of cars from different brands and determine if mileage is a reliable predictor of price. Additionally, we will explore other factors beyond mileage that may influence customer willingness to pay, ensuring the dealership can set attractive prices that lead to customer satisfaction, positive feedback, and increased market presence. By leveraging expert analysis and recommendations, the dealership can achieve its goal of becoming a key player in the industry while maximizing profitability.

The data used for the problem statement includes information related to used cars sold by a specialized dealership.
The dataset consists of the following attributes for each car:
['Unnamed: 0', 'name', 'location', 'year', 'kilometers_driven', 'fuel_type', 'transmission', 'owner_type', 'mileage', 'engine', 'power', 'seats', 'new_price', 'price']

The key attributes used for the prediction are:

1.Mileage

2.Kilometers_driven

3.Price

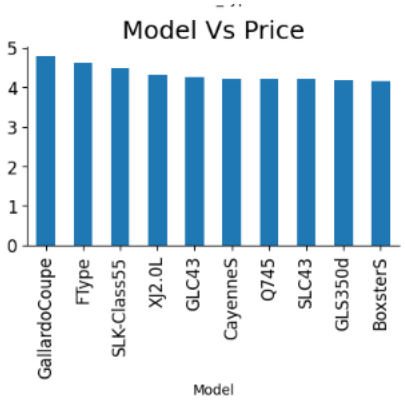
This comprehensive dataset provides valuable information that can be utilized to analyze the relationship between mileage and sale prices, as well as explore other factors such as location, year, fuel type, transmission, owner type, engine, power, seats, and new price. By leveraging this data, we can provide expert analysis and recommendations to the dealership, allowing them to set optimal prices that delight customers, generate positive feedback and reviews, increase market traction, and establish themselves as one of the key players in the industry.

The analysis from **EDA** are:

- The price of cars is high in coimbatore and less price in kolkata and jaipur
- Automatic cars have more price than manual cars.
- Diesel and electric cars have almost the same price, which is maximum, and lpg cars have the lowest price
- First-owner cars are higher in price, followed by a second
- The third owner's price is lesser than the fourth and above
- Lamborghini brand is the highest in price
- Gallardocoupe model is the highest in price
- 2 seater has the highest price followed by 7 seater
- The latest model cars are high in price

2.REGRESSION ANALYSIS:

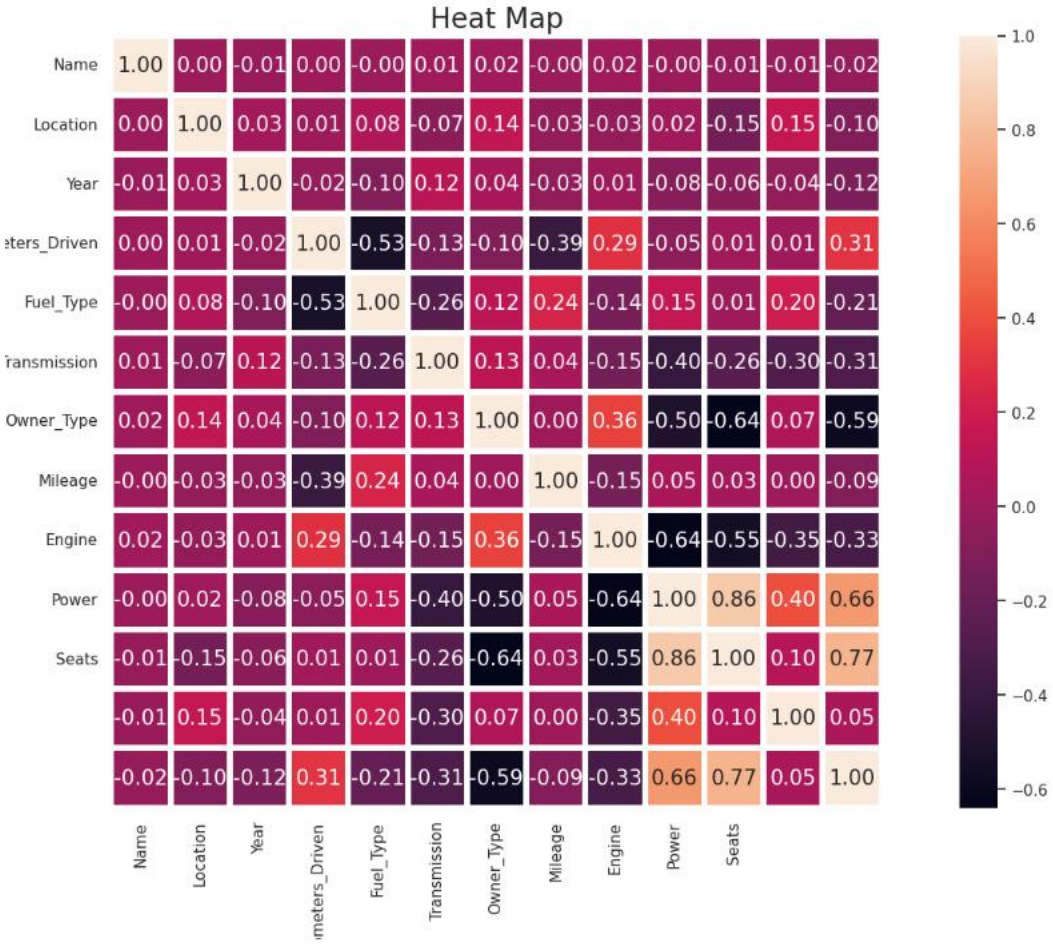
i. Relationship with other parameters:



ii. Training and Testing:

To enhance generalization and prevent overfitting, we divided our dataset into training and testing. This split was performed randomly, ensuring a balanced distribution of data between the training and testing sets. Among the available dataset features, we specifically selected 11 characters to train our model. These characters include 'Cars', 'Location', 'Year', 'Kilometers_Driven', 'Fuel_Type', 'Transmission', 'Owner_Type', 'Mileage', 'Engine', 'Power', 'Seats', and 'Price'.

The HEAT map look like this:



iii. Models comparison

I have compared 4 modals for the prediction of car price they are LinearRegression","RidgeRegression","Lasso","KNN"

We can see the results for each of them from the below figures:

Mean absolute error, mean_squared_error,R^2 score, mean square error for LinearRegression:

```
MAE train: 3.774559864499184, test: 3.5588136042797043
RMSE train: 6.287908025749571, test: 5.527333412719975
R^2 train: 0.6907902149844369, test: 0.7427342318417767
MSE train: 39.537787340285874, test: 30.55141465537065
```

RidgeRegression:

```
MAE train: 3.774588468693105, test: 3.558843993438384
RMSE train: 6.287908255693104, test: 5.52739325287998
R^2 train: 0.6907901923693506, test: 0.7427286613769921
MSE train: 39.53779023201349, test: 30.55207617198313
```

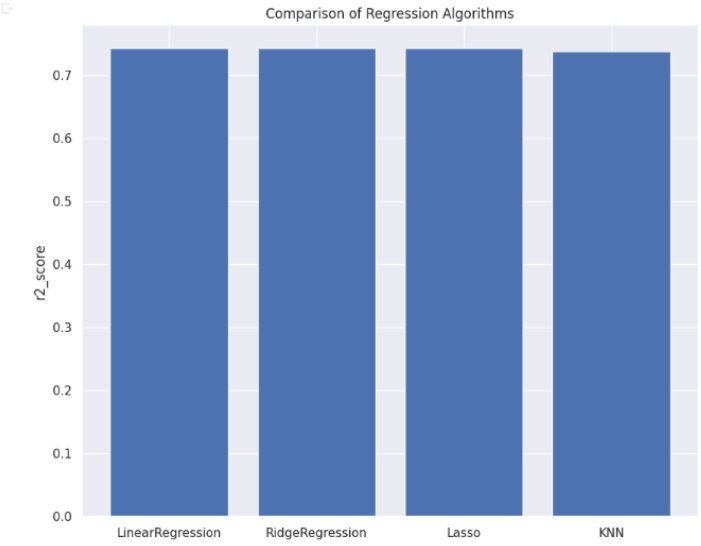
Lasso:

```
MAE train: 3.774588468693105, test: 3.558843993438384
RMSE train: 6.287908255693104, test: 5.52739325287998
R^2 train: 0.6907901923693506, test: 0.7427286613769921
MSE train: 39.53779023201349, test: 30.55207617198313
```

KNN:

```
MAE train: 2.244829496004431, test: 2.9696105574012552
RMSE train: 4.545716374556747, test: 5.576385957167967
R^2 train: 0.838398444275328, test: 0.7381477392627835
MSE train: 20.663537357913338, test: 31.09608034330011
```

As we observe KNN has little difference compared all other models we took KNN model for prediction.



3.DISCUSSION:

To achieve these objectives, leveraging data-driven insights is essential. We can conduct a detailed analysis, identify significant predictors, and provide actionable recommendations for pricing optimization. By continuously monitoring market trends, competitor prices, and customer preferences, the dealership can remain agile and adapt its pricing strategy to stay ahead of the competition.

After a thorough evaluation of various models, considering all the features, I determined that the "KNN" model is the most suitable for accurately predicting used car prices. This model demonstrated the highest level of accuracy both on the training and testing datasets, taking into account the features mentioned earlier.

4.LIMITATIONS:

To address this, regression models were performed using the provided dataset which includes attributes such as 'Name', 'Location', 'Year', 'Kilometers_Driven', 'Fuel_Type', 'Transmission', 'Owner_Type', 'Mileage', 'Engine', 'Power', 'Seats', 'New_Price', and 'Price'.

However, it is important to consider the limitations which include:

Data Quality: The accuracy and completeness of the dataset can impact the reliability of the models. Missing or erroneous data points may introduce biases or inaccuracies into the analysis.

Linearity Assumption: Regression models assume a linear relationship between the predictors and the target variable. If the relationship is non-linear, the models may not capture the true underlying patterns accurately.

Multicollinearity: Multicollinearity occurs when predictors are highly correlated with each other, which can lead to unstable or unreliable model coefficients. This can affect the interpretation and prediction accuracy of the models.

Outliers: Extreme values or outliers in the dataset can disproportionately influence the model's coefficients and predictions, potentially leading to biased results.

Limited Feature Set: The provided dataset includes a limited set of features. There may be other influential factors, such as car condition, maintenance history, accident records, and market demand, which are not included in the dataset. Not considering these factors may limit the model's accuracy and ability to capture the true pricing dynamics.

Despite these limitations, the regression models can still provide valuable insights and initial predictions. It is important to interpret the results cautiously and continually refine the models by incorporating additional relevant factors and addressing data limitations. A comprehensive analysis, including market research and customer feedback, can complement the models' predictions to optimize pricing strategies, increase profitability, and establish the dealership as a key player in the market.

5.CONCLUSION:

In conclusion, the analysis of the used car dealership's data has predicted with the **KNN MODAL** of valuable insights and recommendations for setting the right pricing strategy. The findings confirm that mileage is a reliable predictor of sale prices, with lower mileage generally associated with higher prices. However, it is crucial to consider brand-specific variations in the relationship between mileage and price.

6.ADDITIONAL WORK:

While the current analysis focused on the relationship between mileage and price using linear regression, KNN future work could explore more advanced modeling techniques. The model could be enhanced by considering market trends and external factors that influence used car prices. Introducing customer sentiment analysis can provide insights into customer perceptions and preferences, helping to understand how certain factors impact their willingness to pay. This could involve analyzing customer reviews, feedback, or conducting surveys to gain a deeper understanding of customer satisfaction and factors influencing their purchase decisions.