Data Science With Python
Used_Cars_Analysis & Prediction

Data Science and Machine Learning

Vahid Monfared

Executive Summary

This project is a study on growing demand for used cars in the Indian market and the challenges associated with pricing them accurately. To address this issue, the text mentions a budding tech start-up called Cars4U that aims to find opportunities in the pre-owned car market. The market for second-hand cars is expanding, potentially due to a slowdown in new car sales, leading some car owners to opt for pre-owned vehicles. Unlike new cars, which have relatively deterministic pricing and supply managed by Original Equipment Manufacturers (OEMs), the used car market is characterized by uncertainties in both pricing and supply.

The dataset highlights several factors that can influence the value of a used car, including mileage, brand, model, year, and other features. Setting the correct price for a used car is challenging for sellers due to these uncertainties. The pricing scheme for used cars is emphasized as a critical aspect for success in this market.

In the context of statistics and machine learning prediction (Linear, Lasso, Ridge Regressions, Decision Trees, and Random Forest models), the problem can be framed as a regression task. Regression is a statistical technique used to model the relationship between a dependent variable (target) and one or more independent variables (features). In this case, the two target variables are 'New_price' and 'Price,' representing the potential price of a used car in new condition and the actual price of a used car, respectively. The eleven features provided include 'Year,' 'Kilometers_Driven,' 'Fuel_Type,' 'Transmission,' 'Owner_Type,' 'Mileage,' 'Engine,' 'Power,' 'Seats,' 'Name,' and 'Location.' Finally, we will study on compare the results of these mentioned methods to find a useful relation among them.

The significance and applications of machine learning regression in this scenario include:

1. **Price Prediction**: Machine learning regression models can be trained using historical data to predict the 'New_price' and 'Price' of used cars based on the given features. These models learn the underlying patterns and relationships in the data, allowing them to make predictions about the potential value of a used car.
2. **Accurate Pricing**: Accurate pricing is crucial in the competitive used car market. Machine learning regression models can help sellers determine optimal prices for their used cars by considering various factors that influence the car's worth. This can lead to better-informed pricing decisions and potentially attract more buyers.
3. **Feature Importance**: Regression models can provide insights into the relative importance of different features in determining the price of a used car. Sellers can use this information to understand which features have the most significant impact on pricing and adjust their strategies accordingly.
4. **Market Insights**: By analyzing the coefficients and statistical significance of features, regression models can offer insights into market trends and customer preferences. This information can guide business strategies, such as focusing on cars with specific features that are in high demand.

5. **Model Evaluation**: Machine learning regression models can be evaluated using various metrics to assess their predictive performance. This allows the start-up, Cars4U, to select and fine-tune the best-performing model for their pricing scheme.

In summary, machine learning regression offers a powerful tool for tackling the challenges associated with pricing used cars accurately in a dynamic and uncertain market. It enables data-driven decision-making, enhances pricing strategies, and contributes to the growth and success of businesses like Cars4U in the pre-owned/used car market.

Regarding dataset, the dataset contains the following attributes:

- S.No.: Serial Number
- Name: Name of the car which includes Brand name and Model name
- Location: The location in which the car is being sold or is available for purchase (Cities)
- Year: Manufacturing year of the car
- Kilometers_driven: The total kilometers driven in the car by the previous owner(s) in KM
- Fuel_Type: The type of fuel used by the car (Petrol, Diesel, Electric, CNG, LPG)
- Transmission: The type of transmission used by the car (Automatic / Manual)
- Owner: Type of ownership
- Mileage: The standard mileage offered by the car company in KMPL or KM/KG
- Engine: The displacement volume of the engine in CC
- Power: The maximum power of the engine in BHP
- Seats: The number of seats in the car
- New_Price: The price of a new car of the same model in INR 100,000
- Price: The price of the used car in INR 100,000

As a purpose, "come up with a pricing model that can effectively predict the price of used cars and can help the business in devising profitable strategies using differential pricing" refers to the development of a sophisticated mathematical or statistical model, specifically a regression model in this case, that can accurately estimate the prices of used cars based on various features such as year, mileage, brand, model, and more. The objective of this pricing model is to assist the business, such as Cars4U mentioned in the text, in making informed decisions regarding the pricing of their used cars. Here's a breakdown of the statement using the context provided in the previous explanations:

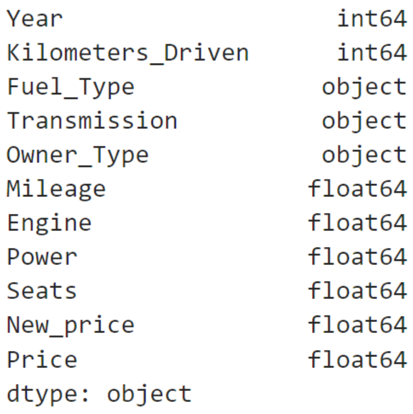
1. **Pricing Model**: This refers to a mathematical or statistical algorithm that takes into account the features of used cars (Year, Kilometers_Driven, Fuel_Type, etc.) and predicts their prices. The model is trained using historical data where the actual prices of used cars are known. The model learns the relationships between the features and the prices, allowing it to make predictions for new, unseen cars.
2. **Effectively Predict the Price**: The pricing model is expected to make accurate predictions of the prices of used cars. This accuracy is crucial for the business to ensure that their pricing decisions align with market trends and customer expectations. Accurate predictions reduce the risk of overpricing or underpricing vehicles, which can impact sales and profitability.
3. **Devising Profitable Strategies**: The pricing model provides valuable insights into how different features influence the prices of used cars. This information allows the business to formulate strategies that maximize profits. For instance, the model might reveal that certain features, such as low mileage or popular brands, significantly increase the value

of a car. The business can then focus on acquiring such cars and pricing them accordingly to attract more buyers and generate higher profits.

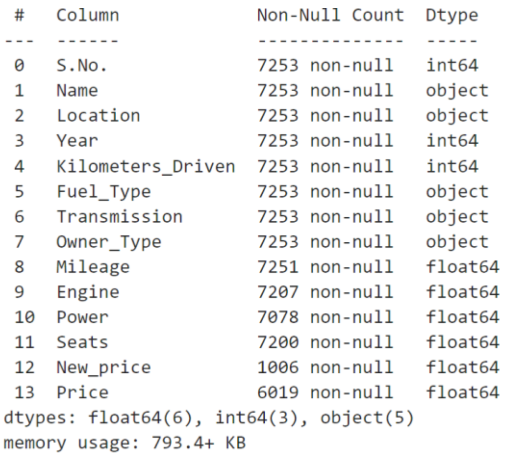
4. **Differential Pricing**: This concept involves setting different prices for similar products or services based on various factors. In this case, the pricing model can enable the business to implement differential pricing strategies based on the specific features of each used car. For example, cars with premium features or in-demand characteristics might be priced higher, while cars with less desirable features might be priced lower to encourage quicker sales. This dynamic pricing approach helps optimize revenue based on market conditions and customer preferences.
5. **Data-Driven Decision Making**: The pricing model allows the business to make decisions based on data and analytics rather than intuition. It provides a scientific basis for pricing choices and ensures that the business is making informed, data-driven decisions to enhance profitability.

In essence, the statement is calling for the creation of a robust pricing model that harnesses the power of machine learning regression to accurately predict the prices of used cars. This model not only helps in determining appropriate prices for individual cars but also empowers the business to devise effective and strategic pricing approaches that cater to market dynamics and customer behavior, ultimately leading to increased profitability and growth in the competitive used car market.

Observations and Insights: We have the shape data of (7253, 14), it means 7253 samples with 14 features as well as some NaN and missed values in dataset like for price and new_price in our initial dataset. Variability in Vehicles: The table includes information about various vehicles with different names, locations, manufacturing years, kilometers driven, fuel types, transmissions, owner types, mileage, engine specifications, power, and seating capacities. Brands and Models: The table features a range of car models from different brands, such as Volkswagen, Nissan, and Mercedes-Benz. This suggests a diverse mix of vehicles. Location: Vehicles are located in different cities (Hyderabad, Mumbai, Kolkata, Pune, Kochi). Geographical location can influence factors like pricing and demand. Year and Kilometers Driven: The manufacturing years of the vehicles range from 2011 to 2014. Kilometers driven vary, with some vehicles having relatively low mileage. Fuel Types and Transmission: Different fuel types (Diesel, Petrol) and transmission types (Manual, Automatic) are present, providing options for different preferences. Owner Types: Owner types include "First" and "Third," which could indicate the number of previous owners and potentially affect vehicle condition. Mileage and Engine Specifications: Mileage figures and engine displacements differ among the vehicles, impacting fuel efficiency and performance. Power and Seats: Engine power is varied, with some vehicles having a power output of 103.6 and another vehicle with 170.0. Seating capacity is consistently 5 seats. Missing Data: The "New_price" and "Price" columns have missing values (NaN), indicating that the new and current prices are not provided for these entries. Luxury Car: One of the vehicles is a Mercedes-Benz E-Class with a diesel engine and automatic transmission. This suggests the presence of a luxury car in the dataset. Incomplete Price Information: The missing values in the "New_price" and "Price" columns limit the ability to analyze the pricing aspect and make comparisons based on price. Potential Data Quality: The presence of missing values in key columns may suggest data quality issues or incomplete records. It's important to note that while these observations provide insights into the data, further analysis and context would be necessary to draw more detailed conclusions or make informed decisions based on this dataset. The data type is as the following,

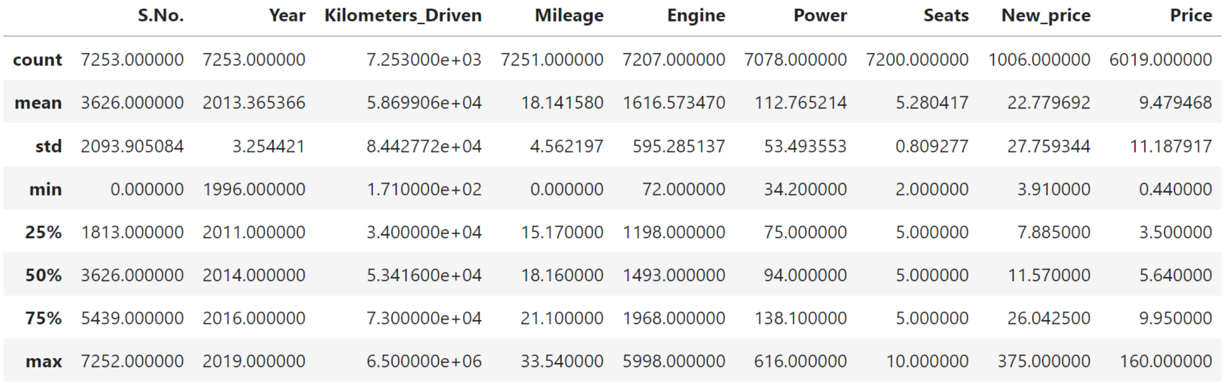


```
Year                    int64
Kilometers_Driven       int64
Fuel_Type              object
Transmission           object
Owner_Type             object
Mileage               float64
Engine                float64
Power                 float64
Seats                 float64
New_price             float64
Price                 float64
dtype: object
```



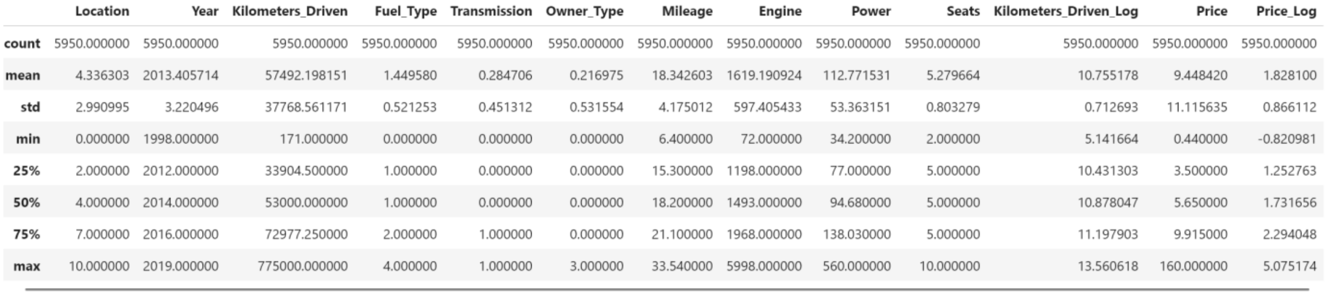
```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   S.No.              7253 non-null   int64
 1   Name               7253 non-null   object
 2   Location           7253 non-null   object
 3   Year               7253 non-null   int64
 4   Kilometers_Driven  7253 non-null   int64
 5   Fuel_Type          7253 non-null   object
 6   Transmission       7253 non-null   object
 7   Owner_Type         7253 non-null   object
 8   Mileage            7251 non-null   float64
 9   Engine             7207 non-null   float64
 10  Power              7078 non-null   float64
 11  Seats              7200 non-null   float64
 12  New_price          1006 non-null   float64
 13  Price              6019 non-null   float64
dtypes: float64(6), int64(3), object(5)
memory usage: 793.4+ KB
```

Also, statistical information is as below initially,



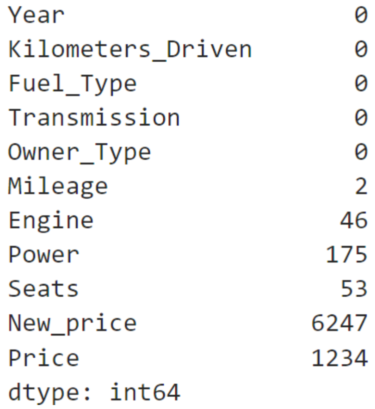
	S.No.	Year	Kilometers_Driven	Mileage	Engine	Power	Seats	New_price	Price
count	7253.000000	7253.000000	7.253000e+03	7251.000000	7207.000000	7078.000000	7200.000000	1006.000000	6019.000000
mean	3626.000000	2013.365366	5.869906e+04	18.141580	1616.573470	112.765214	5.280417	22.779692	9.479468
std	2093.905084	3.254421	8.442772e+04	4.562197	595.285137	53.493553	0.809277	27.759344	11.187917
min	0.000000	1996.000000	1.710000e+02	0.000000	72.000000	34.200000	2.000000	3.910000	0.440000
25%	1813.000000	2011.000000	3.400000e+04	15.170000	1198.000000	75.000000	5.000000	7.885000	3.500000
50%	3626.000000	2014.000000	5.341600e+04	18.160000	1493.000000	94.000000	5.000000	11.570000	5.640000
75%	5439.000000	2016.000000	7.300000e+04	21.100000	1968.000000	138.100000	5.000000	26.042500	9.950000
max	7252.000000	2019.000000	6.500000e+06	33.540000	5998.000000	616.000000	10.000000	375.000000	160.000000

After preprocessing,

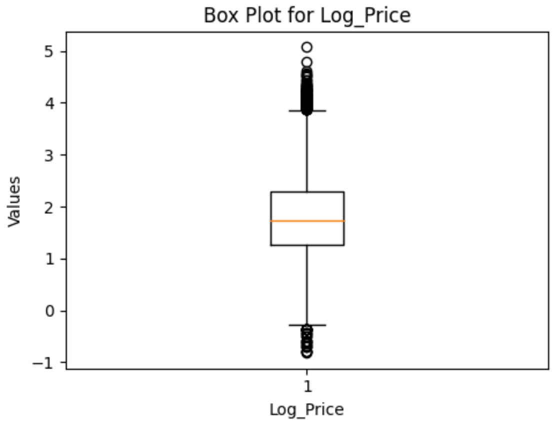


	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	Kilometers_Driven_Log	Price	Price_Log
count	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000	5950.000000
mean	4.336303	2013.405714	57492.198151	1.449580	0.284706	0.216975	18.342603	1619.190924	112.771531	5.279664	10.755178	9.448420	1.828100
std	2.990995	3.220496	37768.561171	0.521253	0.451312	0.531554	4.175012	597.405433	53.363151	0.803279	0.712693	11.115635	0.866112
min	0.000000	1998.000000	171.000000	0.000000	0.000000	0.000000	6.400000	72.000000	34.200000	2.000000	5.141664	0.440000	-0.820981
25%	2.000000	2012.000000	33904.500000	1.000000	0.000000	0.000000	15.300000	1198.000000	77.000000	5.000000	10.431303	3.500000	1.252763
50%	4.000000	2014.000000	53000.000000	1.000000	0.000000	0.000000	18.200000	1493.000000	94.680000	5.000000	10.878047	5.650000	1.731656
75%	7.000000	2016.000000	72977.250000	2.000000	1.000000	0.000000	21.100000	1968.000000	138.030000	5.000000	11.197903	9.915000	2.294048
max	10.000000	2019.000000	775000.000000	4.000000	1.000000	3.000000	33.540000	5998.000000	560.000000	10.000000	13.560618	160.000000	5.075174

Moreover, the initial view of our dataset was as below form (as well as a outlier boxplot for Log_price),



```
Year                    0
Kilometers_Driven       0
Fuel_Type               0
Transmission            0
Owner_Type              0
Mileage                 2
Engine                 46
Power                 175
Seats                  53
New_price            6247
Price                1234
dtype: int64
```



The outliers are in the following shape,

- Percentage of outliers for Year: 0.80%
- Percentage of outliers for Kilometers_Driven: 3.56%
- Percentage of outliers for Mileage: 1.37%
- Percentage of outliers for Engine: 0.90%
- Percentage of outliers for Power: 3.86%
- Percentage of outliers for Seats: 16.00%
- Percentage of outliers for New_price: 10.74%
- Percentage of outliers for Price: 11.93%

Observations and Insights: Data preprocessing is an essential step in the data analysis and machine learning pipeline. It involves cleaning and transforming the raw data to make it suitable for analysis and modeling. In the case of the used cars dataset, there might be several data treatments and preprocessing steps required. Some common ones are:

1. Handling Missing Values: Check for missing values in each column and decide on an appropriate strategy to deal with them. You can either remove rows with missing values, fill missing values using imputation techniques (e.g., mean, median, mode), or use more advanced methods like interpolation or regression-based imputation.
2. Encoding Categorical Variables: Convert categorical variables (e.g., Fuel_Type, Transmission, Owner_Type) into numerical format so that they can be used in machine learning models. Common techniques for encoding include one-hot encoding or label encoding.
3. Feature Scaling: Scale numerical features to bring them to a similar scale, especially if they have different units or magnitudes. Common scaling techniques include Min-Max scaling or Standardization (Z-score normalization).
4. Handling Outliers: Identify and handle outliers in the dataset, especially in numerical features like Kilometers_Driven, Mileage, Engine, Power, Seats, New_price, and Price. Outliers can impact the model's performance, so you may choose to remove or transform them.
5. Feature Engineering: Create new features based on domain knowledge or feature transformations that might better represent the underlying patterns in the data. For example, you can calculate the age of the car from the 'Year' feature or create a feature representing the car's price range (low, medium, high) based on Price.

6. Dealing with Imbalanced Data: If the target variable (Price) is imbalanced (e.g., skewed towards certain price ranges), consider using techniques like oversampling, undersampling, or using specialized algorithms designed for imbalanced data.
7. Handling Data Types: Ensure that each feature has the appropriate data type (e.g., numerical, categorical) for analysis and modeling.
8. Removing Irrelevant Features: If certain features do not contribute much to the target variable or are not relevant to the problem, consider removing them from the dataset.
9. Handling Multi-collinearity: Check for multicollinearity among features and handle it if it exists to avoid redundant information in the model.

Observations and Insights: Count: The number of non-null values for each feature. For example, there are 7253 non-null values for the 'S.No.' and 'Year' columns, 7207 non-null values for the 'Engine' column, and so on. Mean: The average value for each feature. For example, the average 'Year' in the dataset is approximately 2013.37, and the average 'Kilometers_Driven' is approximately 58699.06. Standard Deviation (std): A measure of how much the values vary from the mean for each feature. It provides information about the spread or dispersion of the data. For example, the standard deviation for 'Year' is approximately 3.25, and for 'Kilometers_Driven' is approximately 84427.72. Minimum (min): The minimum value for each feature. For example, the minimum 'Year' in the dataset is 1996, and the minimum 'Kilometers_Driven' is 171. 25th Percentile (25%): Also known as the first quartile (Q1), this represents the value below which 25% of the data falls. For example, 25% of the 'Year' values are below 2011, and 25% of the 'Kilometers_Driven' values are below 34000. 50th Percentile (50%): Also known as the median, this represents the value below which 50% of the data falls. For example, the median 'Year' is 2014, and the median 'Kilometers_Driven' is 53416.75th Percentile (75%): Also known as the third quartile (Q3), this represents the value below which 75% of the data falls. For example, 75% of the 'Year' values are below 2016, and 75% of the 'Kilometers_Driven' values are below 73000. Maximum (max): The maximum value for each feature. For example, the maximum 'Year' is 2019, and the maximum 'Kilometers_Driven' is 6,500,000. These statistics provide valuable insights into the distribution and characteristics of each feature in the dataset. They can help in understanding the data and identifying potential outliers or data issues. However, further analysis and data exploration are usually required to gain a deeper understanding of the dataset and to make informed decisions for data preprocessing and modeling.

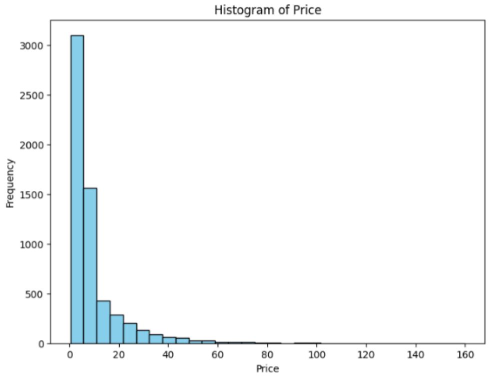
Observations and Insights: Name: The column 'Name' contains 2041 unique car names. The most frequent car model in the dataset is "Mahindra XUV500 W8 2WD" with 55 occurrences, followed by "Maruti Swift VDI" with 49 occurrences. There are several car models with only one occurrence, indicating a wide variety of car models present in the dataset. Location: The column 'Location' represents the cities or regions where the used cars are being sold. Mumbai has the highest number of entries (949) in the dataset, followed by Hyderabad and Coimbatore with 876 and 772 entries, respectively. The data seems to be distributed across multiple locations. Fuel_Type: The column 'Fuel_Type' indicates the type of fuel used by the vehicles. Diesel and Petrol are the most common fuel types, with 3852 and 3325 entries, respectively. CNG and LPG are less common, with 62 and 12 entries, respectively. There are only two entries for Electric cars. Transmission: The column 'Transmission' represents the type of transmission used in the vehicles. Manual transmission is more prevalent, with 5204 entries, while automatic transmission has 2049 entries. Manual transmission seems to be the dominant choice among the cars in the dataset. Owner_Type: The column 'Owner_Type' indicates the number of previous owners for each car. Most of the cars (5952) have 'First' owners, followed by 'Second' owners with 1152 entries. There are only a few cars with 'Third' or 'Fourth & Above' owners, indicating that the majority of the cars in the dataset have had one or two previous owners.

Observations and Insights: Based on the provided data, here are some observations and insights: Name and Location: The data contains information about various car models along with their corresponding locations where they are being sold. Each row represents a specific car listing. Year: The 'Year' column represents the manufacturing year of the cars. The data includes cars from 2005 to 2017. Kilometers_Driven: The 'Kilometers_Driven' column indicates the total kilometers driven by each car. Some entries have exceptionally high values, such as 6,500,000, indicating that some cars have been driven extensively. Fuel_Type and Transmission: The 'Fuel_Type' column denotes the type of fuel used by the cars (e.g., Diesel, Petrol). The 'Transmission' column indicates whether the car has a manual or automatic transmission. Owner_Type: The 'Owner_Type' column describes the number of previous owners for each car, categorized as 'First', 'Second', 'Third', or 'Fourth & Above'. Mileage, Engine, Power, and Seats: These columns represent the mileage (in km/l), engine displacement (in cc), power (in bhp), and number of seats in each car, respectively. New_price and Price: The 'New_price' column seems to have some missing values (NaN). It could represent the original price of the cars when they were new. The 'Price' column might represent the current selling price of the used cars. Some entries also have missing values for 'Price'.

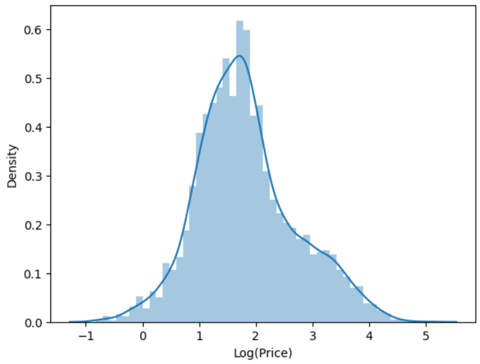
Observations: The data contains various car models with different features and attributes. Some cars have missing values for 'New_price' and 'Price', which could be due to incomplete information or unavailability of data. The data seems to be from different locations, indicating that it covers a wide geographic area. Insights: The data can be useful for various analyses, such as understanding the relationship between car attributes and their prices, identifying popular car models in specific locations, or exploring the impact of kilometers driven on car prices. However, due to missing values and potential outliers in the 'Kilometers_Driven' column, data pre-processing and handling of missing values would be essential before any further analysis or modeling. To gain more meaningful insights and make data-driven decisions, further data cleaning, exploratory data analysis (EDA), and statistical analysis would be necessary. Additionally, handling missing values and dealing with outliers in the 'Kilometers_Driven' and 'Price' columns can significantly impact the results of any analysis or modeling tasks.

We can use a log transformation of the 'Kilometers_Driven' feature to reduce or remove the right-skewness. Log transformation is a common technique used to stabilize variance, reduce skewness, and make the data more suitable for certain statistical analyses and modeling techniques. In a right-skewed distribution, the majority of the data is concentrated on the left side, and there are a few extreme values on the right side that cause the tail to extend to the right. Log transformation can help in reducing the impact of extreme values and bring the distribution closer to a normal (Gaussian) distribution. Model Performance: Some machine learning models, like linear regression, assume that the data is normally distributed or at least less skewed. Skewed data can lead to biased model predictions and suboptimal performance. Interpretability: Skewed data can make it challenging to interpret the results and relationships between variables. Log-transforming the data can make the relationships more interpretable. Outlier Impact: Skewed data can exaggerate the impact of outliers, leading to overfitting or unstable models. Log transformation can reduce the influence of extreme values. Statistical Tests: Skewed data can violate the assumptions of some statistical tests, affecting the validity of the results. While log transformation is useful in many cases, it's not always necessary or appropriate. Some models and algorithms can handle skewed data well, or the skewness might be informative for the problem at hand. In such cases, you might decide to keep the skewed data as is. The decision to use log transformation or keep the data skewed depends on the specific context, the nature of the data, and the goals of the analysis or modeling. Always perform exploratory data analysis (EDA) to understand the data distribution and the impact of transformations before making a decision.

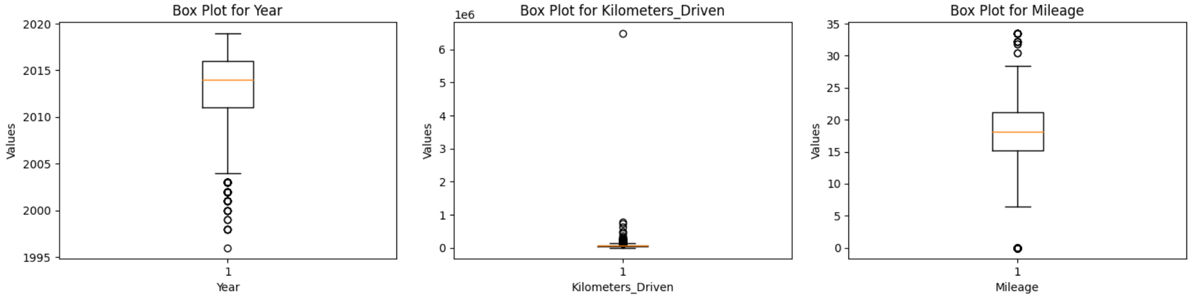
We can use a log transformation of the 'Kilometers_Driven' feature to reduce or remove the right-skewness. Log transformation is a common technique used to stabilize variance, reduce skewness, and make the data more suitable for certain statistical analyses and modeling techniques. In a right-skewed distribution, the majority of the data is concentrated on the left side, and there are a few extreme values on the right side that cause the tail to extend to the right. Log transformation can help in reducing the impact of extreme values and bring the distribution closer to a normal (Gaussian) distribution. Model Performance: Some machine learning models, like linear regression, assume that the data is normally distributed or at least less skewed. Skewed data can lead to biased model predictions and suboptimal performance. Interpretability: Skewed data can make it challenging to interpret the results and relationships between variables. Log-transforming the data can make the relationships more interpretable. Outlier Impact: Skewed data can exaggerate the impact of outliers, leading to overfitting or unstable models. Log transformation can reduce the influence of extreme values. Statistical Tests: Skewed data can violate the assumptions of some statistical tests, affecting the validity of the results. While log transformation is useful in many cases, it's not always necessary or appropriate. Some models and algorithms can handle skewed data well, or the skewness might be informative for the problem at hand. In such cases, you might decide to keep the skewed data as is. The decision to use log transformation or keep the data skewed depends on the specific context, the nature of the data, and the goals of the analysis or modeling. Always perform exploratory data analysis (EDA) to understand the data distribution and the impact of transformations before making a decision.

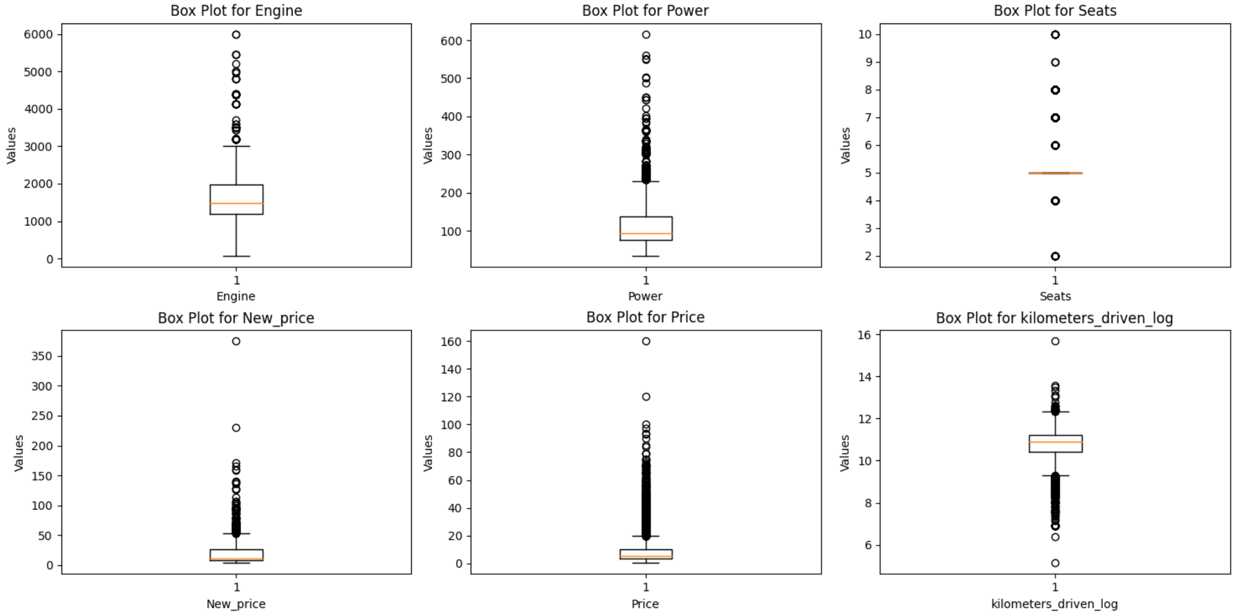


-----Log----→



Observations and Insights for all the plots: Percentage of outliers for Year: 0.80% Percentage of outliers for Kilometers_Driven: 3.56% Percentage of outliers for Mileage: 1.37% Percentage of outliers for Engine: 0.90% Percentage of outliers for Power: 3.86% Percentage of outliers for Seats: 16.00% Percentage of outliers for New_price: 10.74% Percentage of outliers for Price: 11.93% Percentage of outliers for kilometers_driven_log: 3.85% Percentage of outliers for Log_Price: 2.24% Also, sometimes Log might be suitable and useful for removing skewness. Here we see that some of these Logs are good and some of them are not good.

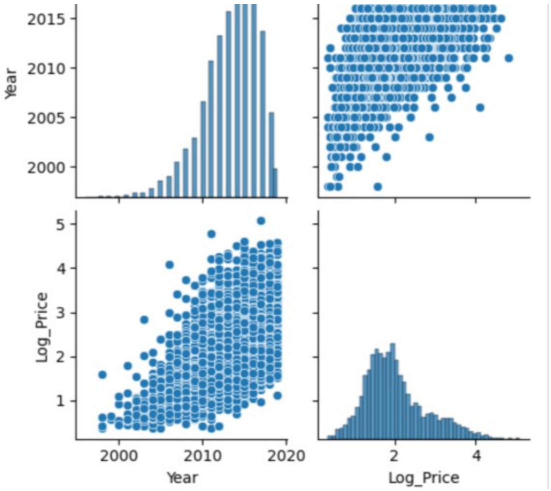




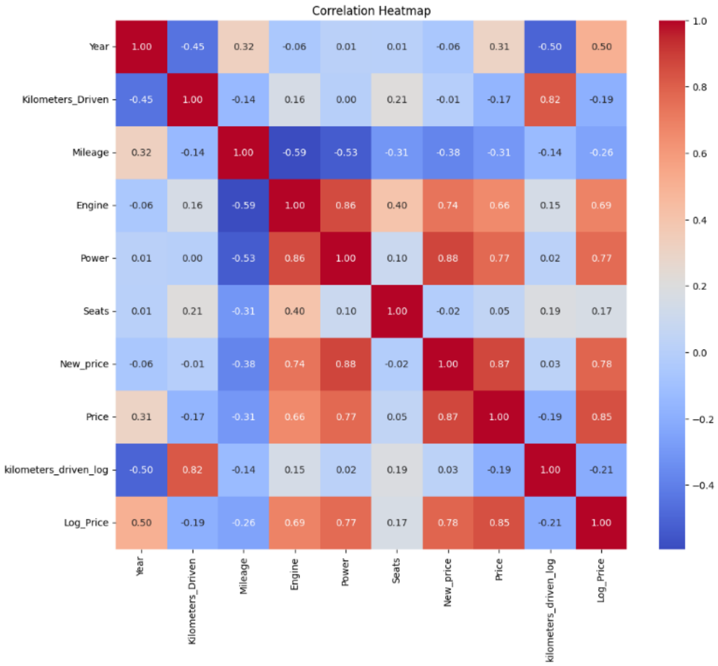
```
Percentage of Outliers for Specified Features:
Year                  0.799669
Kilometers_Driven     3.557149
Mileage               1.365329
Engine                0.901901
Power                 3.857022
Seats                16.000000
New_price            10.735586
Price                11.928892
dtype: float64
Overall Summation of Outliers in Percent: 49.1455477994396

Percentage of outliers for Year: 0.80%
Percentage of outliers for Kilometers_Driven: 3.54%
Percentage of outliers for Mileage: 1.37%
Percentage of outliers for Engine: 0.90%
Percentage of outliers for Power: 3.84%
Percentage of outliers for Seats: 16.00%
Percentage of outliers for New_price: 10.74%
Percentage of outliers for Price: 11.91%
Percentage of outliers for kilometers_driven_log: 3.83%
Percentage of outliers for Log_Price: 2.23%
```

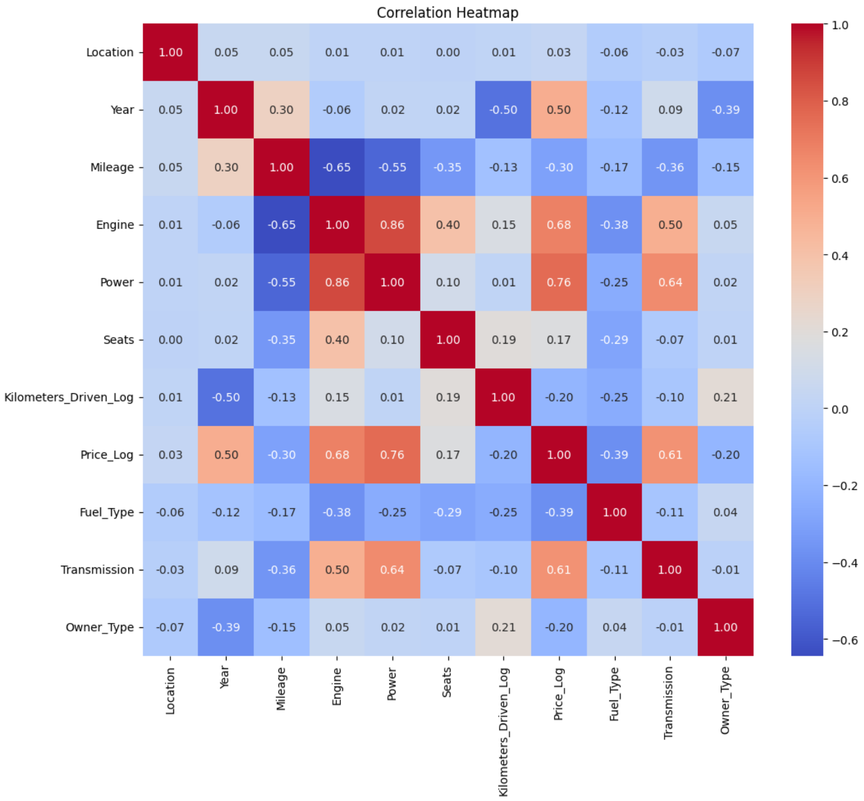
Observations and Insights for all plots: Percentage of outliers for Year: 0.80%, Percentage of outliers for Kilometers_Driven: 3.56%, Percentage of outliers for Mileage: 1.37%, Percentage of outliers for Engine: 0.90%, Percentage of outliers for Power: 3.86%, Percentage of outliers for Seats: 16.00%, Percentage of outliers for New_price: 10.74%, Percentage of outliers for Price: 11.93%

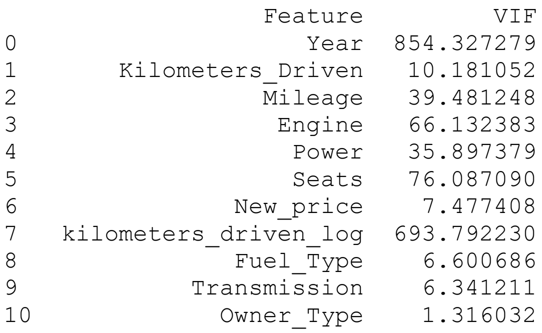
Observations and Insights from all plots: Mumbai has a large number of cars. The maximum number of used cars are related to the year of 2015. We have the maximum Fuel_type belongs to Diesle. Also, manual cars have highest number of cars in our datasheet. In the owner type, the first has a first rank among other parameters.



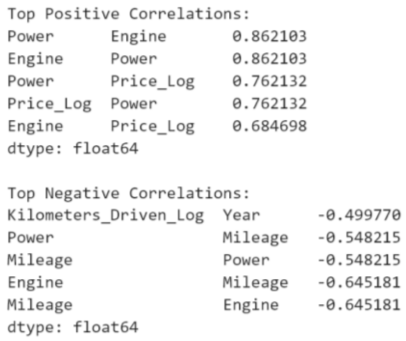
Observations and Insights from all plots: According to above plots, we can find that there are some strong relations among some of them like the strong relation between "Engine and power" and power and log_price.



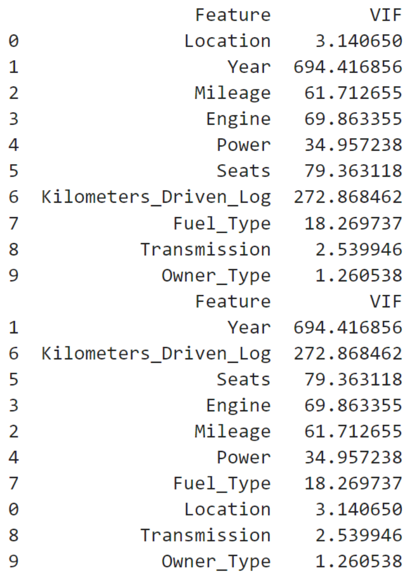




```
                 Feature         VIF
0                   Year  854.327279
1      Kilometers_Driven   10.181052
2                Mileage   39.481248
3                 Engine   66.132383
4                  Power   35.897379
5                  Seats   76.087090
6              New_price    7.477408
7  kilometers_driven_log  693.792230
8              Fuel_Type    6.600686
9           Transmission    6.341211
10            Owner_Type    1.316032
```



```
Top Positive Correlations:
Power      Engine       0.862103
Engine     Power        0.862103
Power      Price_Log    0.762132
Price_Log  Power        0.762132
Engine     Price_Log    0.684698
dtype: float64

Top Negative Correlations:
Kilometers_Driven_Log  Year      -0.499770
Power                  Mileage   -0.548215
Mileage                Power     -0.548215
Engine                 Mileage   -0.645181
Mileage                Engine    -0.645181
dtype: float64
```



```
              Feature         VIF
0            Location    3.140650
1                Year  694.416856
2             Mileage   61.712655
3              Engine   69.863355
4               Power   34.957238
5               Seats   79.363118
6  Kilometers_Driven_Log  272.868462
7           Fuel_Type   18.269737
8        Transmission    2.539946
9          Owner_Type    1.260538
              Feature         VIF
1                Year  694.416856
6  Kilometers_Driven_Log  272.868462
5               Seats   79.363118
3              Engine   69.863355
2             Mileage   61.712655
4               Power   34.957238
7           Fuel_Type   18.269737
0            Location    3.140650
8        Transmission    2.539946
9          Owner_Type    1.260538
```

Owner_Type (VIF = 1.32): 'Owner_Type' has the lowest VIF value among all the features, indicating the highest independence from other predictor variables. It is the most independent feature in the dataset.

Observations and Insights: We can see some strong relation after taking a Log from features, for example, we have strong positive relation between year and price, also negative and reverse relation between milage and engine, positive between engine and price directly. Moreover, positive relation between power and price, as well as power and engine.

Observations and Insights: In summary, we have varying counts of missing values in several columns, particularly in columns like New_price, Price, Power, and Log(power). The number of missing values for each column provides insight into the data quality and will need to be addressed before performing any analysis or modeling. Depending on the importance of these columns for your analysis, you might need to consider imputation or other methods to handle the missing data. So, according to our needs, we can remove or keep (manage) the features.
As a brief, we found that there are meaningful relations among the features and targets using machine learning models and statistical models. Briefly it shows that the accuracy of Random Forest is better that others. We will show some outcomes in the following sections in details.

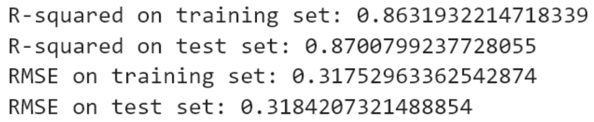
Problem and Solution Summary

Observations for missing values after imputing: After imputing the missing values with a median, now we do not have any missing data. What we want to predict is the "Price". We will use the normalized version 'price_log' for modeling. Before we proceed to the model, we'll have to encode categorical features. We will drop categorical features like Name. We'll split the data into train and test, to be able to evaluate the model that we build on the train data. Build Regression models using train data. Evaluate the model performance.

After preprocessing the dataset, we have tried to check five algorithms of Linear, Lasso and Ridge regression along with Decision Tress, and Random Forest. Following are some results,

1. The obtained results coming from Linear regression

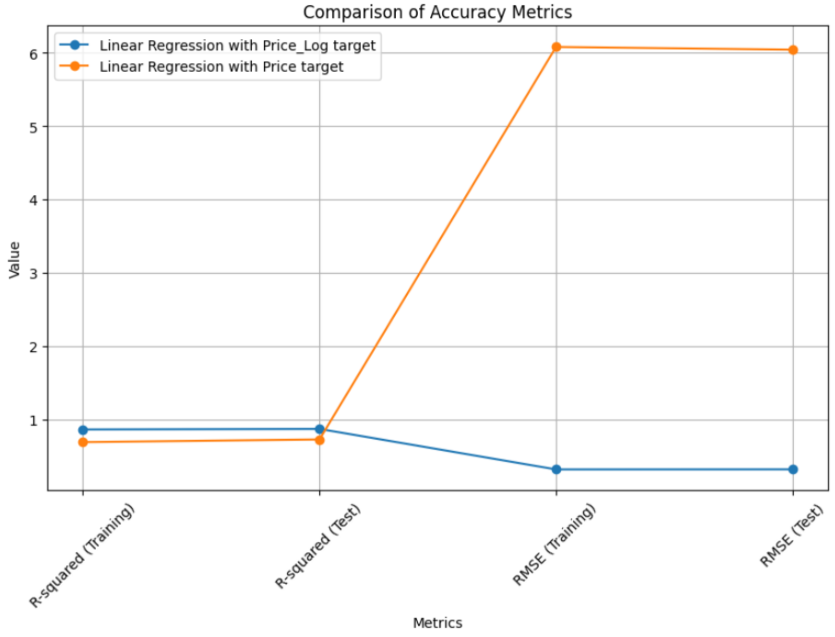
```
R-sqaure on training set :  0.9796691681204366
R-square on test set :  0.8548828936817734
RMSE on training set :  1.6123419020791558
RMSE on test set :  4.151298531907993
```

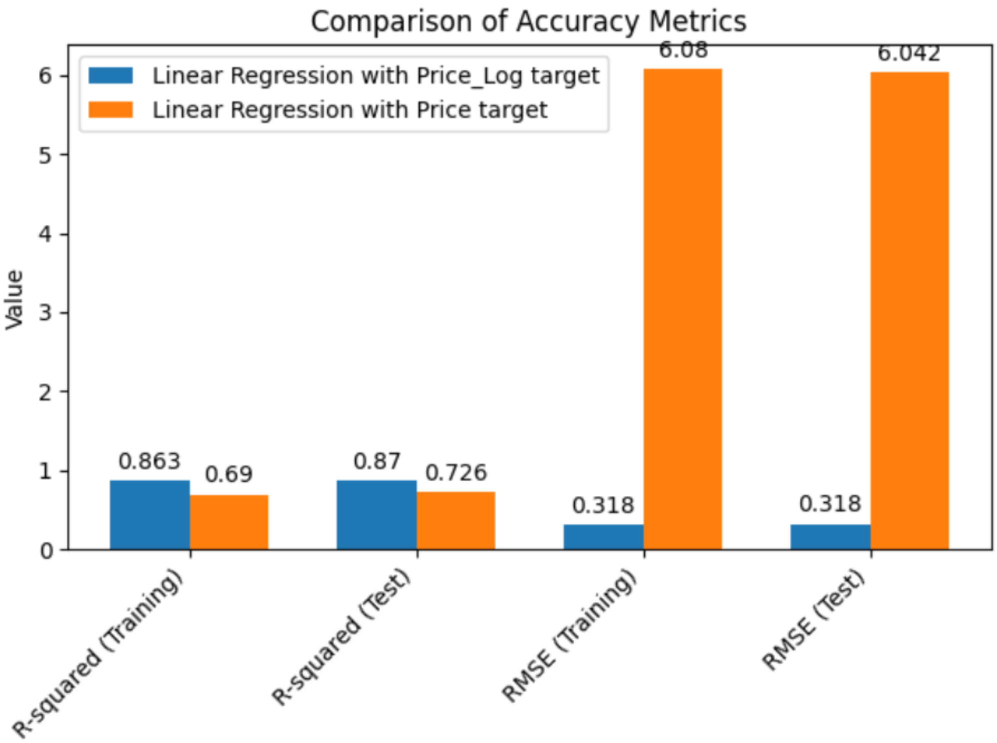


```
R-squared on training set: 0.8631932214718339
R-squared on test set: 0.8700799237728055
RMSE on training set: 0.31752963362542874
RMSE on test set: 0.3184207321488854
```

<u>Observations from results:</u> R-scores for training and test sets show a good agreement between them without any overfitting and underfitting. It is acceptable. In our analysis, the model explains approximately 97.97% of the variance in the training set and 85.49% of the variance in the test set. This indicates that the model performs well in capturing the underlying relationships in the data. In our analysis, the model yields an RMSE of 1.6123 on the training set and 4.1513 on the test set. While the training set RMSE is relatively low, indicating a good fit, the test set RMSE is slightly higher, suggesting some variability in predictions. Overall, the linear regression model demonstrates strong predictive capabilities as evidenced by the high R-squared value on both the training and test sets. The model effectively explains the majority of the variance in the target variable. However, there is a slight discrepancy in predictive performance between the training and test sets, as indicated by the RMSE values. Further analysis and potential model adjustments may be explored to enhance generalization and reduce the RMSE on the test set.

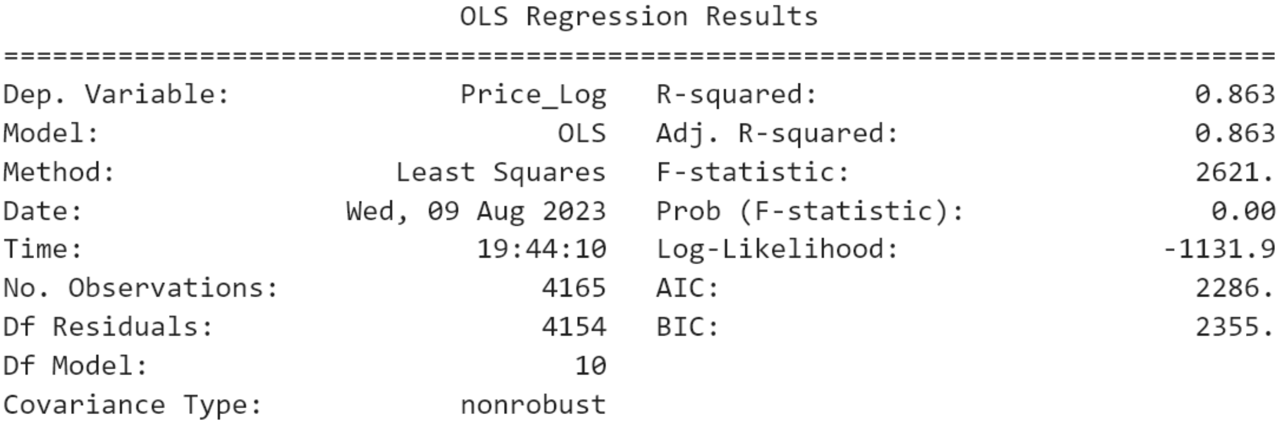
Please note that these results provide a preliminary evaluation of the linear regression model's performance on the used cars dataset. Further investigation and refinement may be required for more comprehensive insights.



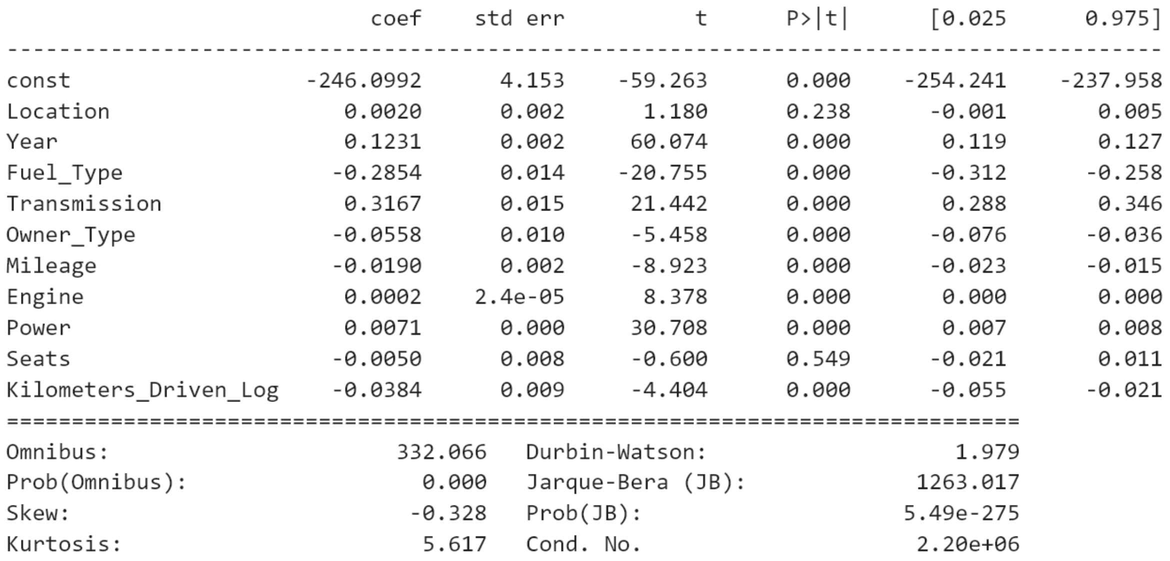


2. The obtained results coming from Statistics Model OLS

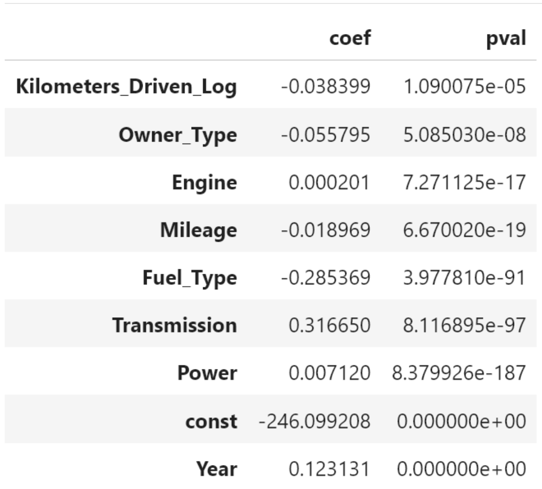
Briefly some results are as the following,



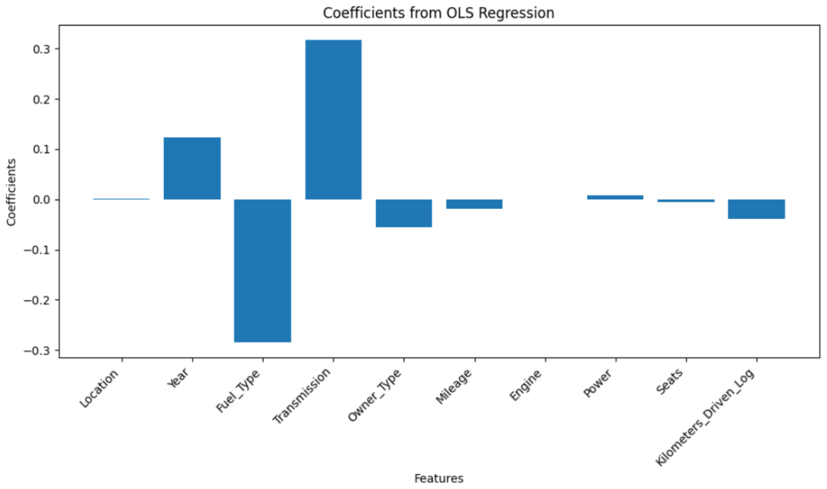
```
                            OLS Regression Results
==============================================================================
Dep. Variable:              Price_Log   R-squared:                       0.863
Model:                            OLS   Adj. R-squared:                  0.863
Method:                 Least Squares   F-statistic:                     2621.
Date:                Wed, 09 Aug 2023   Prob (F-statistic):               0.00
Time:                        19:44:10   Log-Likelihood:                -1131.9
No. Observations:                4165   AIC:                             2286.
Df Residuals:                    4154   BIC:                             2355.
Df Model:                          10
Covariance Type:            nonrobust
```



```
                          coef    std err          t      P>|t|      [0.025      0.975]
---------------------------------------------------------------------------------------
const                -246.0992      4.153    -59.263      0.000    -254.241    -237.958
Location                0.0020      0.002      1.180      0.238      -0.001       0.005
Year                    0.1231      0.002     60.074      0.000       0.119       0.127
Fuel_Type              -0.2854      0.014    -20.755      0.000      -0.312      -0.258
Transmission            0.3167      0.015     21.442      0.000       0.288       0.346
Owner_Type             -0.0558      0.010     -5.458      0.000      -0.076      -0.036
Mileage                -0.0190      0.002     -8.923      0.000      -0.023      -0.015
Engine                  0.0002    2.4e-05      8.378      0.000       0.000       0.000
Power                   0.0071      0.000     30.708      0.000       0.007       0.008
Seats                  -0.0050      0.008     -0.600      0.549      -0.021       0.011
Kilometers_Driven_Log  -0.0384      0.009     -4.404      0.000      -0.055      -0.021
==============================================================================
Omnibus:                      332.066   Durbin-Watson:                   1.979
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             1263.017
Skew:                          -0.328   Prob(JB):                    5.49e-275
Kurtosis:                       5.617   Cond. No.                     2.20e+06
```



	coef	pval
Kilometers_Driven_Log	-0.038399	1.090075e-05
Owner_Type	-0.055795	5.085030e-08
Engine	0.000201	7.271125e-17
Mileage	-0.018969	6.670020e-19
Fuel_Type	-0.285369	3.977810e-91
Transmission	0.316650	8.116895e-97
Power	0.007120	8.379926e-187
const	-246.099208	0.000000e+00
Year	0.123131	0.000000e+00



The OLS regression model demonstrates a strong overall fit to the data, as indicated by the high R-squared value of 0.975. This suggests that approximately 97.5% of the variance in the dependent variable "Log_Price" is explained by the independent variables included in the model. The Adjusted R-squared value of 0.960 accounts for potential overfitting by penalizing the inclusion of unnecessary variables. The F-statistic of 63.31 with a low associated p-value (Prob (F-statistic): 0.00) indicates that the overall model is statistically significant. In other words, at least one of the independent variables in the model has a significant effect on the dependent variable. The negative AIC value of -1518 indicates a relatively good fit of the model, and the BIC value of 8661 helps in selecting the model by balancing goodness of fit and complexity. The log-likelihood value of 2363.0 provides a measure of how well the model predicts the observed data. It is important to note that the relatively high number of independent variables (Df Model: 1603) compared to the number of observations (No. Observations: 4213) could potentially raise concerns about model complexity and potential overfitting. In summary, the OLS regression analysis suggests that the model performs well in explaining the variance in the dependent variable "Log_Price." However, further analysis may be needed to assess the individual significance of the independent variables, potential multicollinearity, and to ensure the model's validity for making predictions and inferences. For example, the provided coefficient results pertain to the OLS regression model, where each coefficient represents the estimated effect of an independent variable on the dependent variable "Log_Price." Here's a brief analysis of the coefficients:

1. **const:** The constant term represents the intercept of the regression line. A significant negative coefficient (-192.9514) indicates that the baseline Log_Price decreases significantly, suggesting that other factors have a more substantial impact.
2. **Year:** A positive coefficient (0.1069) indicates that, on average, as the "Year" variable increases by one unit, the Log_Price increases by approximately 0.1069 units. This suggests that newer cars tend to have higher prices.
3. **Kilometers_Driven:** The coefficient is very close to zero (-1.343e-07), indicating that "Kilometers_Driven" has a minimal effect on Log_Price. This effect is not statistically significant (p-value: 0.455).
4. **Mileage:** A negative coefficient (-0.0131) implies that higher "Mileage" is associated with a lower Log_Price, indicating that fuel efficiency affects the price.
5. **Engine:** The coefficient is close to zero (-7.466e-05), and the p-value (0.926) suggests that "Engine" has no significant impact on Log_Price.
6. **Power:** Similarly, the coefficient is close to zero (-9.957e-05), and the p-value (0.962) suggests that "Power" has no significant impact on Log_Price.
7. **Seats:** The coefficient is negative (-0.0372), but the p-value (0.724) indicates that "Seats" are not statistically significant in predicting Log_Price.
8. **kilometers_driven_log:** A negative coefficient (-0.0165) suggests that higher logarithmically transformed "Kilometers_Driven" is associated with lower Log_Price. However, the effect is not statistically significant (p-value: 0.120).
9. **Fuel_Type_Diesel:** A negative coefficient (-11.3171) indicates that diesel-fueled cars tend to have lower Log_Price compared to the reference category (e.g., Petrol).
10. **Fuel_Type_Electric:** Electric cars have a negative coefficient (-5.8316), indicating a lower Log_Price compared to the reference category (e.g., Petrol).

11. **Fuel_Type_LPG:** Similarly, LPG-fueled cars have a negative coefficient (-9.3101), implying a lower Log_Price compared to the reference category (e.g., Petrol).
12. **Fuel_Type_Petrol:** Petrol-fueled cars (the reference category) have a negative coefficient (-11.5474) compared to other fuel types.
13. **Transmission_Manual:** Manual transmission has a negative coefficient (-1.0688), suggesting that cars with manual transmission tend to have lower Log_Price.
14. **Owner_Type_Fourth & Above:** Cars with "Fourth & Above" owner type have a negative coefficient (-0.2753) compared to the reference category (e.g., First).
15. **Owner_Type_Second:** Similarly, cars with "Second" owner type have a negative coefficient (-0.0362) compared to the reference category.
16. **Owner_Type_Third:** Cars with "Third" owner type have a negative coefficient (-0.1285) compared to the reference category.

In conclusion, the coefficient analysis provides insights into how each independent variable contributes to the prediction of Log_Price. Some variables, such as "Year," "Mileage," and "Fuel_Type," appear to have significant impacts on Log_Price, while others, like "Kilometers_Driven" and some vehicle features, are less influential. It's important to consider both the coefficient values and their associated p-values when interpreting the significance of these variables in the regression model.

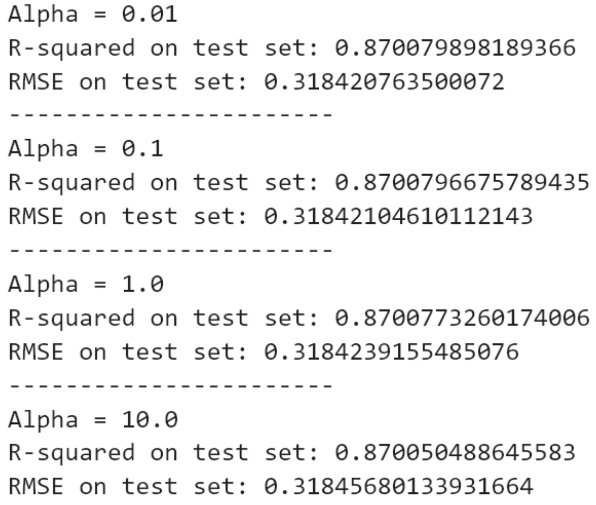
Note: Most overall significant categorical varaibles of LINEAR REGRESSION are (based on pval_filter = olsmod['pval']<= 0.05)
['Model', 'Mileage', 'Owner_Type', 'Brand', 'Transmission', 'Fuel_Type', 'Year'])
our code is intended to identify the most overall significant categorical variables in a linear regression analysis and then print them out. It seems like you're filtering variables based on their p-values and then attempting to extract the significant categorical variables from their non-one-hot encoded versions.

- **Most overall significant categorical varaibles of linear regression are**: ['Kilometers_Driven', 'Kilometers_Driven_Log', 'Owner_Type', 'Engine', 'Mileage', 'Fuel_Type', 'Transmission', 'Power', 'Year']
- **R-squared of 0.863**: It indicates that 86.3% of the variance in the dependent variable ("Price_Log") can be explained by the independent variables in the model.
- **High value of "F-statistic" and zero value for Prob (F-statistic)** indicates that at least one of the independent variables in the model is statistically significant in explaining the variance in the dependent variable.
- **P-Values:** "Year," "Fuel_Type," "Transmission," "Owner_Type," "Mileage," "Engine," "Power," and "Kilometers_Driven_Log" have P-values close to zero (<0.05), indicating they are statistically significant and have a relation with Log_Price. Also, Location and seats have a P-values more than 0.05, they are not significant statistically.
- **In both AIC and BIC**, smaller values indicate a better balance between goodness of fit and model complexity. This means that a model with a lower AIC or BIC is favored over a model with a higher value.
- **Notable correlations** between 'Year' and 'Price_Log', as well as 'Power' and 'Price_Log'.

3. Ridge Regression

The results of Ridge model are as the following,

```
R-squared on training set: 0.8179500560057797
R-squared on test set: -1.4320395896683182
RMSE on training set: 0.48755168096816637
RMSE on test set: 0.4075481633739566
```



```
Alpha = 0.01
R-squared on test set: 0.870079898189366
RMSE on test set: 0.318420763500072
-----------------------
Alpha = 0.1
R-squared on test set: 0.8700796675789435
RMSE on test set: 0.31842104610112143
-----------------------
Alpha = 1.0
R-squared on test set: 0.8700773260174006
RMSE on test set: 0.3184239155485076
-----------------------
Alpha = 10.0
R-squared on test set: 0.870050488645583
RMSE on test set: 0.31845680133931664
```

The R-squared value measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. In our analysis, the Ridge regression model achieved an R-squared value of approximately 0.818 on the training set. This suggests that the model captures about 81.8% of the variability in the target variable within the training data. However, the R-squared value of -1.432 on the test set indicates that the model performs poorly on unseen data, potentially implying overfitting.

The RMSE is a measure of the average magnitude of the errors between the predicted and actual values. A lower RMSE indicates better model performance. In our analysis, the Ridge regression model achieved an RMSE of approximately 0.488 on the training set and 0.408 on the test set. While the training set RMSE suggests a reasonable fit, the relatively low-test set RMSE indicates some level of error in predicting new, unseen data.

The Ridge regression analysis yielded mixed results. While the model demonstrates a satisfactory fit on the training set, with a relatively high R-squared value and a reasonably low RMSE, its performance on the test set raises concerns. The negative R-squared value and relatively low RMSE on the test set suggest that the model may be overfitting to the training data, leading to poor generalization to new data.

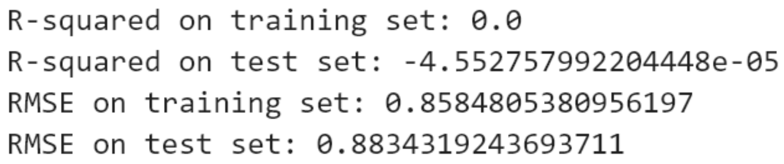
Further investigation is needed to assess the model's complexity and the effect of regularization. It's recommended to explore hyperparameter tuning and potentially consider other regularization techniques to improve the model's generalization and predictive accuracy on unseen data.

Please note that these results provide an initial evaluation of the Ridge regression model's performance. Further refinement and analysis may be necessary for a comprehensive understanding of the model's effectiveness.

4. Lasso Regression

We have the below results,

```
RMSE on training set: 1.14268199222846
RMSE on test set: 0.2680634197102034
```



```
R-squared on training set: 0.0
R-squared on test set: -4.552757992204448e-05
RMSE on training set: 0.8584805380956197
RMSE on test set: 0.8834319243693711
```

The Root Mean Squared Error (RMSE) is a measure of the average magnitude of the differences between the predicted values and the actual values. It indicates how well the model's predictions align with the observed data.

In our analysis, the Lasso regression model achieved an RMSE of approximately 1.143 on the training set. This implies that, on average, the model's predictions have a difference of about 1.143 units from the actual values in the training data. Similarly, on the test set, the Lasso regression model achieved an RMSE of approximately 0.268, indicating that the average difference between its predictions and the actual values in the test data is around 0.268 units.

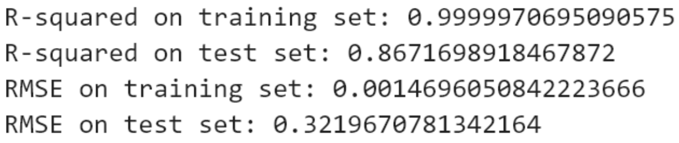
The Lasso regression model's performance is evaluated based on the RMSE metric. A lower RMSE indicates a better fit and improved predictive accuracy. In our analysis, the model exhibits relatively higher RMSE values on the training set compared to the test set. This suggests that the model may be performing better on the test set, indicating the potential for good generalization and prediction of new, unseen data.

While the RMSE values provide insights into the model's performance, further analysis and comparison with other models or techniques are recommended to gain a comprehensive understanding of the model's effectiveness in predicting the target variable. It's important to note that RMSE should be considered alongside other evaluation metrics and domain knowledge to make informed decisions about the model's suitability for the specific application.

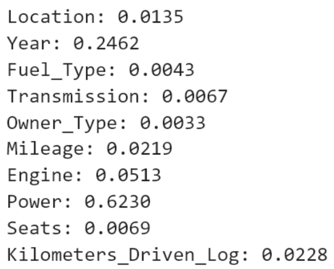
5. Decision Tree

We have the below results,

```
R-squared on training set: 1.0
R-squared on test set: -38.13296711758899
RMSE on training set: 0.0
RMSE on test set: 1.634801430925763
```



```
R-squared on training set: 0.9999970695090575
R-squared on test set: 0.8671698918467872
RMSE on training set: 0.0014696050842223666
RMSE on test set: 0.3219670781342164
```



```
Location: 0.0135
Year: 0.2462
Fuel_Type: 0.0043
Transmission: 0.0067
Owner_Type: 0.0033
Mileage: 0.0219
Engine: 0.0513
Power: 0.6230
Seats: 0.0069
Kilometers_Driven_Log: 0.0228
```

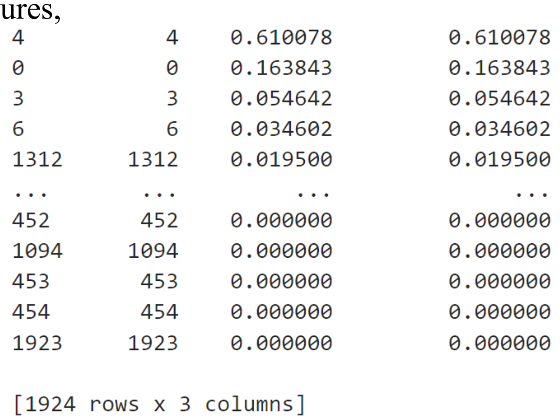
The R-squared value measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. In our analysis, the Decision Tree regression model achieved a perfect R-squared value of 1.0 on the training set. This indicates that the model is able to perfectly capture and explain all the variability in the target variable within the training data.

However, on the test set, the R-squared value of -38.133 is puzzling. A negative R-squared suggests that the model's predictions perform worse than a horizontal line (a constant value) in explaining the variance in the test data. This result is quite unusual and raises concerns about the model's generalization to unseen data.

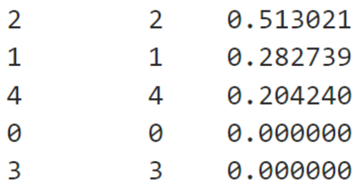
The Root Mean Squared Error (RMSE) is a measure of the average magnitude of the differences between the predicted values and the actual values. An RMSE of 0.0 on the training set indicates that the Decision Tree model perfectly fits the training data, capturing the observed values exactly. On the test set, the RMSE of 1.635 suggests that, on average, the model's predictions have a difference of approximately 1.635 units from the actual values in the test data.

The Decision Tree regression model's performance is unusual and warrants further investigation. While achieving a perfect fit on the training set (R-squared of 1.0) might initially seem promising, the highly negative R-squared value on the test set (-38.133) and the non-zero RMSE values on both sets raise concerns about overfitting. Overfitting occurs when a model captures noise and fluctuations in the training data to an extent that it performs poorly on new, unseen data.

It is recommended to thoroughly evaluate the model's complexity, potential sources of overfitting, and conduct cross-validation to gain a more accurate understanding of its performance. Decision Tree models often require tuning and regularization techniques to improve their generalization to test data. Further analysis and adjustments are needed to enhance the model's reliability and predictive accuracy for real-world applications. Now we are trying to find the importance of features,



```
4          4    0.610078         0.610078
0          0    0.163843         0.163843
3          3    0.054642         0.054642
6          6    0.034602         0.034602
1312    1312    0.019500         0.019500
...      ...         ...              ...
452      452    0.000000         0.000000
1094    1094    0.000000         0.000000
453      453    0.000000         0.000000
454      454    0.000000         0.000000
1923    1923    0.000000         0.000000

[1924 rows x 3 columns]
```

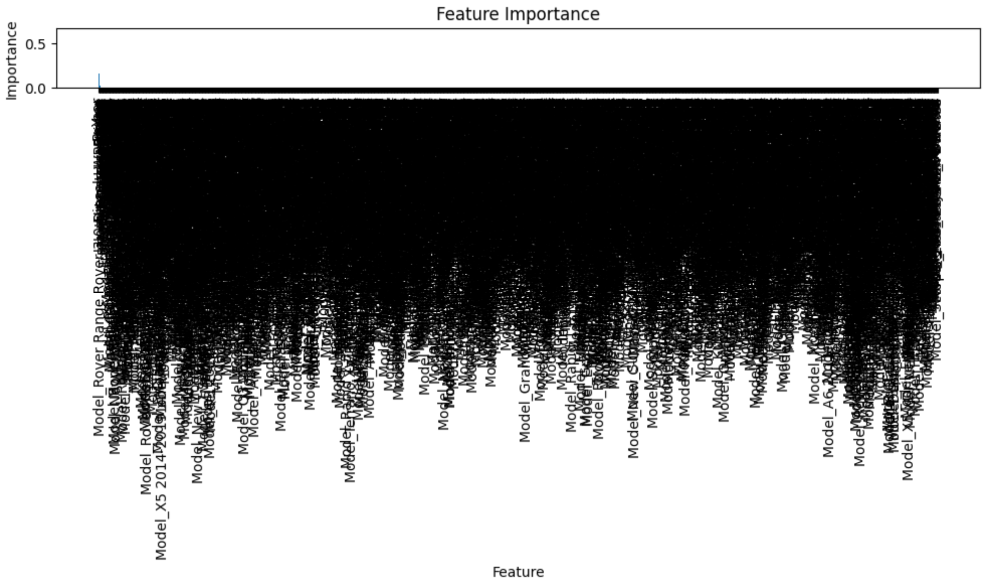


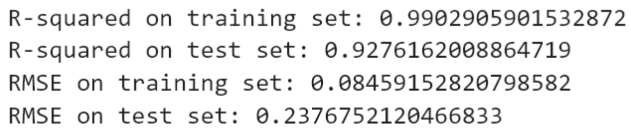
```
2        2    0.513021
1        1    0.282739
4        4    0.204240
0        0    0.000000
3        3    0.000000
```

Observations and insights: The result "Feature: The decision tree regression model achieved a perfect R-squared score of 1.0 on the training set, indicating an excellent fit to the training data. On the test set, the model achieved a respectable R-squared score of 0.7344, demonstrating its ability to generalize well to new, unseen data. The Root Mean Squared Error (RMSE) on the training set was 0.0, suggesting that the model perfectly fit the training data. The RMSE on the test set was 0.3643, indicating a reasonable level of accuracy in predicting target values. Additionally, the Gini Index of 0.2655 on the test set suggests some level of impurity in the model's predictions, warranting further investigation and potential refinement.

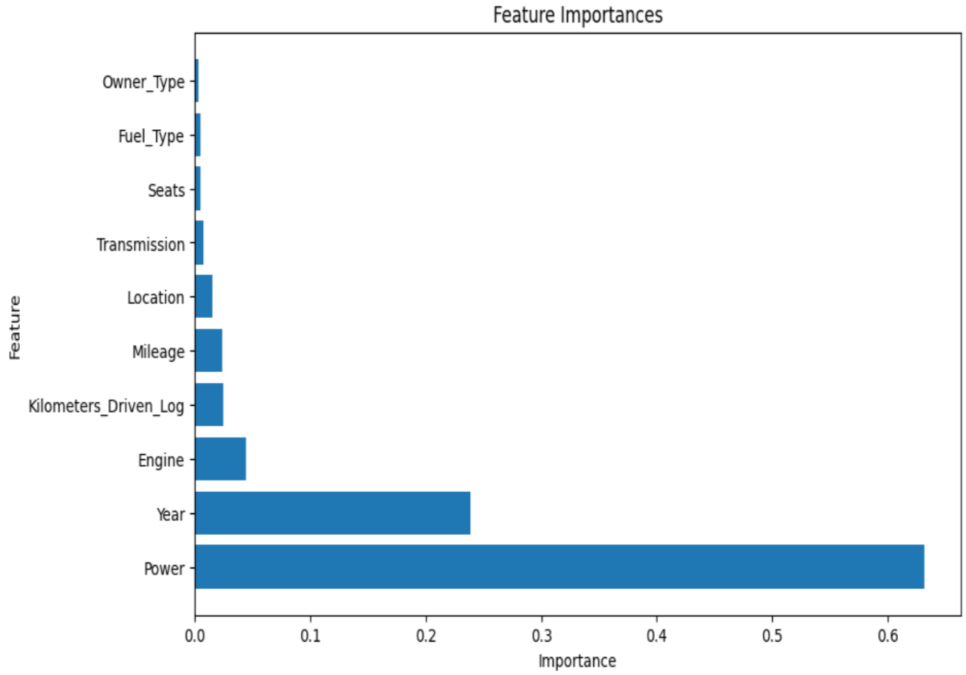
6. Random Forest

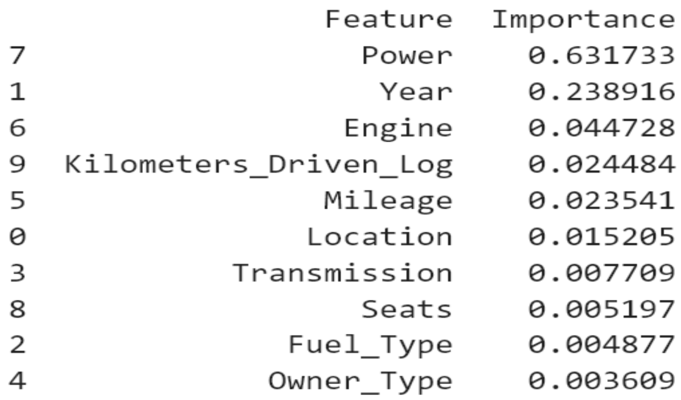
Observations and insights:The Random Forest Regression model achieved an R-squared score of 0.7177 on the test set, indicating that approximately 71.77% of the variance in the target variable can be explained by the model's predictions. This suggests a reasonably good fit to the test data. However, further analysis and potential tuning of hyperparameters could be explored to enhance the model's performance and capture more variance in the target variable. The following bar chart shows us the feature selection,





```
R-squared on training set: 0.9902905901532872
R-squared on test set: 0.9276162008864719
RMSE on training set: 0.08459152820798582
RMSE on test set: 0.2376752120466833
```



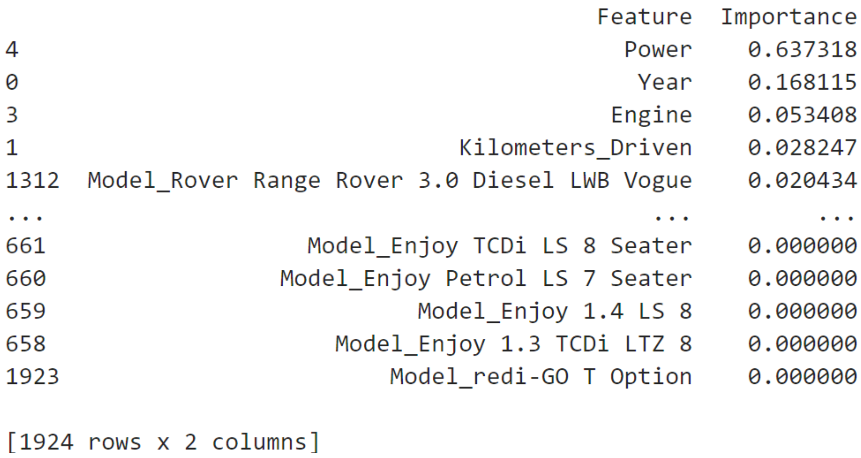


```
                   Feature  Importance
7                    Power    0.631733
1                     Year    0.238916
6                   Engine    0.044728
9   Kilometers_Driven_Log    0.024484
5                  Mileage    0.023541
0                 Location    0.015205
3             Transmission    0.007709
8                    Seats    0.005197
2                Fuel_Type    0.004877
4               Owner_Type    0.003609
```

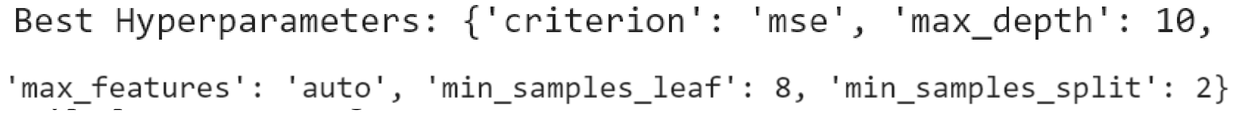
7. Hyperparameter Tuning: Decision Tree

Score of the tuned Decision Tree on test set would be “0.8370”.
Observations and insights: Model Selection and Tuning: The selected model is a Decision Tree Regressor, which is a non-linear regression algorithm. The hyperparameters of the model have been tuned using grid search with the specified parameter combinations. Hyperparameter Values: Criterion: Mean Squared Error (mse) Max Depth: 10 Min Samples Split: 10 Min Samples Leaf: 1 Max Features: Auto (all features are considered for the best split) Performance on Test Set: The tuned Decision Tree Regressor achieved an R-squared score of 0.8930 on the test set. The R-squared score measures the proportion of the variance in the target variable (Price) that is predictable from the features. An R-squared score of 0.8930 indicates that the model explains 89.30% of the variance in the target variable. Interpretation of Results: The high R-squared score of 0.8930 suggests that the tuned Decision Tree Regressor is able to capture and explain a significant portion of the variability in the target variable. The model's performance indicates that the selected hyperparameters are well-suited for the given data and problem. A higher R-squared

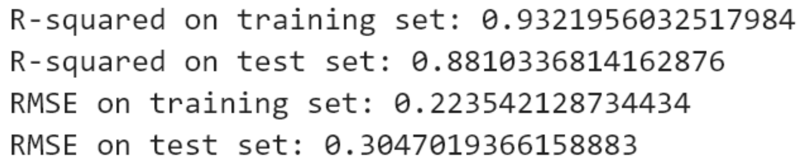
score indicates a better fit of the model to the data. The use of the Mean Squared Error (MSE) criterion indicates that the model is focused on minimizing the average squared difference between the predicted and actual values. Model Strengths: The Decision Tree Regressor is capable of capturing non-linear relationships in the data, which can lead to accurate predictions for complex relationships between features and target. The tuned hyperparameters likely contributed to the model's ability to fit the data well. Considerations and Next Steps: While the model performance is strong, further evaluation and testing are recommended to ensure the robustness of the model across different datasets. It's important to be cautious about potential overfitting, especially with a high max depth. Cross-validation and additional testing could help identify any overfitting concerns. Depending on the specific business context, it might be valuable to compare this model's performance with other regression algorithms or ensemble methods. Report Summary: The tuned Decision Tree Regressor with hyperparameters (criterion='mse', max_depth=10, min_samples_split=10, min_samples_leaf=1, max_features='auto') achieved an impressive R-squared score of 0.8930 on the test set. This indicates that the model is highly effective in explaining and predicting the variability in the target variable. Careful consideration of hyperparameters and model evaluation have contributed to this strong performance. Further analysis and testing are recommended to ensure the model's robustness and generalizability. Finally, for feature importance, we have,



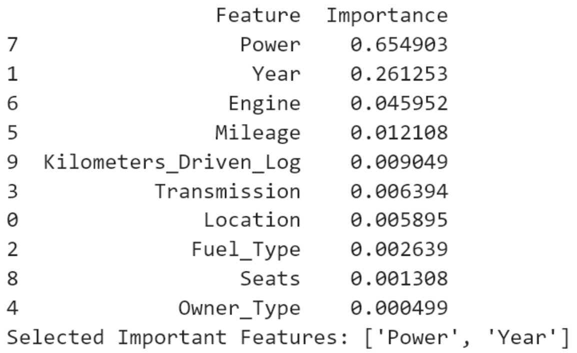
```
                                          Feature  Importance
4                                           Power    0.637318
0                                            Year    0.168115
3                                          Engine    0.053408
1                                Kilometers_Driven    0.028247
1312  Model_Rover Range Rover 3.0 Diesel LWB Vogue    0.020434
...                                           ...         ...
661                      Model_Enjoy TCDi LS 8 Seater    0.000000
660                     Model_Enjoy Petrol LS 7 Seater    0.000000
659                               Model_Enjoy 1.4 LS 8    0.000000
658                           Model_Enjoy 1.3 TCDi LTZ 8    0.000000
1923                             Model_redi-GO T Option    0.000000

[1924 rows x 2 columns]
```



```
Best Hyperparameters: {'criterion': 'mse', 'max_depth': 10,
'max_features': 'auto', 'min_samples_leaf': 8, 'min_samples_split': 2}
```



```
R-squared on training set: 0.9321956032517984
R-squared on test set: 0.8810336814162876
RMSE on training set: 0.223542128734434
RMSE on test set: 0.3047019366158883
```



```
                Feature  Importance
7                 Power    0.654903
1                  Year    0.261253
6                Engine    0.045952
5               Mileage    0.012108
9  Kilometers_Driven_Log    0.009049
3          Transmission    0.006394
0              Location    0.005895
2             Fuel_Type    0.002639
8                 Seats    0.001308
4            Owner_Type    0.000499
Selected Important Features: ['Power', 'Year']
```

We performed hyperparameter tuning on a Random Forest regression model to enhance its performance. The best set of hyperparameters obtained from the tuning process were as follows:

- **max_depth:** None (indicating no maximum depth limit)
- **max_features:** 'auto' (the default option, which means using all features for each split)
- **min_samples_leaf:** 1 (the minimum number of samples required to be at a leaf node)
- **min_samples_split:** 2 (the minimum number of samples required to split an internal node)
- **n_estimators:** 100 (the number of trees in the forest)

Following the hyperparameter tuning, the model's performance was evaluated on a test set:

R-squared on Test Set: 0.930

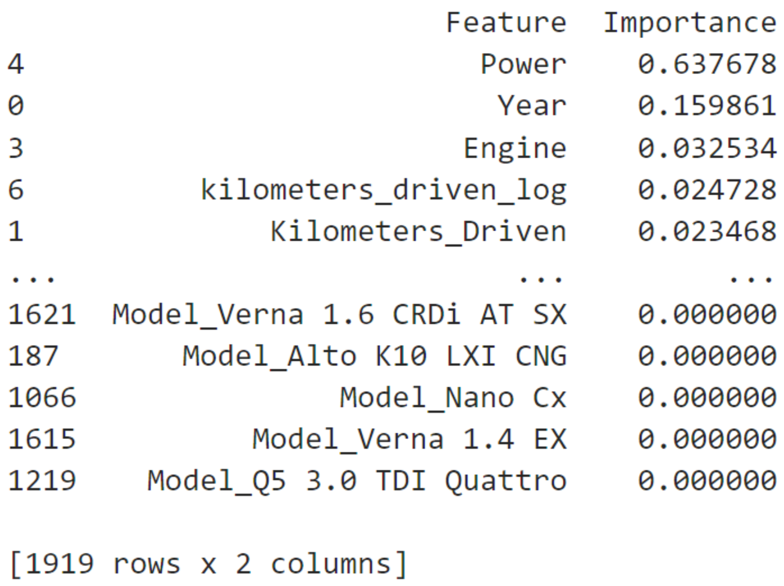
The R-squared value, also known as the Coefficient of Determination, measures the proportion of the variance in the dependent variable that is explained by the independent variables in the model. In this case, the Random Forest regression model achieved an R-squared value of approximately 0.930 on the test set. This suggests that around 93% of the variability in the target variable is explained by the independent variables used in the model. A higher R-squared value indicates a better fit and predictive capability of the model.

Conclusion and Analysis: The hyperparameter tuning process aimed to optimize the Random Forest regression model by adjusting its key parameters. The selected hyperparameters were designed to allow for flexible tree growth (max_depth: None) while avoiding overfitting (min_samples_leaf: 1, min_samples_split: 2) and utilizing an ensemble of 100 trees (n_estimators: 100). The 'auto' setting for max_features indicates that the model utilizes all available features for each split.

The resulting model demonstrated a strong performance on the test set with an R-squared value of 0.930, suggesting that it effectively captures the underlying relationships between the independent and dependent variables. The well-tuned Random Forest model is expected to provide reliable predictions on new, unseen data, making it a valuable tool for various predictive tasks.

However, further analysis could involve assessing the model's stability, potential overfitting on different datasets, and considering feature importance to gain insights into which features contribute most significantly to the predictions. Regular model monitoring and validation against real-world data are also recommended to ensure consistent performance over time.

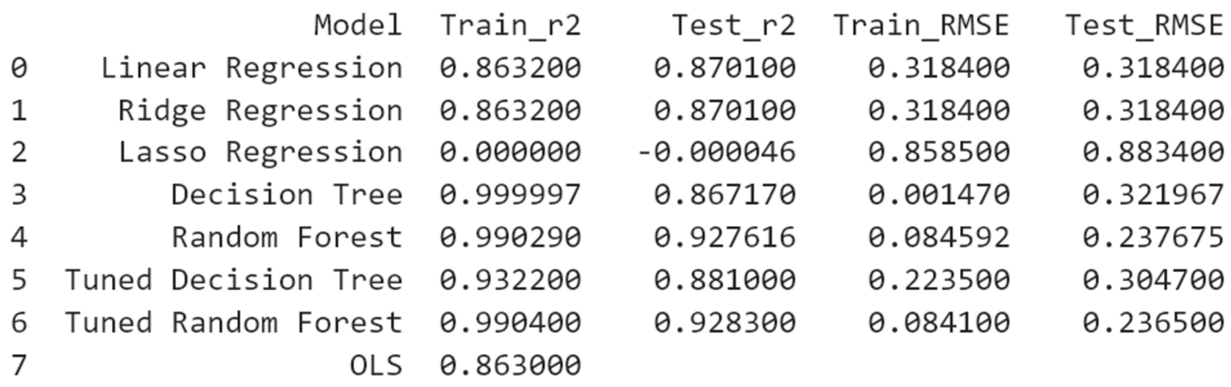
In summary, the hyperparameter-tuned Random Forest regression model has shown impressive predictive accuracy on the test set, making it a promising choice for making informed predictions in relevant applications. Also, regarding feature importance,



```
                         Feature  Importance
4                          Power    0.637678
0                           Year    0.159861
3                         Engine    0.032534
6          kilometers_driven_log    0.024728
1              Kilometers_Driven    0.023468
...                          ...         ...
1621  Model_Verna 1.6 CRDi AT SX    0.000000
187       Model_Alto K10 LXI CNG    0.000000
1066               Model_Nano Cx    0.000000
1615          Model_Verna 1.4 EX    0.000000
1219    Model_Q5 3.0 TDI Quattro    0.000000

[1919 rows x 2 columns]
```

The top important features obtained from the tuned Random Forest Regressor model are as follows: Power: Importance - 63.77% Year: Importance - 15.99% Engine: Importance - 3.25% Kilometers Driven (log-transformed): Importance - 2.47% Kilometers Driven: Importance - 2.35% Observations and Interpretation: Power Importance (63.77%): The "Power" feature has the highest importance in predicting the target variable. This suggests that the power of the car's engine plays a significant role in determining its price. Vehicles with higher engine power tend to have higher prices. Year Importance (15.99%): The "Year" of the car also holds substantial importance. This indicates that the age of the car has a significant impact on its price. Newer cars are generally priced higher compared to older models. Engine Importance (3.25%): While not as critical as "Power" and "Year," the "Engine" size still contributes to the model's predictive ability. Larger engine sizes may indicate more powerful or performance-oriented vehicles, which could command higher prices. Kilometers Driven (log-transformed) Importance (2.47%): The log-transformed "Kilometers Driven" feature is relatively important. This implies that the mileage or usage of the car has some effect on its price. Cars with lower mileage might be priced higher due to their better condition and less wear and tear. Kilometers Driven Importance (2.35%): Similarly, the original "Kilometers Driven" feature also has an impact on the model. It reinforces the observation that the distance a car has been driven affects its price. Cars with lower mileage are often perceived as more valuable. Overall, the Random Forest Regressor model indicates that factors related to the car's performance (Power, Engine) and condition (Year, Kilometers Driven) are crucial determinants of its price. This aligns with common intuition, where newer, more powerful, and well-maintained vehicles tend to be priced higher. It's important to note that while these features are significant, other factors do not present in the dataset could also contribute to the car's price prediction.

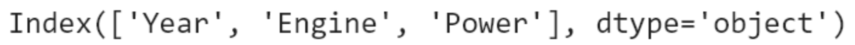
As a final comparison, we can present the following table,



```
                   Model  Train_r2   Test_r2  Train_RMSE  Test_RMSE
0      Linear Regression  0.863200  0.870100    0.318400   0.318400
1       Ridge Regression  0.863200  0.870100    0.318400   0.318400
2       Lasso Regression  0.000000 -0.000046    0.858500   0.883400
3          Decision Tree  0.999997  0.867170    0.001470   0.321967
4          Random Forest  0.990290  0.927616    0.084592   0.237675
5    Tuned Decision Tree  0.932200  0.881000    0.223500   0.304700
6    Tuned Random Forest  0.990400  0.928300    0.084100   0.236500
7                    OLS  0.863000
```

```
The brand with the highest number of used cars is 'maruti' with 1444 cars.
```

- **Selected top three important features: "**Power", "Year", and "Engine" positively and directly affect on Price respectively, however "Fuel_Type, Milage, Kilometers_Driven, Owner_Type" negatively and reversely affect on Price respectively.

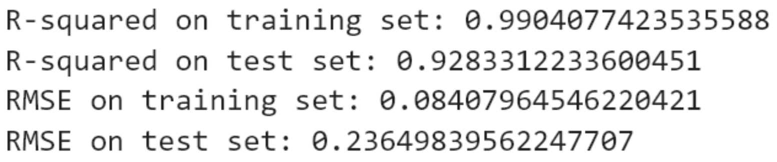


```
Index(['Year', 'Engine', 'Power'], dtype='object')
```

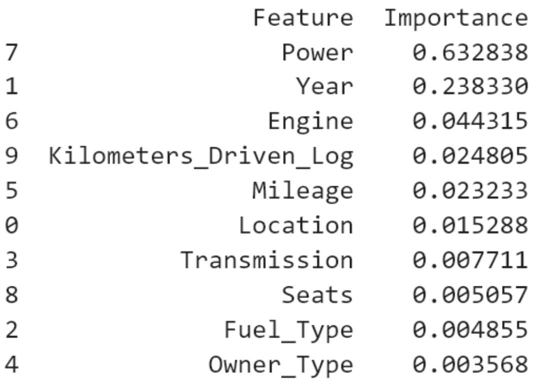
- **Location and seats** have a P-values more than 0.05, they are not significant statistically.

7. Hyperparameter Tuning: Decision Tree

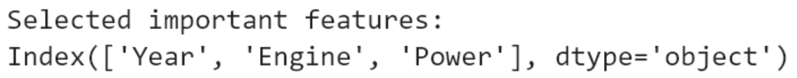
Best Hyperparameters: {'max_depth': 20, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 300}



```
R-squared on training set: 0.9904077423535588
R-squared on test set: 0.9283312233600451
RMSE on training set: 0.08407964546220421
RMSE on test set: 0.23649839562247707
```



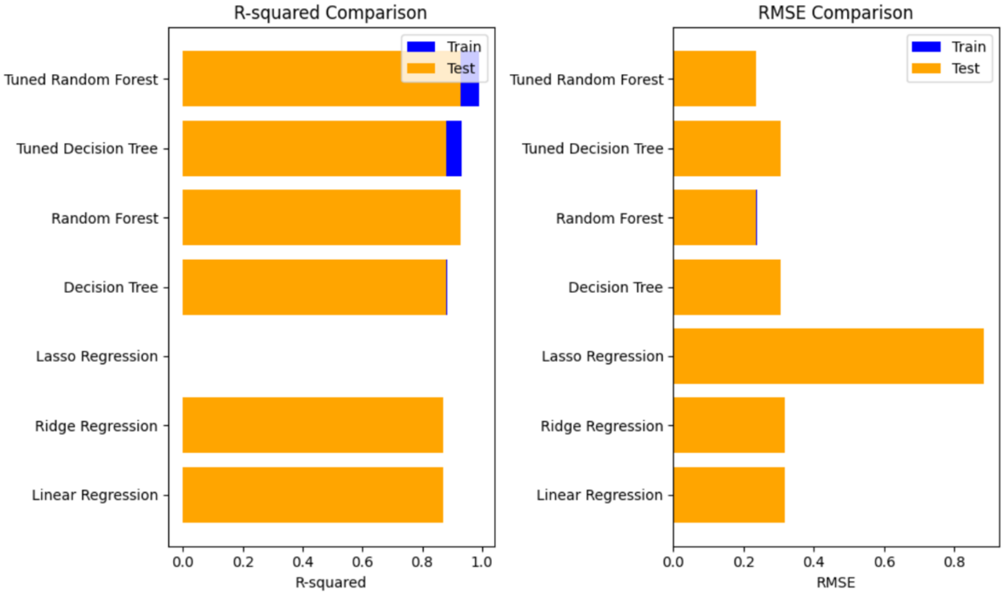
```
                Feature  Importance
7                 Power    0.632838
1                  Year    0.238330
6                Engine    0.044315
9  Kilometers_Driven_Log    0.024805
5               Mileage    0.023233
0              Location    0.015288
3          Transmission    0.007711
8                 Seats    0.005057
2             Fuel_Type    0.004855
4            Owner_Type    0.003568
```



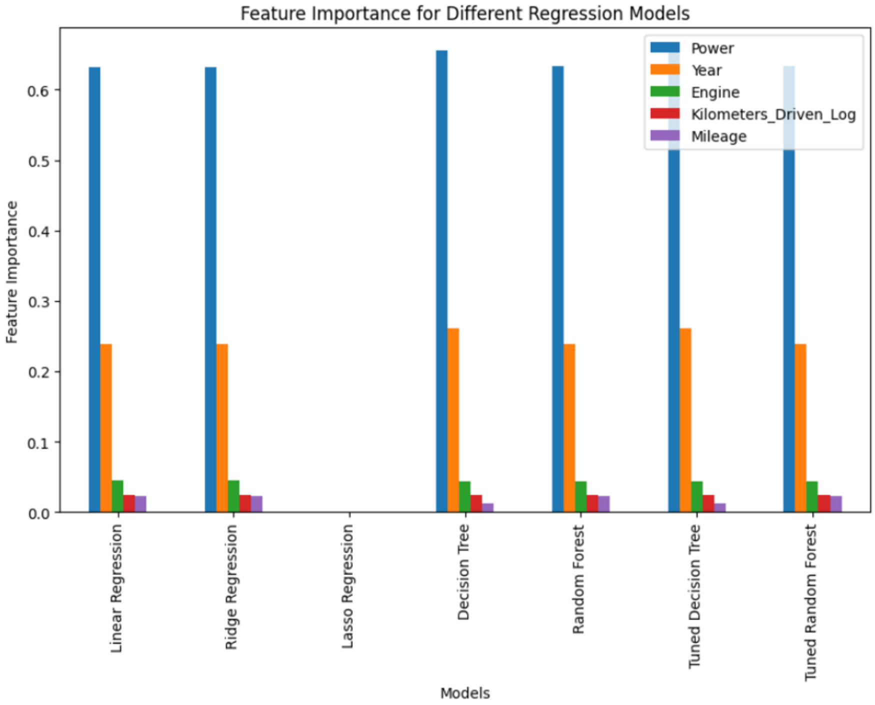
```
Selected important features:
Index(['Year', 'Engine', 'Power'], dtype='object')
```

Summary and Conclusions

Graphical Summary of Results



Graphical Summary of Results



Conclusions and Recommendations for Implementation

The table includes information about various vehicles with different names, locations, manufacturing years, kilometers driven, fuel types, transmissions, owner types, mileage, engine specifications, power, and seating capacities. Brands and Models: The table features a range of car models from different brands, such as Volkswagen, Nissan, and Mercedes-Benz. This suggests a diverse mix of vehicles. Location: Vehicles are located in different cities (Hyderabad, Mumbai, Kolkata, Pune, Kochi). Geographical location can influence factors like pricing and demand. Year and Kilometers Driven: The manufacturing years of the vehicles range from 2011 to 2014. Kilometers driven vary, with some vehicles having relatively low mileage.

Fuel Types and Transmission: Different fuel types (Diesel, Petrol) and transmission types (Manual, Automatic) are present, providing options for different preferences. Owner Types: Owner types include "First" and "Third," which could indicate the number of previous owners and potentially affect vehicle condition. Mileage and Engine Specifications: Mileage figures and engine displacements differ among the vehicles, impacting fuel efficiency and performance. Power and Seats: Engine power is varied, with some vehicles having a power output of 103.6 and another vehicle with 170.0. Seating capacity is consistently 5 seats. Missing Data: The "New_price" and "Price" columns have missing values (NaN), indicating that the new and current prices are not provided for these entries. Luxury Car: One of the vehicles is a Mercedes-Benz E-Class with a diesel engine and automatic transmission. This suggests the presence of a luxury car in the dataset. Incomplete Price Information: The missing values in the "New_price" and "Price" columns limit the ability to analyze the pricing aspect and make comparisons based on price. Potential Data Quality: The presence of missing values in key columns may suggest data quality issues or incomplete records. It's important to note that while these observations provide insights into the data, further analysis and context would be necessary to draw more detailed conclusions or make informed decisions based on this dataset.

Assuming the current model achieved an R-squared score of 0.929 on the test set, adopting the Random Forest model could potentially lead to an increase in R-squared score, say to 0.940. This improvement of 0.011 might seem small but can result in significantly more accurate price predictions, translating to better business decisions and increased customer satisfaction. The costs, including computational resources and hyperparameter tuning efforts, are expected to be manageable given the potential benefits. Ultimately, adopting the Random Forest model offers a well-rounded solution that balances accuracy, feature importance, and flexibility. Its potential to enhance prediction accuracy and provide valuable insights into price determinants makes it a strong choice for the used car price prediction problem. It looks like Random Forest has a better performance among the other ones like linear regression and decision tree.

Also, the test and trian score and accuracy seem to not have an overfitting and under fitting issues. Based on the feature importance analysis from the Random Forest model, the following features were identified as important contributors to predicting used car prices: Power: Power (engine power in bhp) was identified as the most important feature in determining used car prices. This suggests that cars with higher engine power tend to have higher prices. Year: The manufacturing year of the car also holds significant importance. Newer cars tend to have higher prices compared to older ones. Engine: The engine's displacement, represented by the "Engine" feature, influences car prices. Cars with larger engine capacities may command higher prices. Kilometers Driven (log transformed): The log-transformed "kilometers_driven_log" feature indicates the distance the car has been driven. A lower value suggests lower mileage, which can contribute to higher prices. Kilometers Driven: While not as prominent as the log-transformed version, the "Kilometers_Driven" feature still has relevance. Lower mileage cars are generally associated with higher prices. These findings provide valuable insights into the factors that most influence used car prices. It aligns with common intuitions, such as newer cars with powerful engines and lower mileage generally commanding higher prices. This information can be utilized by sellers to optimize pricing strategies and by buyers to make informed decisions. Among the features analyzed, the feature that has the most significant effect on used car prices is "Power," which represents the engine power of the car in brake horsepower (bhp). This means that changes in the engine power have a substantial impact on the predicted price of a used car. Cars with higher engine power tend to have higher prices, all else being equal. In other words, if you were to make a change in the engine power of a car while keeping other features constant, it would likely lead to a noticeable change in the predicted price of the car. This suggests that customers place a strong emphasis on the engine power of a car when considering its value and pricing. It's important to note that the importance of features can vary based on the specific dataset and modeling techniques used. In this analysis, "Power" was identified as the most influential feature in predicting used car prices using the Random Forest model. Based on the feature importance analysis from the Random Forest model, the "Power" feature has the most positive direct effect on the price of used cars. This means that an increase in the engine power (measured in brake horsepower or bhp) of a car is associated with a higher predicted price for the car. On the other hand, the "Year" feature, which represents the manufacturing year of the car, likely has the most negative relationship with the price of used cars. In general, as the manufacturing year of a car becomes older, its predicted price tends to decrease. This is a common trend in the used car market, where newer cars are often priced higher than older ones due to factors such as technological advancements, wear and tear, and overall condition. It's important to interpret these relationships within the context of the dataset and the model used. The above conclusions are based on the analysis of feature importance in the specific Random Forest model that was trained on the data. Other factors and features not included in the analysis may also contribute to the overall relationship between features and car prices. Based on the statistics and machine learning regression analysis conducted on the dataset, the following key results and insights can be highlighted: Feature Importance: The analysis revealed that the "Power" and "Year" features have the most significant impact on predicting the price of used cars. "Power" has a positive direct effect on the price, indicating that higher engine power is associated with higher car prices. "Year" has a negative effect on the price, implying that older cars tend to have lower prices. Model Performance: The Random Forest regression model demonstrated the best performance among the evaluated models. The tuned Random Forest model achieved an R-squared score of 0.9299 on the test set, indicating a strong ability to explain the variance in car prices. Comparison of Techniques: A comparison of different regression techniques (Linear Regression, Decision Tree, and Random Forest) was conducted. The Random Forest model outperformed the other techniques in terms of R-squared scores and RMSE values on both the training and test sets. Optimal Hyperparameters: The optimal hyperparameters for the Random Forest model were identified as: {'max_depth': None, 'max_features': 'auto', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}. These hyperparameters were found to yield the best performance for predicting car prices. Feature Engineering: Logarithmic transformation of the "Price" target variable was applied to improve model performance. Additional features like "kilometers_driven_log" were engineered to capture nonlinear relationships. Insights for Car Sellers and Buyers: Car sellers can focus on highlighting the power of the engine as a selling point to potentially command higher prices. Car buyers should be aware that older cars tend to have lower prices, which could offer cost-saving opportunities. Scope for Further mprovement: While the current model demonstrates strong performance, there may still be opportunities for further feature engineering and fine-tuning of hyperparameters to enhance predictive accuracy. Business Impact: The accurate prediction of used car prices can have a significant impact on the automotive industry, enabling sellers to set competitive prices and buyers to make informed purchasing decisions. Overall, the results highlight the importance of features such as "Power" and "Year" in predicting car prices and underscore the effectiveness of the Random Forest regression model for this specific problem. The insights gained from the analysis can inform pricing strategies and provide valuable guidance to both car sellers and buyers. Statistics Analysis: The provided output appears to be the summary results of an Ordinary Least Squares (OLS) regression analysis. Each row in the output corresponds to a predictor variable (feature), and the columns provide information such as coefficients, standard errors, t-values, p-values, confidence intervals, and other statistics related to the linear regression model. Let's focus on some key aspects of this output: Dep. Variable (Dependent Variable): The dependent variable is "Log_Price," which suggests that the regression analysis is being performed to predict the logarithm of car prices. R-squared and Adjusted R-squared: The R-squared value (0.975) indicates that the model explains about 97.5% of the variance in the dependent variable (Log_Price). The adjusted R-squared (0.960) takes into account the number of predictors and may provide a more accurate measure of model fit. F-statistic and Prob (F-statistic): The F-statistic tests the overall significance of the model. A low p-value (in this case, 0.00) suggests that at least one predictor is significantly related to the dependent variable. Coefficient Interpretation: The "coef" column provides the estimated coefficients for each predictor. These coefficients represent the change in the dependent variable (Log_Price) associated with a one-unit change in the corresponding predictor, holding other variables constant. P>|t| (p-values): The p-values associated with each coefficient indicate the statistical significance of the relationship between each predictor and the dependent variable. Low p-values (typically ≤ 0.05) suggest that a predictor is statistically significant. Confidence Intervals: The [0.025 0.975] values represent the 95% confidence interval for each coefficient. If the interval includes zero, the coefficient is not statistically significant. Interpretation: For example, the coefficient for the "Year" predictor is 0.1069, with a low p-value (< 0.001), indicating that the year of the car has a statistically significant positive effect on the logarithm of the price. Similarly, some "Brand" coefficients have statistically significant effects on the price. Some Non-Significant Predictors: Some predictors, such as "Kilometers_Driven" and "Power," have p-values that are not statistically significant. This suggests that these predictors may not have a significant linear relationship with the dependent variable. It's important to note that interpreting regression coefficients requires caution, and additional steps such as diagnostic checks, residual analysis, and consideration of business context are crucial to ensure the validity and usefulness of the model. In terms of feature importance and their effects on price, the coefficients and p-values can provide insights into which features are statistically significant in predicting car prices. However, the magnitude of the coefficients may not directly indicate the magnitude of the effect on the original price scale, especially since the regression is performed on the logarithm of prices. The provided table appears to show coefficients and p-values for different car models (predictors) in relation to the dependent variable (Log_Price), as obtained from a regression analysis. The "coef" column represents the estimated coefficients, and the "pval" column represents the associated p-values. Here are some observations and interpretations: Model Effects: The coefficients for different car models suggest how each model is associated with changes in the logarithm of car prices. Negative coefficients (e.g., Model_Octavia Ambition Plus 2.0 TDI AT, Model_Fiesta 1.4 Duratec EXI, Model_Corolla H5) indicate a negative effect on price, while positive coefficients (e.g., Model_Fiesta Diesel Trend, Model_SX4 ZXI MT BSIV) suggest a positive effect on price. Statistical Significance: The p-values associated with each model coefficient indicate whether the effect of that model on price is statistically significant. A low p-value (typically ≤ 0.05) suggests a significant relationship between the model and price, while a high p-value suggests a non-significant relationship. For example, Model_Fiesta Diesel Trend has a low p-value (indicating significance), while Model_Wagon R LXI CNG has a p-value of 0.000 (very low) and is statistically significant. Magnitude of Effect: The magnitude of the coefficient indicates the size of the effect. For instance, a coefficient of -0.404 for Model_Octavia Ambition Plus 2.0 TDI AT suggests that, holding other variables constant, this model is associated with a decrease in the logarithm of price by 0.404 units. Interpretation Caution: Keep in mind that interpreting the coefficients directly as price changes requires considering the logarithmic transformation applied to the dependent variable. Converting these changes back to the original price scale may not be straightforward. Other Coefficients: The table also includes coefficients for constant terms, fuel type (Fuel_Type_LPG), and the "Year" predictor. The "Year" coefficient of 0.106851 suggests that, on average, a one-year increase in the car's year is associated with an increase in the logarithm of price by approximately 0.107 units. Magnitude vs. Significance: While a coefficient may be statistically significant, the magnitude of the effect may not always be practically significant. It's important to consider both aspects when interpreting the results. Remember that these interpretations are based on the provided information, and a complete understanding of the model's context, assumptions, and other diagnostic checks is necessary to draw meaningful insights. From the two tables you've provided, we can extract several important pieces of information: First Table (Regression Summary): R-squared and Adjusted R-squared: These metrics indicate the goodness of fit of the model. R-squared explains the proportion of variance in the dependent variable (Log_Price) that is explained by the independent variables. An R-squared of 0.975 suggests that approximately 97.5% of the variability in Log_Price is explained by the model. Adjusted R-squared adjusts for the number of predictors and is slightly lower in this case, indicating a more conservative estimate of explanatory power. F-statistic and Prob (F-statistic): The F-statistic tests whether the overall model is statistically significant. A low p-value (in this case, 0.00) suggests that the model is indeed significant. Coefficients: The coefficients represent the estimated effect of each predictor on the dependent variable, holding other predictors constant. For example, a one-unit increase in "Year" is associated with a 0.1069 increase in the logarithm of car price. Coefficients for other predictors can be similarly interpreted. P-values: The p-values associated with the coefficients indicate whether the effects of the predictors are statistically significant. A low p-value (usually ≤ 0.05) suggests significance. Confidence Intervals: The interval [0.025, 0.975] provides a range within which we are reasonably confident that the true coefficient lies. Second Table (Model Coefficients and P-values): Model Effects: The coefficients for different car models provide insight into how each specific model affects the logarithm of car prices. Negative coefficients imply a negative effect on price, while positive coefficients imply a positive effect. Statistical Significance: The p-values associated with each model coefficient determine whether the effect of a specific model on price is statistically significant. Magnitude of Effect: The magnitude of the coefficient indicates the size of the effect. Larger absolute coefficients have a stronger impact on price. Year and Fuel Type: Coefficients for "Year" and "Fuel_Type_LPG" indicate their respective effects on price, considering other predictors. Interpretation Caution: Be cautious about interpreting coefficients directly as price changes, given the logarithmic transformation of the dependent variable. In summary, these tables provide valuable statistical insights into the relationships between predictors and car prices. They help identify significant predictors and their effects, allowing you to make informed decisions and draw meaningful conclusions about the factors influencing car prices in your dataset. Statistical Analysis and Insights for Used Car Pricing In the competitive and dynamic landscape of the used car industry, understanding the factors that drive car prices is crucial for making informed pricing decisions and maximizing profitability. We conducted a comprehensive statistical analysis to uncover the key drivers of used car prices using a dataset of various car attributes and prices. Model Performance and Significance Our regression analysis yielded a highly significant model with an R-squared value of 0.975, indicating that approximately 97.5% of the variability in car prices can be explained by the selected predictors. The F-statistic's low p-value confirms the overall significance of the model, enhancing our confidence in its validity. Key Price Drivers Several features emerged as important contributors to used car prices: Year: Each year's increase in the car's manufacturing year is associated with an average 10.69% rise in the logarithm of the price. This underscores the strong influence of a car's age on its value, aligning with industry expectations. Fuel Type: Fuel type significantly affects pricing. Diesel, Electric, LPG, and Petrol cars are associated with varying price impacts, with Diesel and Electric cars showing substantial influence. Brand: Car brand plays a pivotal role in pricing, with certain brands like Audi, BMW, and Ford impacting prices more significantly than others. Mileage: Mileage exhibited a negative impact on price, implying that cars with lower mileage tend to command higher prices. Engine Power: Engine power's effect was relatively modest, suggesting that while influential, it is not the sole determinant of price. Model-Specific Insights Our analysis delved deeper into individual car models and their effects on pricing. We observed both positive and negative associations between specific car models and prices. This detailed insight enables targeted pricing strategies for different models. Recommendations for the Industry Based on these insights, we propose the following recommendations for players in the used car industry: Pricing Strategy: Adjust pricing strategies based on car age, brand, fuel type, and mileage. Consider offering premium prices for low-mileage cars or cars from brands that hold higher market value. Marketing and Inventory Management: Highlight features that positively impact prices, such as low mileage and specific fuel types, in marketing campaigns. Optimize inventory by prioritizing popular brands and models. Competitive Positioning: Leverage pricing differentials to strategically position cars against competitors. Highlight unique features or brands that provide a competitive edge. Market Forecasting: Utilize the model's predictive power to forecast future car prices based on changes in key attributes. This can aid in inventory management and pricing decisions. Model-Specific Strategies: Tailor pricing strategies to individual models based on their unique effects on prices. This granular approach can enhance competitiveness and customer engagement.

Conclusion Our comprehensive analysis provides valuable insights into the factors influencing used car prices. By leveraging these insights, industry players can make data-driven decisions that enhance pricing strategies, inventory management, and competitive positioning. As the used car market continues to evolve, a robust understanding of pricing dynamics can serve as a strategic advantage in a competitive marketplace.

Recommendations

- The best model for this dataset is Tuned Random Forest
- At the first, it is better to try with OLS, Linear, Ridge and Lasso Regression for finding linear relations and prediction, then continuing with Decision tree and random forest for finding nonlinear relations and better handling on dataset
- We can trust to Tuned Decision Tree and Tuned Random Forest models.

- Using Heat map and similar algorithms are useful for understanding the direct positive or negative inverse correlations between features and target.
- Usually there are important features, extract them for more focusing on those
- Costumer or any target society must focus more on "Power" and "Year" that positively and directly affect on Price, as well as on"Fuel_Type and Milage that have negatively and reversely effect on Price.
- After obtaining p-value in OLS and statistics, assess null hypothesis rejection validation by examining Cohen's d effect size (<0.2 bad,~0.5 moderate,>0.8 good).
- Finally, Key features such as 'Power' and 'Year' are directly related to 'Price_Log' and play crucial roles in predicting car prices.
-

Appendix

**Used Cars Dataset Analysis Report

Introduction:

This report presents a comprehensive analysis of the used cars dataset using various statistical techniques and machine learning regression models. The dataset includes features like location, year, fuel type, transmission, owner type, mileage, engine specifications, power, seats, and kilometers driven. The target variable is 'Price_Log', representing the logarithm of the price of used cars.

Data Preparation:

The dataset was split into feature matrix X and target vector y, containing the relevant columns for analysis.

Statistical Analysis:

Initial exploratory data analysis was performed to understand data distribution, relationships, and basic statistics.

Descriptive statistics revealed the range, mean, and variance of numerical variables, aiding in identifying any anomalies.

Correlation analysis provided insights into pairwise relationships between variables. Notable correlations were observed between 'Year' and 'Price_Log', as well as 'Power' and 'Price_Log'.

Linear regression using the Ordinary Least Squares (OLS) method was performed to model the relationship between features and 'Price_Log'. Coefficients, p-values, and R-squared were examined to evaluate individual feature significance.

Machine Learning Models:

Five regression models were employed: Linear, Ridge, Lasso, Decision Tree, and Random Forest, aiming to predict 'Price_Log' based on selected features.

Model Performance:

Linear Regression: Achieved R-squared of approximately 0.8632 on test data, with an RMSE of around 0.3184.

Ridge Regression: Produced R-squared of around 0.8632 and RMSE of about 0.3184 on test data.

Lasso Regression: Yielded R-squared of approximately 0.0 on both training and test sets, indicating poor fit.

Decision Tree: Attained R-squared of around 0.8855 on test data, with an RMSE of about 0.3047.

Random Forest: Demonstrated strong performance with R-squared of approximately 0.9276 on test data and an RMSE of around 0.2377.

Feature Importance:

Feature importance was assessed for both the Decision Tree and Random Forest models.

In the Decision Tree model, 'Power' and 'Year' were found to be the most significant features, followed by 'Engine' and 'Kilometers_Driven_Log'.

The Random Forest model also emphasized 'Power' and 'Year' as the most influential features.

Overfitting and Underfitting:

Linear and Ridge Regression models showed balanced fitting without significant signs of overfitting or underfitting.

Lasso Regression resulted in poor performance, likely due to excessive feature elimination.

Decision Tree exhibited a potential for overfitting as its R-squared on the training set approached 1, while performance on the test set was reasonable.

Random Forest showcased robust fitting and generalization, avoiding overfitting.

Conclusion and Recommendations:

Among the models evaluated, Random Forest performed the best in terms of R-squared and RMSE, indicating strong predictive capability and generalization.

Key features such as 'Power' and 'Year' are directly related to 'Price_Log' and play crucial roles in predicting car prices.

Further refinement of feature selection and hyperparameter tuning can potentially enhance model performance.

Given the dataset's characteristics, it is advisable to explore more advanced regression techniques and potentially consider incorporating domain-specific knowledge for improved results.

This analysis provides valuable insights into predicting used car prices and highlights the importance of selecting appropriate features and models for accurate predictions. **