Used Bike Prices - Feature Engineering and EDA

Tools : Visual Studio code / jupyter notebook

Domain : Finance Analyst

Project Difficulties level : Advance

Insights of the project

1. Which type of cars are sold maximum?

- we can conclude that cars with body : Sedan, Engine Type : Petrol, Drive : Front are sold maximum.

2. What is the co-relation between price and mileage?

- Price and Mileage are negatively co-related with corelation coefficient = -0.28
- we can conclude that :
- Most of the cars gives mileage in [1, 300] with price < 150000
- There exists a few excpetions / outliers in case of Petrol cars.

3. How many cars are registered?

- Total 8611 Cars are registered and 552 are not registered.

4. Price distribution between registered and non-registered cars.

- we can say that all the cars which are registered have IQR of price between 6000 to 18000 with median at 9900. And, cars which are not registered have price IQR between 2500 to 4500 with median at 3000. So, not-registered cars have significantly lower prices compared to the registered cars.

5. What is the car price distribution based on Engine Value?

- This data contains outliers. So, we can better conclude the results after outlier treatement. As of now, the corr value between these two variables is 0.05.
- As per the observation, price range for most of the cars is in [0, 100000] with Engine Value in [1, 6] with a sighlty positiive trend in the data.

6. Which Engine Type of cars users preferred maximum?

- user preference order is : Petrol > Diesel > Gas > Other

7. Establish corelation between all features using heatmap.

- Price is positively co-related with Engine Value & Year and negatively co-related with Mileage.
- Mileage is positively co-related with Engine Value and negatively co-related with Year.
- Engine Value is negatively co-related with Year.
- None of the co-relations is strong. Here, Mileage and Year having the maximum co-relation with coefficient = -0.46.

8. Distribution of Price

- We can see that this data contains lots of outliers.
- The price data is Right-skewed with skewness value = 6.127
- Logarithmic distribution follows close to Normal distribution.
- The Price Distribution has its IQR around [5000, 15000] range with median around 10000.
- The Price varies from 0 to 40000 for maximum no. of instances.

Thank You !!