Alternative Project – Futurization & Model Tuening

Data Description:

This dataset contains information about used cars listed on www.cardekho.com This data can be used for a lot of purposes such as price prediction to exemplify the use of linear regression in Machine Learning.

Domain:

Used car market

Context:

Given the attribute below, we want to see how well we can predict the selling price of a car and also figure out which attributes are stronger factors in making predictions.

Attribute Information:

- Car_Name - name of the car
- Year - year in which the car was bought.
- Selling_Price - price the owner wants to sell the car at
- Present_Price - current ex-showroom price of the car
- Kms_Driven - distance completed by the car in km
- Fuel_TypeFuel - type of the car
- Seller_Type - Defines whether the seller is a dealer or an individual
- Transmission - Defines whether the car is manual or automatic
- Owner - Defines the number of owners the car has previously had

Learning Outcomes:

- Exploratory Data Analysis
- Preparing the data to train a model
- Training and making predictions using a Regression model
- Model evaluation

Objective:

Train a regression model to predict the Selling price of a given car, evaluate the performance of the model and make attempts to make better predictions by Feature Engineering.

Steps and tasks:

1. Read the column description and ensure you understand each attribute well
2. Check the presence of null values in the data (2 marks)
3. Study the data distribution in each attribute, share your findings (15 marks)
4. Check the presence of outliers in Present_Price and Selling_Price (3 marks)

5. Label encode the categorical variables wherever necessary (3 marks)
6. Separate the data into dependent and independent variables and split the data into training and test set (2 marks)
7. Train a Linear regression model using the training data and make predictions on the test data (10 marks)
8. Calculate the RMSE by comparing the predictions with the actual values (5 marks)
9. Plot a scatter-plot with the predictions on one axis and the actuals on the other and write down your insights (5marks)
10. Repeat steps 5 to 9, but with the Fuel_Type one hot encoded and comment on your new results (10 marks)
11. Impute the outliers in Present_Price and Kms_Driven with median, perform step 10 again and comment on the new results (5 marks)
12. Drop at least one variable that adds least value to the model, repeat step 10 and comment on the results. And write a detailed conclusion on what worked in improving the model and what did not. (extra appreciation for trying out more than the above approaches) (15 marks)