Predictive Analytics: Predicting Used Car Prices

Introduction

CarDekho.com is a website and mobile application which facilitates its users to buy right vehicles for them. It does this through a user-friendly website and mobile app that carry rich automotive content such as expert reviews, details of all specs and prices, comparisons as well as videos and pictures of all brands and models available in India.

Objective

Follow the Data Science Methodology that we studied in Module 3. Analyse the data set and identify the most relevant factors that influence customer to buy car based on price, kilometres driven, fuel type, transmission type, etc.

Dataset

The dataset, **car data.csv**, is taken from Kaggle. It has been uploaded to canvas.

Tasks in this assignment

1. Write a Data Science Proposal for achieving the objective mentioned.

2. Perform exploratory analysis on the data and describe your understanding of the data.

3. Perform data wrangling / pre-processing. E.g., missing data, normalization, discretization, etc.,

4. Apply any two feature selection engineering techniques.

5. Compare the two selected feature engineering techniques.

6. Plot top 5, 6, and 8 features.

7. Provide a high-level description of Machine Learning models – association rules and random forest to predict.

8. Describe the entire data set and identify your observations from this.

9. Plot the histogram of various data sets (at least for 3 parameters i.e. year in which the car was bought., selling price, showroom price)

10. Present the conclusions/results (Answers for the given objectives) in the format shared.

# **Expected Submissions**

Two files are expected as the assignment submission.

1. The summary of the work in the template provided. (You may fill only the boxes relevant to this problem statement)
2. The executed ipynb file with clear subdivision of the codes and brief description of the purpose of respective code. All the executed tables or graphs and results should be present in the ipynb file. The ipynb file maybe submitted as a single .pdf file.

**Additional Instructions:**

Please take the dataset.

Do the preliminary data pre-processing, (Scope: As discussed in class)

Find the best features influencing the target variable using PCA or correlation.

Apply regression and classification (using decision tree).

Submit your assignment with explanations for the choice of the ML techniques and

summarize your findings from the analysis of the data.

Use the submission template and rubric as reference and include only what is applicable for your problem.