Cairo University
Faculty of Engineering
Computer Engineering Department

CMPN451
Spring 2023

Data Mining, Big Data and Analytics
Assignment

Description:

You are given a dataset for car prices based on some features provided in the dataset. Using this dataset, you are required to do the following:

- Identify the useful and irrelevant features in the dataset.
- Clean the dataset, which can include the following:
  - Identifying relevant and irrelevant columns.
  - Handling of missing data.
  - Dealing with outlier data.
  - Handling different forms of features (Categorical / Numerical).
  - Handling wide ranges of numerical values.

  These are just suggestions feel free to add whatever you see fit. However, you should have a valid argument for each decision you take in this step.
- Analyze the features in the dataset and identify their significance and correlation.
- Identify candidate models for this problem and train them on the given training set.
- Choose some appropriate evaluation criteria for the models and use it to compare results.
- Use your model to predict car prices in the test set.

Deliverables:

1. **Code**

   The script used for solving this problem either in R or in Python.

   Make sure the code is clear and commented.

2. **Report**

   Final Document containing:

   i. High level Project pipeline.
   ii. Clear description of the data cleaning process.
   iii. Data visualization and Plots.
   iv. Models trained.
   v. Results and Evaluation.

3. **Submission.txt**

   This file contains your predictions for each sample in the test set. It should have the following format:

   ID,price
   1,500
   2,632

Notes:

- You should work on this assignment in teams of **2 members**.
- There is a penalty for late submissions.
- Since you are all given the same dataset with the same inputs you are subject to competition between each other. The criteria for the comparison between the teams shall be:
  - The model accuracy and final results.
  - The quality of the project pipeline and the steps followed to achieve the final results.
- We will evaluate your submissions with Root Mean Squared Log Error (RMSLE).
- **Any sign of cheating or plagiarism will not be tolerated and will be graded ZERO in the assignment.**