Data Science Evaluation

The objective is to assess the data science techniques you employ, rather than focusing on the model's accuracy. We do not anticipate the model to be accurate.

Please provide your code in a Jupyter notebook with inline comments and interpretations.

Task 1

Objective:

Design a robust data science solution to predict the visual grade condition of a car before it reaches a WBAC site, enabling pricing analysts to make more informed purchasing decisions.

Overview:

At We Buy Any Car (WBAC), each car is assigned a visual grade condition ranging from 1 to 5, which helps auction buyers assess the condition of the vehicle.

Scenario:

The pricing analysts, responsible for overseeing stock purchasing, want to predict the visual grade condition of a car before it arrives at a WBAC site. They aim to utilize previous purchasing data to make these predictions more accurately.

Your Task:

1. **Dataset:** Use the provided Grades.xlsx dataset, which contains historical data on car purchases and their assigned visual grades.
2. **Required Outputs:**
    - Develop a machine learning model to predict the visual grade condition (1-5) of cars.
    - Structure your approach in an organized notebook that demonstrates your methodology.
    - Compare multiple machine learning algorithms and techniques and justify your choices.
    - Use a validation approach with a hold-out set separate from the test set to validate your model's performance.

- Ensure that the final 10% of the dataset is reserved for testing your predictions.

Points to Consider:

- Clearly outline your data preprocessing steps, feature engineering, and model selection criteria.
- Explain your choice of algorithms and any hyperparameter tuning you perform.
- Discuss the results of different models and combinations you have tried, and how you determine the best-performing model.
- Use appropriate metrics to evaluate and compare model performance.
- Interpret the results with visualizations and explanations to support your findings and choices.

Hint: MotExpireDate indicates the date when the MOT of the vehicle expires. Assume that the current date is 21/02/2024 to derive daysLeftTillMotExpiry.

Task 2

Objective:

Design a data science solution to help the auction team focus on the most in-demand stock, balancing competitiveness and performance at auctions.

Overview:

At We Buy Any Car (WBAC), we want to identify the most in-demand and profitable stock based on auction performance. Each stock is grouped by Make and Model, and auction performance is evaluated across four key dimensions provided in the dataset Cluster.csv:

- Bids: Average number of bids that each Make-Model group received.
- UniqueBidders: Average number of unique bidders that each Make-Model group received.
- ConversionPct: Sales conversion rate for each Make-Model group.
- PerformancePct: Sales performance as a percentage of the Guide Value for each Make-Model group.

Scenario:

The auction team aims to foster a more competitive bidding environment by selectively pushing bidders. However, they want to minimize risk by focusing on stock that is already in high demand.

Dataset: Utilize the provided AuctionData.xlsx dataset.

Required Outputs:

- **Part 1:** Cluster the Make-Model combinations based on the four dimensions mentioned above.
- **Part 2:** Produce the following:
  - A 2D Voronoi diagram illustrating the clusters, including centroids and vertices. Note: The diagram should not be limited to selecting only 2 out of the 4 dimensions; you should use an appropriate dimensionality reduction method to achieve this.
  - A ranked list of each Make-Model group with its respective cluster.
  - A ranked list of clusters (1 to n, where 1 represents the best-performing cluster and n represents the worst).
  - Descriptive statistics for each cluster (mean, median, standard deviation, etc.).

Points to Consider:

- If clustering produces clusters in an unranked form, you will need to develop a strategy to rank clusters from 1 to n, where 1 represents the best-performing cluster.
- Explain the method you use for dimensionality reduction and why it is appropriate.
- Discuss your choice of the number of clusters and how you determine the optimal value.
- Provide a well-organized notebook that demonstrates your approach and findings, including all required visualizations and outputs.