CIT – connected innovation topics
Assessment - Data Scientist

Overview
As an automotive supplier, we are interested in predicting critical factors that affect our business. For this task, you are provided with a real-world dataset related to the automotive industry.

Objective & Tasks
Your objective is to perform a thorough analysis of this data and build a machine learning model to predict a target variable. This should include:

1. Exploratory Data Analysis (EDA): Conduct a thorough exploratory data analysis. This should include understanding the distribution of data, detecting outliers, and exploring relationships between features. Visualize important features and correlations.

2. Feature Engineering and Selection: Based on your EDA, engineer new features and select the most relevant ones for your model. Justify your choices.

3. Machine Learning Model: Build a machine learning model to predict the "price" variable. Explain your choice of model and any hyperparameters you tune. Use appropriate validation techniques.

4. Evaluation and Interpretation: Evaluate the performance of your model using appropriate metrics. Interpret your model's predictions, and discuss its strengths and weaknesses.

Dataset
Use the following dataset for this task:
UCI Machine Learning Repository: Automobile Dataset

Documentation
Include a MS PowerPoint presentation that:

1. Explains your EDA process and findings.
2. Describes your feature engineering and selection process.
3. Details your machine learning model building process, including how you validated the model and tuned any parameters.
4. Evaluates the model's performance and interprets its predictions.
5. Discusses any challenges you faced and how you addressed them.

Delivery
Please provide your code, the final processed data, your model, and your report in a Jupyter notebook. Upload all these materials to a public GitHub repository and share the link with us.

Evaluation Criteria
We will evaluate your work based on:

1. **Communication (30%)**: How well you explain your process and findings, and how effectively you visualize your data and results.
2. **Correctness (25%)**: The appropriateness and correctness of your methodology.
3. **Efficiency (20%)**: The efficiency of your code and use of computational resources.
4. **Creativity (15%)**: Your innovation in feature engineering, model building, and interpretation.
5. **Robustness (10%)**: Your consideration of potential pitfalls and how you validated your model.

Remember, while efficiency is important, it is more crucial for us to understand your thought process, your problem-solving skills, and your ability to clearly communicate your methods and results. Good luck!

BOSCH