Birla Institute of Technology & Science, Pilani

Feature Engineering - Assignment

Course No. : AIML_Oct_2023_C2_PCAMZC111

Course Title : PG Program in AI & Machine Learning

Total Marks: 50

= =

Assignment Title: Comprehensive Feature Engineering for Predictive Modelling

Objective:

The aim of this assignment is to deepen your understanding of various feature engineering techniques and their impact on the performance of predictive models. You will work with datasets containing a mix of continuous, categorical, and text variables, and implement various feature transformation and creation strategies.

Part 1: Data Preparation and Exploration on iris data (7 Marks)

Dataset :
https://drive.google.com/file/d/17dBC_LyFiYb3Unkt_1Fw2VveJHR9BjnR/view?usp=sharing

- **Q1 [5 Marks]:** Load iris data set that is shared as part of the assignment. Data contains a mix of continuous and categorical. **Provide a brief exploratory data analysis**, including statistics, distributions, and any patterns observed.
- **Q2 [2 Marks]:** Handle missing values and outliers in your dataset. Document your strategy and justify your choices.

Part 2: Data Preparation and Exploration on wine data (8 Marks)

- **Q1 [5 Marks]:** Load data set that with below line of code. Provide a brief exploratory data analysis, including statistics, distributions, and any patterns observed. Handle missing values and outliers.

  from sklearn.datasets import load_wine
- **Q2 [3 Marks]:** Apply relevant feature engineering and transformation techniques like Normalization, scaling. Apply any feature reduction technique on wine data set and highlight what all features are considered important by the technique used.

Part 3: Model Building and Evaluation (20 Marks)

- **Q1 [5 Marks]:** Split data in training and testing set. Train a base classification model for part 1 and regression model for part 2. Use relevant evaluation metrics and note the performance.
- **Q2 [15 Marks]:** Re-train the model with transformed data in Part 1 and Part 2 in various iterations and note evaluation metric. Document all the improvements or deterioration observed after the various Feature engineering techniques employed.

Part 4: Dimensionality Reduction (5 Marks)

- **Q1 [5 Marks]:** Apply dimensionality reduction on Part 1 and Part 2, specifically PCA with above 90% variance and re-train the model again and note & document all the improvements or deterioration

Part 5: Video Recording (10 Marks)

- **Q1 [10 Marks]:** Video record and provide detailed explanation of everything done as part of the assignment.

Submission Guidelines:

- Provide a well-documented Jupyter Notebook containing the code, outputs, and a brief explanation of your findings for each question.
- Your explanations should clearly demonstrate your understanding of how and why each technique affects the model's performance.
- For custom implementation questions, provide a comparison with Scikit-learn’s built-in functions to validate your implementation.
- Ensure that the code is clean, well-commented, and reproducible.
- Only 3 files should be uploaded in canvas without zipping them.
    - Single IPYNB file (Part 1 to Part 4)
    - HTML output of the ipynb file.
    - Video File (Part 5)

Evaluation Criteria:

- Correctness and completeness of the implementation.
- Depth of analysis and explanation for each technique's impact on the model.
- Creativity in feature engineering and problem-solving approach.
- Quality and clarity of the code and explanations.

Note: This assignment is designed to challenge you and deepen your understanding of feature engineering. You're encouraged to explore additional resources and literature to aid in your implementation and analysis. Happy coding!

****