PG-DBDA (March 2023 Batch)

Date : 22/07/2023

Time : 10:00 AM to 10:00 PM

Subject: Practical Machine Learning

Note: There are two files to be created for this exam:

- **.ipynb file (with proper observations)**
- **.docx file (your findings and observations)**

Business Context:

- **In this case, trainees are tasked with building a predictive model to identify fraudulent transactions for a financial company. The dataset provided contains 6,362,620 rows and 10 columns in CSV format. The goal is to develop a machine learning model that can effectively distinguish between legitimate and fraudulent transactions.**

  To clarify the process, here's a step-by-step explanation:

- **Objective: The main objective is to create a predictive model that can accurately classify transactions as either legitimate or fraudulent. This model will be based on the available data and will be used to make predictions on new, unseen data.**

- **Data: The dataset consists of 6,362,620 rows (transactions) and 10 columns (features) in CSV format. The specific features included in the dataset are not mentioned in the description, but they are essential for training the model. They could include transaction amount, location, time, customer information, and any other relevant data points.**

- **Model Development: Trainees are given the freedom to choose any suitable method for developing their machine learning model. This could include using various algorithms such as logistic regression, random forests, support vector machines, or neural networks.**

- **Model Training: The model will be trained on the calibration data, which is a subset of the provided dataset. During training, the model will learn patterns and relationships from the labeled data (transactions marked as fraudulent or legitimate).**

- **Model Validation: After training, the model's performance will be evaluated on the validation data, which is another subset of the dataset that the model has not seen during training. This step helps assess how well the model generalizes to new, unseen data.**

- **Model Comparison: To ensure the best possible predictive performance, trainees are encouraged to experiment with various machine learning algorithms and techniques. This could involve trying different classifiers (e.g., logistic regression, decision trees, random forests, gradient boosting, neural networks) or employing different preprocessing and feature engineering methods.**

- **Performance Metrics: To evaluate the models' effectiveness, appropriate performance metrics such**

as accuracy, precision, recall, F1 score, and area under the receiver operating characteristic curve (AUC-ROC) should be calculated for each model on the validation data. These metrics will provide insights into how well each model performs in identifying fraudulent transactions.

- **Model Selection: Based on the performance metrics, trainees will be able to determine which model is the most suitable for the task. The model with the highest accuracy or best trade-off between precision and recall, for example, may be considered the most appropriate for fraud detection.**

- **Interpretation of Results: Apart from performance metrics, it's crucial to interpret the results and understand why certain models perform better than others. Trainees should investigate the features and decision boundaries used by each model to gain insights into how they make predictions.**

- **In summary, the case requires trainees to build multiple machine learning models and compare their performance using various metrics. By doing so, they can determine the best model for detecting fraudulent transactions. The process of model comparison and interpretation helps enhance the overall accuracy and reliability of the fraud detection system and guides the development of an actionable plan based on the chosen model's strengths.**