Data Analytics Officer Assessment

INSTRUCTION

This assessment consists of three sections: Data Formatting, Machine Learning Application, and Communication & Presentation. Complete all sections and submit your assessment within 3 days hours. Your responses will be evaluated based on accuracy, technical skills, problem-solving ability, and communication clarity.

DATASET DESCRIPTION

This dataset shared in the file called **“ILTESTData_Raw_20240909.xlsx”** contains information about 2,000 insurance transactions. Each row in the dataset represents a single transaction and includes details about the transaction, the customer involved, and additional attributes that might affect credit scoring.

SUBMISSION

- **Due Date & Time**: 13th Friday 9:00 AM
- **Your submission should include:**
  - Python scripts used in data formatting and model development.
  - A presentation for each of the problem statements using a PowerPoint or PDF slide deck for each of your presentations. Focus on clarity, visual appeal, and effective communication of your insights. Include charts or visuals where relevant.
- Ensure all work is original.

DATA DICTIONARY

Here’s what each column in the dataset represents:

a) **Transaction ID**: A unique number assigned to each transaction for identification.

b) **Class of Business**: The broad category of insurance related to the transaction. For example, it could be aviation, engineering, or domestic fire insurance.

c) **Sub-Class of Business**: A more specific type within the broader category. For example, within domestic fire insurance, it might include various sub-types.

d) **Transaction Type**: How the transaction is classified. This could be a direct transaction, inward reinsurance (from other insurers), or outward reinsurance (to other insurers).

e) **Amount**: The amount of money involved in the transaction. This can be a positive number, a negative number (indicating refunds or adjustments), or missing information.

f) **Currency**: The type of currency used in the transaction, like Kenyan Shilling (KES) or US Dollar (USD). Sometimes, this information might be missing.

g) **Customer Name**: The name of the customer who made the transaction.

h) **Date**: The date when the transaction occurred.

i) **Credit Score**: A number showing the creditworthiness of the customer. Higher numbers mean better credit.

j) **Age**: The age of the customer at the time of the transaction.

k) **Income**: The annual income of the customer.

l) **Number of Claims**: How many insurance claims the customer has made in the past.

m) **Account Balance**: The current balance of the customer's account, which can be positive or negative.

n) **Default**: Indicates whether the customer defaulted on their obligations. It's a simple yes (1) or no (0).

SECTION 1

Data Formatting to IRA Template (30 points)

You are provided with raw financial data from various sources, and your task is to format this data to comply with a specified IRA template named "**ILTESTData_Raw_20240909 – Template.xslx**".

- Write a Python script that reads the raw data from the excel file provided, clean it, and formats it according to the given IRA template. Include code snippets, explanations of key functions, and any assumptions made.
- Enhance your script to be reusable for future datasets that follow similar formatting requirements.
- Describe how you would automate the script to run periodically, including logging and error handling.
- Explain how you would test the accuracy of your formatted data against the IRA template. Provide code examples of tests you would implement (e.g., unit tests or data validation checks).

SECTION 2

Machine Learning - Credit Scoring in Insurance (50 points)

Develop a Machine Learning model to create a credit scoring system tailored for insurance applicants. The goal is to assess the risk of clients based on their financial data and insurance history.

- Choose a suitable Machine Learning model (e.g., logistic regression, decision tree, gradient boosting) for credit scoring.
- Describe your approach to feature engineering, model training, and hyperparameter tuning. Include Python code snippets demonstrating your process.
- Define the evaluation metrics (e.g., AUC-ROC, F1 score, precision, recall) and justify why these are suitable for the credit scoring task.
- Provide an analysis of how the model's predictions can impact underwriting decisions, risk assessment, and overall financial performance.
- Discuss how you would identify and mitigate potential biases in your credit scoring model. Include any ethical considerations specific to insurance.

SECTION 3

Communication & Presentation (20 points)

Prepare a presentation to communicate your findings and model results to the management team. Your goal is to clearly explain the technical aspects in a way that is understandable to non-technical stakeholders.