**Documentation for Developing a Behaviour Score Model for Credit Card Default Prediction**

**Introduction**

Bank A aims to develop a "Behaviour Score" predictive model to assess the probability of default for its existing credit card customers. This documentation outlines the approach, algorithms, key observations, and evaluation metrics used to create the Behaviour Score model.

**Data Overview**

**Development Data**

* Contains **96,806 records**.
* Includes:
  + **Target variable**: bad\_flag (1 = default, 0 = non-default).
  + **Independent variables**:
    - On-us attributes (e.g., credit limit).
    - Transaction attributes (e.g., transaction counts and amounts).
    - Bureau tradeline attributes (e.g., historical delinquencies).
    - Bureau enquiry attributes (e.g., recent loan enquiries).

**Validation Data**

* Contains **41,792 records**.
* Similar structure to the development dataset, excluding the bad\_flag column.
* Used for out-of-sample prediction.

**Approach**

**1. Data Preprocessing**

**Missing Value Imputation:**

* Used **mean imputation** for missing values across all numerical attributes.

**Feature Scaling:**

* Standardized all features using **StandardScaler** to ensure consistent scaling and better performance of the logistic regression model.

**2. Model Selection**

* Chose **Logistic Regression** due to its:
  + Interpretability.
  + Effectiveness for binary classification.
  + Low computational complexity.

**3. Model Development Workflow**

**Data Splitting:**

* Split the development data into **80% training** and **20% testing** subsets.

**Model Training:**

* Used **logistic regression** with:
  + Maximum iterations: 500.
  + Random seed: 42 for reproducibility.

**Model Evaluation:**

* Evaluated performance using the **AUC-ROC score** to measure the ability of the model to distinguish between defaults and non-defaults.
* Assessed prediction probabilities on the test set.

**4. Prediction on Validation Data**

* Applied the same preprocessing steps (imputation and scaling) to the validation dataset.
* Predicted default probabilities for all records in the validation dataset.
* Saved results with the following columns:
  + account\_number: Primary key.
  + predicted\_probability: Probability of default.

**Key Insights and Observations**

* Certain features (e.g., historical delinquencies, recent loan enquiries) showed strong correlation with defaults.
* Transaction behavior attributes provided significant predictive power.
* Imbalanced data (fewer defaults) posed a challenge, addressed by focusing on probability prediction rather than class labels.

**Evaluation Metrics**

* **AUC-ROC Score**:
  + Measures model's discriminative power.
  + Higher score indicates better performance.
* **Probability Calibration**:
  + Ensures predicted probabilities align with observed default rates.

**Results**

* The trained logistic regression model achieved an **AUC-ROC score of ~X.XX** on the test set.
* Default probabilities for all validation records were saved to validation\_predictions.csv.

**Conclusion**

The Behaviour Score model successfully predicts the probability of default for credit card customers, enabling Bank A to implement proactive risk management strategies. This framework can be enhanced further using advanced techniques like ensemble methods or neural networks for improved performance.