PD Model

1. Target definition

The probability of default is defined as the likelihood that a borrower fails to repay a loan within 21 days from the disbursement date. A binary target variable was constructed as follows: customers with *amount_outstanding_21d* > 0 were labeled as **BAD** (i.e. in default), while those with no outstanding balance were labeled as **GOOD**.

2. Feature Engineering

A comprehensive exploratory data analysis (EDA) was conducted to assess the distributional properties, presence of missing values, outliers, and anomalies (e.g., negative values) across features, which could indicate data integrity issues. Based on insights from the EDA, several derived features were engineered:
card_expiry_days - number of days remaining until the card's expiration, measured from the loan issuance date
num_failed_payments_* - ratio-based variables reflecting historical trends of failed payments
amount_repaid_* - ratio-based variables representing repayment behavior over time
The categorical features (**merchant_group, merchant_category**) were transformed using one hot encoding and target encoding, respectively.

3. Validation split

The dataset was partitioned to ensure robust model development by splitting into 3 groups: DEV/VAL (70:30) and OOT comprising the most recent month, reserved for independent validation.

4. Feature selection

To reduce dimensionality and eliminate irrelevant predictors, a shortlist of variables was created using quasi-Boruta algorithm (random variables introduced into the dataset), a XGBoost model with intentional overfitting, feature importance and shap importance metrics. The final set of features was selected based on their performance relative to the introduced random variables.

5. Final model

The final model was built using an XGBoost classifier with 413 trees, after performing Bayesian optimization for hyperparameter tuning. The model achieved the following results:

Sample	Bad Rate	GINI coef (2 * AUC - 1)
DEV	5.69%	41.88%
VAL	4.44%	40.65%
OOT	5.72%	37.98%

6. Calibration

To align predicted probabilities with observed default rates, isotonic regression was applied to calibrate the model on the full dataset. The calibration curves indicate improved agreement between predicted probabilities and actual outcomes, thereby enhancing the model's suitability for risk estimation tasks.