CASE: Automatic Credit Modeling System

Background

In consumer lending, lenders use credit modeling methods to assess the default risk of borrowers and decide whether to accept an applicant for a loan. Recent advances in machine learning techniques have improved the effectiveness of risk assessment models but also brought practical challenges. A thorough understanding of the input data helps strengthen trust in models, but such analysis tasks usually require heavy coding work and are time-consuming. Complex machine learning models make it difficult to explain how the model works and why an applicant gets a high or low credit score from the model. Further efforts should be made to build trustful models in the financial decision process efficiently.

In this competition, participants are expected to design an automatic analysis and modeling system that provides reliable machine learning models for credit assessment in consumer lending. The system should perform an analysis of the input data to help users have a thorough understanding of both samples and features. The system is expected to give explanations on models and certain predictions rather than simply provide black-box models. Participants are encouraged to use various visualization techniques to provide insights.

Data

1. Label data: It contains the date of loan applications (“APPLICATION_DATE”) and default labels (“DEFAULT_LABEL”). The value of 1 in “DEFAULT_LABEL” stands for default and 0 for normal.
2. Feature data: It contains anonymous features (“V000” - “V999”) of applicants collected at the application date (“APPLICATION_DATE”) which can be used as independent variables to predict the default label. It can be matched to label data with “APPLICATION_ID”.

Requirements

The minimum requirements of the prototype should meet the following criteria:

1. It should provide an analysis function for users to understand training data before building models.
2. It should provide a modeling function to train at least one type of classification model using machine learning algorithms and evaluate the model’s effectiveness. A test dataset is provided for the performance evaluation.
3. It should provide interpretations of models and predictions.