Architectural Decision Document

1. Use Case

My case is to build a model for predicting default payments on bank credit cards. For this study, we use "Default of Credit Card Clients Dataset", see: Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. This dataset can be also downloaded from: https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

The dataset contains information (30,000 records) on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. There are 25 variables (id, 23 features, and target). Here is their brief description: 1) ID: ID of each client; 2) LIMIT_BAL: Amount of given credit (in NT dollars); 3) SEX: Gender (1=male, 2=female); 4) EDUCATION: (1=graduate school, 2=university, 3=high school, 4=others, 5 and 6=unknown); 5) MARRIAGE: Marital status (1=married, 2=single, 3=others); 6) AGE: Age in years; 7) PAY_0, PAY_2, ..., PAY_6: Repayment status in September, August, ..., April (-1=pay duly, 1=payment delay for one month, 2=payment delay for two months, …, 9=payment delay for nine months and above); 8) BILL_AMT1, BILL_AMT2, ..., BILL_AMT6: Amount of bill statement in September, August, ..., April (in NT dollars); 9) PAY_AMT1, PAY_AMT_2, ..., PAY_AMT6: Amount of previous payment in September, August, ..., April (in NT dollars); 10) default.payment.next.month: Default payment (1=yes, 0=no).

2. Data Set

Please, download the dataset from:

https://www.kaggle.com/uciml/default-of-credit-card-clients-dataset

and my jupyter notebook with the description of modeling process from:

https://dataplatform.cloud.ibm.com/analytics/notebooks/v2/a5882ea9-eb7b-48ba-bc8a-0aaa23fb412a/view?access_token=1a3096882e934ac2cc1facd66a32fe4fcc38889a1d4b98ca0a1ec0d8152e7aaa

3. Data Quality Assessment

Problem: some features contain the unknown and undocumented values.

4. Data Exploration

Our initial data exploration consists of:

1) missing values and wrong measurements

2) some data cleansing

3) correlation matrix

5. Data Visualization

To better understand the nature of features and their relationship to the target, I built a lot of histograms and box plots.

6. Feature Engineering

The most obvious features that can be added to a dataset are:

1) PAY_sum -- the total number of overdue payments;

2) PAY_sum_w -- the "weighted" total number of overdue payments (recent overdues have more weight);

3) sum and some statistics for BILL_AMT (min, max, mean, std);

4) sum and some statistics for PAY_AMT (min, max, mean, std).

Since I will use gradient boosting (which is a “tree” model), there is no need to perform normalization or scaling of the dataset.

7. Selection of Model Performance Indicator

Since we have a binary classification problem and the dataset is unbalanced by the target, we will use the ROC AUC metric. Note that banks mainly use ROC AUC or Gini to assess the quality of such models.

8. Machine Learning Algorithm

Among the various implementations of gradient boosting, XGBoost and LightGBM are very popular now, and I use them.

9. Model performance between different models

To assess the quality of models, I divide the dataset into train (5/6) and test (1/6) parts, and to select hyperparameters on the train part, 5-fold cross-validation is used. On the test part, LGBMClassifiers (ROC AUC = 0.789) is a little better than XGBClassifier (ROC AUC = 0.787). For the bank credit default, ROC AUC = 0.79 is very good value.