Charles F. Dolan School of Business
BA 545 – Data Mining, Competition #2, Spring 2019

Default of Credit Card Clients

Overview of the Competition

Understanding the default (cf. https://en.wikipedia.org/wiki/Default_(finance)) of credit card customers is an interesting issue in the banking industry. Default payment refers to the failure of meeting the legal obligations of a loan – in this context, it means that the customer cannot make the payment of the credit card expenditure (on time). Previous analyses have been focused on analyzing the default scenario using demographic features (e.g. gender, age) and credit/payment features (e.g. amount of given credit (credit line), past 6-month historical payments) of the customers.

In this competition, you are going to design, implement, and deploy a classification analysis to predict whether a credit card customer will default the payment. Your client is seeking advanced and novel methods to prepare the collected data, to prepare data for the classification purpose. You will focus on all the phases in the CRISP-DM model, aside from model deployment. Advanced modeling techniques, interactive presentation skills, and elicitation of the analytical problem will be needed for extra points.

General Information about the Competition

- Started: 12:00 am, Tuesday, March 19th, 2019 ETC
- Data Audit Report Due: 11:59 pm, Monday, April 8th, 2019 ETC
- Initial Data Models Due: 11:59 pm, Monday, April 22nd, 2019 ETC
- Final Presentation and Report Deadline: 12:00 am, Monday, May 6th, 2019 ETC – Final Presentations in class
- Points: 350 points in total (50 points for data audit report, 100 points for initial data models, 200 for final presentation and final report)

Competition Rules

Please find the generic competition rules below. Dr. Tao reserves all the rights to further explain the rules.

Participation Rules

- All participants have to be in groups; every group contains 3 students.
- Privately sharing data, codes, or Modeler streams outside of the groups is not permitted – once violated, the group will be disqualified from the competition.
- Group leaders have full authority over the group – the communications between groups, or between groups and Dr. Tao should only be conducted by group leaders.

Submission Rules

- Only one (1) submission can be made by each group.
- Unlimited submissions can be made before the checkpoint deadline – which will not be graded.
- No late submissions (after submission deadline) will be accepted.
- Each submission contains (fail to submit any part may lead to the disqualification of the competition):
    - The complete analytical report (cf. https://github.com/DrJieTao/IS540-Project-2/blob/master/Decision%20Tree%2C%20Random%20Forest%20%26%20SVM.ipynb);
    - Final Processed Data;
    - Final Presentation Slides.

Evaluation Rules

- Submissions are evaluated on two evaluation metrics of the modeling results; each part takes up to 100 points:
    - *Prediction Accuracy*: we will use F-1 score as the measurement of how close the predicted values are to the factual outputs – please refer to this Wikipedia article for the computation of F-1 score (https://en.wikipedia.org/wiki/Precision_and_recall):

$$\text{precision} = \frac{true\ positive}{true\ positive + false\ positive}, recall = \frac{true\ positive}{true\ positive + false\ negative}, f1score = 2 \times \frac{precision \times recall}{precision + recall}$$

    - *Predictive Power*: we will use ROC curve (Receiver Operating Characteristics), or similarly, AUC (Area Under Curve), is a graphical representation that demonstrate the predictive performances of models. Note that ROC/AUC can only be applied to **binary targets**. The rule-of-thumb for ROC is as follows (**the higher the better**):

AUC	Interpretation
1.0	Perfect test
0.9 to 0.99	Excellent test
0.8 to 0.89	Good test
0.7 to 0.79	Fair test
0.51 to 0.69	Poor test
0.5	Worthless test

- Submissions are ranked base on MAE and AUC, respectively. The rank and associated points can be found below (tied rank is allowed):

Rank	Points
1	96
2	93
3	90
4	87
5	84
6	80

- Every submission will be evaluated by Dr. Tao and at least one (1) independent judge – all disagreements need to be resolved before release the results to the participants.

- Extra point opportunities: each group can make use of up to one (1) extra points opportunity – each opportunity worth up to 20 extra points:
    - Advanced modeling techniques: if you use advanced modeling skills, such as Gradient Boosting, Lasso Regression, Random Forest Regression, XGBoost, Elastic Net, or Deep Neural Networks, you can be awarded up to 20 points;
    - Other advanced techniques: using cross-validation, tuning/searching for model hyperparameters, or ensemble learning, you will receive up to 20 extra points;
    - Interactive presentation skills: If you decide to use an interactive scenario during your presentation – for instance, conversation between customer/banker; or between consultant/manager - you will receive up to 20 extra points;
    - You need to notify Dr. Tao your group's decision regarding the extra point opportunities **no later than** 11:59 pm, Monday, April 8th, 2019 ETC.

Competition Guidelines

Research Question

The overarching research question is “What are the determinants of the default of credit card customers?”

Data Dictionary

See the data page (cf. https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients). You will need to conduct an EDA to understand the data better.

Suggested Tasks

Following tasks are normally conduct in a classification project. Not all steps are required/available in this particular dataset – and not in this particular order. Some additional step(s) might be required – which you might need to research on. Keep in mind that your decisions on different strategies below will determine the results.

1. *Business Understanding*: frame your analytical questions: can you have side analytical problems from this project? What is the unit of analysis (customers?)? What is the population of interest (do you need some other information)?
2. *Data Understanding*: EDA is very important – you can reuse a lot of your code from competition #1 – also, Kaggle has a large number of EDA kernels. Also, you should consider how to split your data (fixed training/testing split, or cross-validation). Additionally, you should consider do you need additional data in this analysis.
3. *Data Preparation*: again, you can use a ton of the code from competition #1 here – several things to highlight here, including encoding, feature engineering. Also, some of the features are time-series data (X6 – X23) – how to preserve the sequential information in them is very important.
4. *Modeling*: use as many models as possible is recommended – so that you can look for the best model.
5. *Evaluation/Optimization*: if you want to win, optimization is a must-have. Strategies such as GridSearch, Stacking/Ensemble, and other techniques are worth consideration.

Additional Rules

- If you decide to include/exclude certain variable(s)/observation(s) from the dataset, you will need to get an approval from Dr. Tao;
- You will need to report your workload assignment (within the group) in the Milestone report and the final report – non-equal workload assignment will lead to the penalty to the whole group;
- If cross-group collaboration is identified, both groups will be disqualified from the competition (result in 0 points for this part of the class).