Data Analysis Competition for Undergraduate Teams: University of Georgia
Credit Card Approval Model

March 24 – April 15, 2021

1. Objectives

This competition will test your skills related to data analysis, model development, interpreting the results, documentation, and presentation. You have to analyze the data available in eLC, develop multiple models using different techniques, interpret the results, and document the analysis as well as findings. The data used in this study are simulated and do not correspond to any real accounts.

2. Project Description:

Financial institutions that lend to consumers rely on models to help decide on who to approve or decline for credit (for lending products such as credit cards, automobile loans, or home loans). In this project, your task is to develop models that review credit card applications to determine which ones should be approved. You are given historical data on response (binary default indicator) and 20 predictor variables from credit card accounts for a hypothetical bank XYZ, a regional bank in the south-eastern region. There are three datasets in eLC: a training dataset with 20,000 accounts; a validation dataset with 3,000 accounts, and a test dataset with 5,000 accounts. Information about the variables are given in the Appendix.

As part of the competition, you are asked to do the following and also address specific questions below:

- Conduct an exploratory analysis of the data, provide a summary, do any necessary data pre-processing in preparation for modeling;
- Develop and fit a logistic regression (LR) model, assess its performance, and interpret the results;
- Develop an additional model based on a machine learning (ML) algorithm selected from one of the following: Random Forest, Gradient Boosting (XGBoost or another implementation), or Feedforward Neural Network; assess its performance, and interpret the results. Make sure to explain why you chose this particular algorithm.
- Compare the results from the ML algorithm with those from logistic regression model and discuss their advantages and disadvantages; select one of these models for credit approval; and describe the reasons for your selection;
- Describe how you would use it to make decisions on future credit card applications.
- Do customers who already have an account with the financial institution receive any favorable treatment in your model? Support your answer with appropriate analysis.
- Suppose a credit card application is rejected using your model, and the applicant asks you to provide an explanation on why it was rejected. How would you explain the results to the customer?

The analysis and model development can also cover any other considerations about the data or models that you deem important.

3. Deliverables:

Please submit the following:

a) A report (pdf file) that describes all important steps in your data analysis, model development, model interpretation (for example, which predictors are important, what are the input-output relationships, are there any other interesting structure in the model, etc.), comparison of the models, and answer to the specific questions in addition to justification for your final model selection. The body of the

report should be no more than 15 pages in length (font size 11 and spacing 1.2). In addition, you can include an appendix of no more than 5 pages that contains additional tables and figures. Include the important figures, tables, and discussion in the body of the report.

b) The codes you used for the analysis should have brief but adequate annotations so that we can review it. Using a format like Jupyter Notebook would be ideal. Indicate clearly the software packages and versions (if appropriate) that you used for the analysis;

c) A presentation deck (pdf file) with no more than 12 slides that summarizes your results and conclusions. Note that you will be presenting these results to a panel of judges. Make sure that the presentation is accessible to a general audience. The judges are familiar with the background on credit card application but are not necessarily familiar with the technical details in your models.

d) You are allowed to review textbooks, published papers, websites, and other open literature in preparing for this case study. Note, however, that the material you submit in your report must be based on your own analysis and writing. If you relied on published scholarly work and open-source software for your analysis and findings (beyond what is generally known), you should provide references at the end of the report.

4. Appendix: Description of Dataset

i) Three datasets are attached: a) training dataset A with 20,000 accounts; b) validation dataset B with 3,000 accounts, and c) test dataset C with 5000 accounts. Use dataset A for developing your model, dataset B for hyper-parameter tuning for machine learning model, and dataset C for predictive performance assessment.

ii) Description of variables:

VARIABLE NAMES USED IN THE DATASET	DESCRIPTION OF VARIABLES
Response: Default_Ind	Indicator of Default: Binary: 1 = account defaulted after an account was approved and opened with bank XYZ within a period of 18 months; 0 = not defaulted; (Default means no payments for 3 consecutive months)
Predictors: Applicant's attributes derived from information available from credit bureaus at the time of application	
tot_credit_debt	Total debt (amount owed by applicant at the time of application) on all of their credit products (credit cards, auto-loans, mortgages, etc.)
avg_card_debt	Average monthly debt (amount owed by applicant) on all of their credit cards over last 12 months
credit_age	Age in months of first credit product ((credit cards, auto-loans, mortgages, etc.) obtained by the applicant
credit_good_age	Age in months of first credit product obtained by the applicant that is currently in "good" standing (no past due payments)
card_age	Age in months of first credit card obtained by the applicant
non_mtg_acc_past_due_12_months_num	Number of non-mortgage credit-product accounts by the applicants that are 30 or more days delinquent within last 12 months (Delinquent means payment not made)
non_mtg_acc_past_due_6_months_num	Number of non-mortgage credit-product accounts by the applicant that are 30 or more days delinquent within last 6 months
mortgages_past_due_6_months_num	Number of mortgages by the applicant that are delinquent within last 6 months
credit_past_due_amount	Total amount of money that is currently past due on all credit accounts
inq_12_month_num	Number of credit inquiries in last 12 months (An inquiry occurs when the applicant's credit history is requested by a lender from the credit bureau. This occurs when a consumer applies for credit.)
card_inq_24_month_num	Number of credit card inquiries (on applicant's credit) in last 24 months
card_open_36_month_num	Number of credit cards opened by applicant in last 36 months
auto_open_36_month_num	Number of auto loans opened by applicant in last 36 months
uti_card	Utilization on (all currently available) credit card accounts (Utilization is ratio of balance divided by credit limit)
uti_50plus_pct	Percentage of open credit products (accounts) with over 50% utilization
uti_max_credit_line	Utilization of credit product (account) with highest credit limit
uti_card_50plus_pct	Percentage of open credit cards with over 50% utilization
ind_acc_XYZ	Indicator: 1 if applicant already has some account (checking/savings, etc.) with the bank XYZ; 0 otherwise

rep_income	annual income (self-reported by applicant and not verified)
States	Residence state of applicant (AL, FL, GA, LA, MS, NC, SC)