PROJECT 1: LOAN PREDICTION
DSCI 451
Fall 2024

Financial institutions want to loan money to people who are likely to pay it back. They use a variety of information about an individual to predict whether or not that individual is likely to pay back a loan. People who were approved for a loan are judged more likely to repay the loan. People who are denied are judged more likely to default (i.e., more likely not to repay the loan).

On the CourseSite entry for this assignment is a file titled state_GA_actions_taken.csv. This file includes historical data about home loan applications that were approved or denied. Use this data set to create a binary model that predicts loan repayment. Essentially, the model will use information from this historical data to predict which people in the future are likely to repay their loan (because they were approved for a loan in the past) and those who are more likely to default (because they were denied a loan in the past).

Everything necessary to complete this assignment should have been learned in DSCI 310 (or in equivalent prior coursework). Thus, this assignment provides an opportunity for students to showcase their previously acquired skills.

DOCUMENTATION

These data were made available by the US Consumer Financial Protection Bureau (CFPB) as required by the Home Mortgage Disclosure Act (HMDA). The data file for this project includes a subset of all home loans for which people applied in the state of Georgia during 2023. The action_taken column indicates whether the loan was originated (i.e., the loan was made to the applicant(s)) or denied:

- action_taken == 1, the loan was originated
- action_taken == 3, the loan was denied

Other values for action_taken correspond to other outcomes excluded from this data set (e.g., the lending institution approved the loan to the applicant(s), but the appraised value of the property was less than the requested loan; the lending institution approved the loan, but the applicant(s) withdrew the loan application). These data only include loans that were approved and accepted or that were denied.

The remaining columns describe attributes of the applicant(s), the lending institution, the type of loan, the property, etc. Full details are available from the CFPB's website: https://ffiec.cfpb.gov/documentation/publications/loan-level-datasets/lar-data-fields/.

DELIVERABLES

Students should submit their work as an R Notebook (an .Rmd file). The notebook should load all required libraries and should install any non-standard libraries. The notebook should include code for all analytic steps taken – data cleaning, preprocessing, model training, performance assessment, etc. The notebook should also include justification for all analytic decisions – why the data cleaning, preprocessing, model training, performance assessment, etc. were done in the way

they were. These justifications should accompany the relevant code in text blocks (i.e., outside of the ```{r} [...] ``` code blocks). Justifications can (and likely should) make use of concepts and examples covered in course readings and class discussion. Justifications can also be based on preliminary or exploratory analyses of the data, including the use of visualizations.

In some ways, this Project is comparable to the Final Project from DSCI 310, but with a less extensive write-up and a less complex task. Furthermore, using an R Notebook will provide the opportunity to interweave the analysis itself with the written description thereof. The description and justifications provided in the notebook should provide sufficient detail for another student who successfully completed DSCI 310 to understand what was done and why.

The assignment submission on CourseSite should include both the .Rmd file and the .html file for the notebook.