Commercial Bank Customer Retention Prediction

APSTA-GE 2401: Statistical Consulting Seminar, Mid-Presentaion

Tong Jin, Andy Tan, Zixuan Zhou

New York University

10/26/2020

Topics

- Introduction
  - Overview
  - Tasks
  - Details
- Project Plan

Introduction

Overview

Commercial Bank Customer Retention Prediction

- Active competition on Data Castle
  - Initial: 10/26
  - 1st Round: 10/26 - 12/10
  - 2nd Round: 12/11 - 12/25
  - Final Round: 12/26 - 1/10/2021
- Using the following machine learning techniques to predict whether a customer will **retent** (or **churn**):
  - Exploratory Data Analysis (EDA)
  - Data Mining
  - Supervised Learning

Tasks

Banks need a better understanding of customer demands as they expanding their business. Specifically, banks are interested in learning about customer's churn rate and the change of their financial status. The purpose of predicting customer's retention probability is for precision marketing in order to prevent customers from churning.

1. Build a model to predict retention probability based on real-world data records.

2. Present a business solution for precision marketing based on model results.

Details – Data

- Linked data tables in `.csv` format
  - `cust_no`: ID
- Two datasets:
  - Train
    - `x_train`: features
    - `y_train`: label
  - Test
    - `x_test`: additional data records with label removed

Details – Features

There are 5 feature categories marked with different initials:

- `X`: end-of-month balance, `ncol_X = 8`
  - structural, checking, savings, CDs, funds, loans, ...
- `B`: customer behavior, `ncol_B = 7`
  - transfer (mobile, branch), transaction data, frequency, ...
- `E`: big events, `ncol_E = 18`
  - new account opening, mobile app activation, large transfer, ...
- `C`: savings, `ncol_C = 2`
  - number of saving products and amounts
- `I`: customer information, `ncol_I = 20`
  - age, sex, level, occupation, income, indicators (marital, mobile app, mobile pay) ...

Total number of features: 55

Details - Dimensions

Train:

Number of Rows:

```
## [1] 465441
```

Train validation:

```
## [1] 145296
```

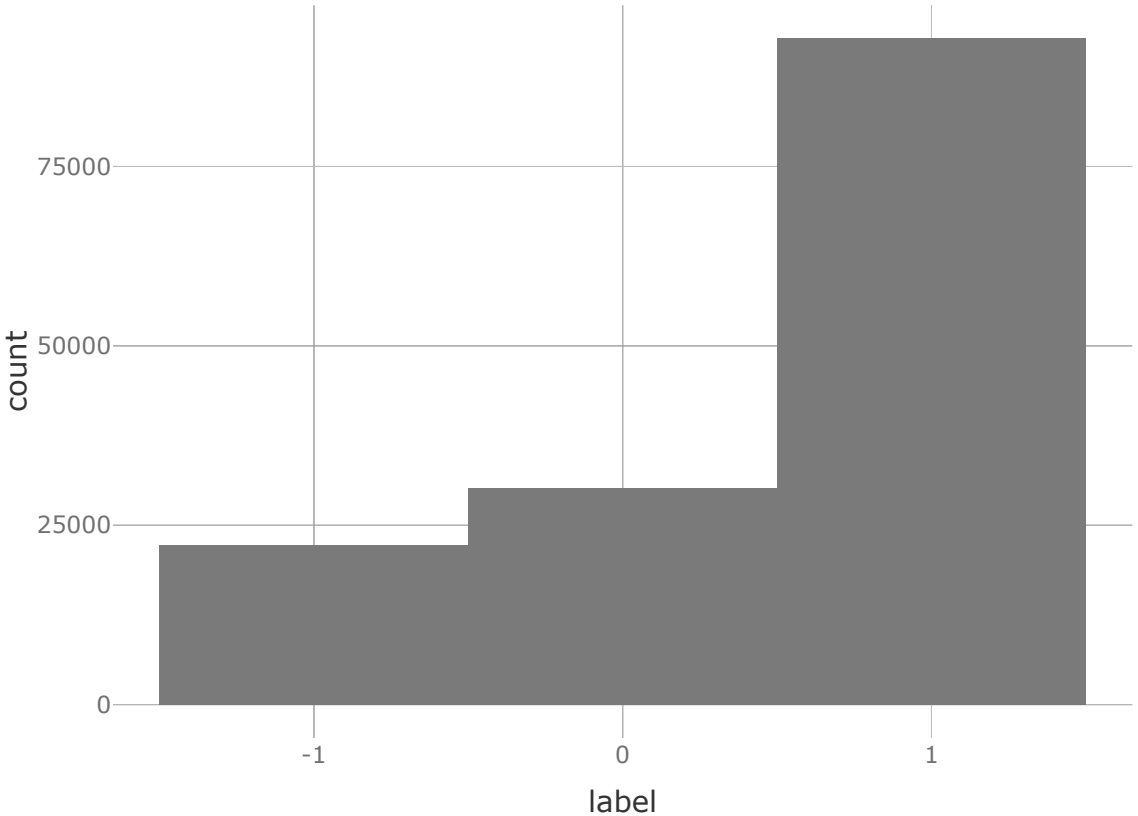
Test:

Number of Rows:

```
## [1] 76722
```

Number of Columns: 1 (label)

Details – Train Churn Rate



Grading

Cohen's Kappa Statistic

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

where `Pr(a)` is the predicted label, `Pr(e)` is the expected label value.

- `< 0`: No agreement
- `0 - 0.2`: slight
- `0.21 - 0.4`: fair
- `0.41 - 0.6`: moderate
- `0.61 - 0.8`: substantial
- `0.81 - 1`: perfect

Project Plan

Project Plan

- Programming Language: `Python`
- Structure:
    - Data processing:
        - Cleaning
        - EDA
        - Feature Engineering (SVD)
    - Baseline Model: Logistic regression with elastic net
    - Test model:
        - Random forest
        - Gradient boosting machine
        - Multilayer perceptron
    - AUC for selection
    - Feature importance

Thank you!