

Commercial Bank Customer Retention Prediction

APSTA-GE 2401: Statistical Consulting Seminar,
Preliminary Results

Tong Jin, Andy Tan, Zixuan Zhou

New York University

11/24/2020

Topics

- Background
- Exploratory Data Analysis
 - Customer Demographic Information
 - Details
- Project Plan

Background

Introduction

This project applies supervised machine learning techniques to predict the customer retention preference of a commercial bank in China. The dataset is derived from an active competition on [Data Castle](#).

Overview

- 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

As we embracing the era of digital finance, many financial institutions adapt themselves to a data-based business mode by applying various data mining and machine learning techniques to their services. The Xiamen International Bank ("The Bank") has been heavily invested in data-driven financial services in recent years. As the bank expanding its business, it requires a better understanding of customer demands, especially revenue-related preference. Coordinated with Data Castle, the bank launched this competition, aiming to invite data scientists and statisticians to provide solutions regarding a real-world service problem: predicting customer retentions.

The task of this competition is to design and implement a supervised machine learning algorithm to predict whether or not customers will continue their businesses with the bank in the near future. Specifically, the bank is interested in learning about customer's [churn rate](#) and the probability of changing their financial status. The purpose of predicting customer's retention probability is for precision marketing that prevents the bank from losing revenues.

To solve this problem, our team accomplished the following tasks:

1. Select and build a model combination that predicts retention probability based on the dataset.
2. Based on model results, present a professional business proposal for precision marketing.

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

Project Plan

Project Plan

- Programming Language: Python
- Structure:
 - Data processing: **[Completed]**
 - Cleaning **[Completed]**
 - EDA **[Completed]**
 - Feature Engineering (SVD) **[In Process]**
 - Baseline Model: Logistic regression with elastic net **[In Process]**
 - Test model: **[In Process]**
 - Random forest **[In Process]**
 - Gradient boosting machine **[In Process]**
 - Multilayer perceptron **[In Process]**
 - AUC for selection **[To Do]**
 - Feature importance **[To Do]**

Contact

Team

Team Name: A3SR

Team Members:

- [Tong Jin](#), NYU
- [Zheng Tan](#), NYU
- [Zixuan Zhou](#), NYU