Taiwan Credit Card Default Predictive Modeling

Ryan Metz

For this analysis I'm choosing a Credit Card default dataset from Taiwan, released in 2016 by I-Cheng Yeh. The business problem I chose was to find characteristics of those that default or not and create a model to accurately predict default or not.

Variable Definitions

X1: Amount of Given Credit

X2: Gender (1= Male, 2=Female)

X3: Education (1= graduate school, 2= university, 3= high school, 4=other)

X4: Marital Status (1= married, 2= single, 3=others)

X5: Age (year)

X6 - X11: History of past payments by month. X6 starting in September. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above.

X12 - X17: Amount of bill statement (dollars). X12 starting in September.

X18 - X23: Amount of previous payment (dollars). X18 starting in September.

Y: Default (1 = Yes, 0= No)

Introduction

I wanted to predict default or not default when considering variables from a credit account. This can affect the credit card business bottom line by increasing the amount loaned out to potential clients and prevent defaults on account loans. This could also lead the company to advertise to potential clients that fit the model of not defaulting and prevent those that have a higher chance of default. The company can also adjust their account credit thresholds to prevent default loss.

Data Cleaning

For my data cleaning process, there weren’t any missing data points (NAs) and there were 245 rows with values outside of the definition range in education variable.

A screenshot of a computer screen

Description automatically generatedExploratory Data Analysis

We had 10 categorical variables and 14 numerical variables. The categorical variables included education, gender, martial status, etc. Numerical variables included monthly account balance, age, amount of credit given, etc.

A screenshot of a computer code

Description automatically generated I’ve included a table to show the account credit given to an account on average grouped by if they’ve defaulted on their account or not. 0 representing no default, 1 representing default. You can see those that don’t default on average have a higher account balance than those that do.

A graph of a number of accounts

Description automatically generatedIn this chart you can see the distribution of accounts at the company and their account holder age, grouped by the target variable; if they’ve defaulted or not default. You can see the bulk of account holders are between 20 and 40 years old, skewing right. I want to highlight that there is a imbalance in target variable classes, which we will address later in the model creation process as this could lead to weaker results.

A graph of a credit availability distribution

Description automatically generatedThis is a boxplot showing the distribution of the education groups and their credit amount by if they defaulted or not. Both Graduate and Other education groups had higher mean Credit Balance. Accounts with high school education had the lowest amount of credit balance, followed by accounts with university education. It looks to seem there were many outliers in University and High School Education groups which could show some interaction between credit balance and another factor, which would give the account higher credit availability.

Modeling

For my supervised approach, I ran a classification regression with XGBoost to create a predictive model with default as my target variable, after comparing to Random Forest and KNN supervised models.

A screenshot of a computer

Description automatically generated

My predictive model accuracy was 81.45%, which I knew needed to be improved. I remembered from the EDA that there was a class imbalance within the dataset. Therefore, I deployed the SMOTE (Synthetic Minioirty Over-sampling Technique) to help balance the dataset. This will create synthetic data points to bring the total count of the underrepresented category to match the majority.

I’ve included the class imbalance in the table below to the right. This shows that the not default classification has around 3.5x the number of observations compared to the default class in the target variable.

A screenshot of a computer screen

Description automatically generatedYou can see, by balancing the classes we increased our predictive accuracy from 81.45% to 94.19%; a 12.74% improvement in predictive accuracy. This is very impactful for the models’ capabilities and can shrink potential costs associated with predicting a default or not default wrongly. Meaning if you gave out a higher loan limit and it defaulted or not giving out a higher loan and missing out of potential interest gain could be avoided at a higher rate. But how can we improve the model more?

A graph showing a plot of numbers

Description automatically generated with medium confidenceI wanted to introduce an unsupervised method into the model with clustering. This method leverages KNN (k nearest neighbors) to create unidentified variables to improve accuracy further. I used a 3 split clustering method when choosing the new variable.

A screenshot of a computer screen

Description automatically generated

After adding the clustering variable to the previous SMOTE dataset, the model improves marginally. This raised the model’s accuracy to 94.35%, only up 0.16% in predictive accuracy from the previous model. Although the slim margin of improvement, this could lead to a better result in characterizing accounts and still leads to a better overall financial impact for the company.

A screenshot of a computer

Description automatically generatedLooking at the variable importance of the last model, would imply that X1 (Credit amount given) is the lead variable when predicting a default or not. Followed by X5 (Age) and X6.5/X7.5 (Monthly account balances). Looking at these variables can help the company to identify defaults or not by looking at these variables early. Looking at the monthly balances, representing the early months in the account balance, could help the lender to adjust their credit limits to those with negative account balances or increase credit limit to those with positive account balance in the early months or expect the company to experience a loss months before the default.

Further Improvements

Ways to improve the model would to be including more observations to the dataset by capturing more client information or purchasing client information from other credit card carriers that have similar structure. Improving data collection methods to increase data accuracy of reflecting the true nature of an account. Improving the range of models with grid search during the training methods. Improving physical hardware to create more robust calculations, this would impact the number of clusters or synthetically increasing the observations, and number of training iterations.