Shangzhou Yin
U63027471

Discussion of American Express Credit Card Default Prediction

Introduction

Modern life relies on the convenience of credit cards to make daily purchases, no matter shopping online or in a physical store, which only needs customer to carry a piece of card over large amount cash. However, based on the Federal Reserve Economic Data[1], credit card delinquency rate increases from 2016 to 2019 and starts to increase again in 2021Q3(a sharp decrease in 2020 is due to Covid19). The credit card is issued by a bank or financial services that allow cardholders to borrow funds to make payments for goods and service[2], so it results in huge risk to card issuers that they do not know when the money will be returned and if the customers have the ability to return. In order to identify high credit card default users for banks and financial services and minimize the loss, Kaggle competition, American Express Default Prediction, is proposed to create machine learning models to predict whether a customer will default in the future or not. Also, this is a classification problem because of binary target variable.

Dataset

Since the dataset released by American Express involves customers' privacy and security, all 190 features are anonymized and normalized and categorized into 5 groups containing delinquency variables, spend variables, payment variables, balance variables, and risk variables. The target is a binary variable, which is resulted by observing 18 months performance window after the latest credit card statement. It is considered as a default event as noted 1 in the dataset if the customer does not pay the due amount in 120 days after their latest statement date, otherwise, it is noted as 0.

Related work

In this section, I reviewed some research papers that related to this topic. The main difference is the dataset. The credit card customer dataset of the following research paper is from April to September 2005 in Taiwan on the UCI website and each feature is known and specific.

Shenghui Yang et al. examined several data mining methods in credit card default prediction including Logistic Regression, Neural Network, Support Vector Machine, Xgboost and LightGBM[3]. They evaluated these methods by comparing AUC, Correct rate, and F1-Score and found that LightGBM has the best classification effect in the experiment.

Yashna Sayjadah et al. used machine learning techniques to predict the credit card default with the same dataset from the UCI website[4]. They applied Logistic Regression, Rpart Decision Tree, and Random Forest to the research by evaluating the Accuracy, True Positive, and AUC, and found that Random Forest had the best performance with the best accuracy and AUC.

Industrial Review

The traditional approach of determining whether a borrower will default or not highly relies on manual or system account monitoring. The account first becomes delinquent if missing a

payment for 30 days and the default happens after 6 months if not make the minimum payment due[5]. The system will report abnormal accounts with the above one or two conditions to the manager. The manager will then contact the customers to discuss their accounts' situations. The current method can only help the bank or financial services to notice the default event and reduce the loss if an account is in an abnormal condition after it happens but cannot prevent it before it happens.

The other way to reduce banks' loss is from the source which is credit card issuing. Banks or financial services usually collect and evaluate applicants' information including their current financial situation, their past loans, and payment history, and give them a score[6]. Based on the credit score, if applicants' have a high credit score, then it will be easy for them to apply for a credit card with high limits, otherwise vice versa. Meanwhile, a high credit score usually means that the customer has a lower probability to default.

Application

The purpose of this project is to design a machine learning model for American Express to challenge the current model in production. In this competition, it trains with 190 anonymized features that are divided into 5 categories. Since not all features are equally important, it is possible to find several categories that affect feature engineering. Thus, in the real world, the bank or financial service can collect more data in these categories and use them to predict the default. To use this system to fully replace the manual account monitoring is not feasible under the current situation, but with this system, it can play a role in assisting banks to assess an individual's possibility of default to improve manual approval efficiency and reduce labor cost, as well as whether issue a credit card to an individual. Combining these two approaches, it will be efficient to minimize or even prevent banks' losses in the future.

Open-Source Review

This section selects one of the notebooks from the competition with the most votes and evaluates the approach[7]. The most impressive point that author did is data preprocessing which rounding up the float column values at 0 and 1 and saving in parquet format to ensure no data loss and reduce the size of the dataset for training. Meanwhile, the author uses DeviceQuantileDMatrix to load data, which uses small GPU memory.

Methodology

Based on the research, the whole process will be divided into four main parts.

Step1: Data Preprocessing

Since the project has not been fully designed, a general data preprocessing approach will be illustrated below. First, any records with missing target column data will be removed as these are necessary for prediction. To handle missing numeric values, it is appropriate to fill them with either the median or mean value of that feature. In terms of categorical values, we can one hot encode these into separate columns so that they could be easily used as an input for the model. To address the class imbalance, undersampling, oversampling, or SMOTE can be applied to get better class distribution. Undersampling is the process of removing instances of

the majority class in order to improve class distribution. Oversampling adds copies of the minority class to balance class distribution. SMOTE stands for synthetic minority oversampling technique, which uses KNN clustering to create new minority class objects that fill in the gap between the minority and majority class.

Step2: Model Training & Internal Test

The entire dataset will be split into three sub-datasets: training set (70%), validation set (20%) and test set (10%). Since the Kaggle test is not accessible only if you upload your final work to the website, I decided to do an internal test to determine the performance of the models. Based on the open-resource and related work found online, Logistic Regression, Neural Network, Support Vector Machine, random Forest, Xgboost and LightGBM were deployed and performed well, I decided to separately apply those models to the training dataset, use validation set to determine which model will be used for the internal test dataset. The internal test accuracy will be compared with the first prize work that has a test accuracy of 80%. If the internal test accuracy does not reach the desired result, one of possible solutions is to tune parameters to improve the models. If the model reaches the desired result, the solution will be used for model test in Kaggle.

Step3: Model Test

In this final step, the final solution can be obtained from Model Training & Internal Test and will be uploaded to Kaggle to test the real test dataset. If the test result does not reach the desired result, then it will go back to research new models and repeat step 2 and 3, and the whole process stops when it reaches the desired accuracy.

Shangzhou Yin
U63027471

Reference

[1] Board of Governors of the Federal Reserve System (US), "Delinquency Rate on Credit Card Loans, All Commercial Banks," *FRED, Federal Reserve Bank of St. Louis*, Jan. 01, 1991. https://fred.stlouisfed.org/series/DRCCLACBS (accessed Aug. 30, 2022).

[2] "What Is a Credit Card?," *Investopedia*. https://www.investopedia.com/terms/c/creditcard.asp (accessed Aug. 30, 2022).

[3] "Comparison of Several Data Mining Methods in Credit Card Default Prediction." https://www.scirp.org/journal/paperinformation.aspx?paperid=87507 (accessed Aug. 30, 2022).

[4] Y. Sayjadah, I. A. T. Hashem, F. Alotaibi, and K. A. Kasmiran, "Credit Card Default Prediction using Machine Learning Techniques," in *2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA)*, Oct. 2018, pp. 1–4. doi: 10.1109/ICACCAF.2018.8776802.

[5] S. Bucci, "Credit Card Default: What to Do About It | Bankrate.com," *Bankrate*. https://www.bankrate.com/finance/credit-cards/credit-card-default/ (accessed Aug. 30, 2022).

[6] "What Do Banks Do To Practice Credit Risk Management." https://www.credolab.com/blog/what-do-banks-do-to-practice-credit-risk-management (accessed Aug. 30, 2022).

[7] "XGBoost Starter - [0.793] | Kaggle." https://www.kaggle.com/code/cdeotte/xgboost-starter-0-793#Save-OOF-Preds (accessed Aug. 30, 2022).