Name: Eagle Xuhui Ying Email: yingx22@wfu.edu Date: 12/09/2022 Section: C

Executive Summary

Problem

Suppose I have recently joined a major financial institution and am tasked with developing and comparing several machine learning models to predict "loan status", Specifically, which loans are likely to default. My task is to build, tune and evaluate several different models, predicting (loan_default), identify any outliers and explain them, explain my top 10 correct predictions of a loan default (loan default = 1), top 10 predictions of loan default = 0, and my top 10 incorrect predictions. I have been provided with two datasets. One is loan_train.csv, which is used to train and evaluate my model. Another is loan_holdout.csv, which is used to assess the accuracy of my prediction.

Key Findings

After I ran my XGBoost model, I found five following most predictors to predict loan status. The coefficient of this predictors can calculated in my logistic regression model.

1. last_pymnt_amnt: Increasing the last total payment amount can reduce the probability of loan default. When the last total payment amount received increases by $10, the log odds of logistic regression will decrease by 92.78 units. This is the most important predictor in my prediction model.

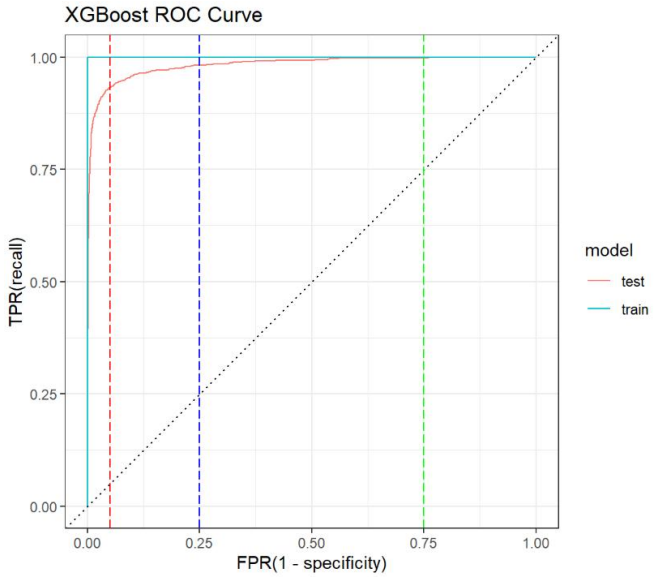
2. last_pymnt_d_year: This variable is the time difference between the last month's payment received and today, which is transformed into years in my model. It has a positive influence on my model. One year increase in this variable will result in a 29.18 unit increase in log odds. In other words, the more recent the last month's payment was received, the less likely the loan will default. This is the second most important predictor in my prediction model.

3. last_credit_pull_d_year: This variable is the time difference between the most recent month LC pulled credit for this loan and today, which is transformed into years in my model. It has a positive influence on my model. One year increase in this variable will result in a 15.83 unit decrease in log odds. In other words, the more recent the credit was pulled, the more likely the loan will default. This is the third most crucial predictor in my prediction model.

4. issue_d_year: This variable is the time difference between The month the loan was funded and today, which is transformed into years in my model. One year increase in this variable will result in a 16.23 unit decrease in log odds. In other words, the more recently the loan is funded, the more likely to loan will go into default.

5. loan_amnt: The listed amount of the loan applied for by the borrower is positively associated with the likelihood of loan default. A $10 increase in loan amounts will result in a 3.11 unit increase in log odds in my logistic regression model.

Model Performance Summary & Interpretation



The ROC Curve is a graph showing the performance of a classification model at all classification thresholds. We use AUC (Area under the ROC Curve) as the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the model's performance in distinguishing between the positive and negative classes. In my final model, the AUC is 0.9824. Since the AUC is close to 1, my model is good at predicting loan status classification.

Recommendations

1. The financial institution should pay attention to the last total payment amount, which is the most critical predictor in my prediction model. With the fewer amounts of money they paid last time, the institution should focus more on the potential risks of loan default to take measures to prevent bankruptcy from happening as much as possible. For example, they should reassess the borrows ability to pay back and limit the amount of money they can borrow at once. Also, since the listed amount of the loan applied for by the borrower is also an important default of loan default, we should focus on borrowers with large amounts of loans and take measures to limit large quantities of borrowing.

2. The financial institution also needs to pay attention to the latest month of some borrow's behaviors. One important indicator is the last month's payment received date. The institution should pay attention to borrowers who didn't pay their debts for a long time, remind them to pay their debts, and take measures to limit their future borrowing. Another significant predictor is the most recent month LC pulled credit for this loan. Therefore, the financial institution should focus on borrowers who didn't pull credit for their loans for a long time to figure out measures to reduce the risk of default loans.