IIMT2641 Introduction to Business Analytics

Assignment 4

Due: 18 Apr 2023, 00:00

In the lending industry, investors provide loans to borrowers in exchange for the promise of repayment with interest. If the borrower repays the loan, then the lender profits from the interest. However, if the borrower is unable to repay the loan, then the lender loses money. Therefore, lenders would like to minimize the risk of a borrower being unable to repay a loan.

In this exercise, we will use publicly available data from LendingClub, a website that connects borrowers and investors over the internet. The dataset is in the file *Loans.csv*. There are 9,578 observations, each representing a 3-year loan that was funded through the LendingClub.com platform between May 2007 and February 2010. There are 18 variables in the dataset, described in Table 1. We will be trying to predict NotFullyPaid, using all of the other variables as independent variables.

(a) Let us start by building a logistic regression model.

(i) First, randomly split the dataset *Loans.csv* into a training set and a testing set. Put 70% of the data in the training set. What is the accuracy on the test set of a simple baseline model that predicts that all loans will be paid back in full (NotFullyPaid = 0)? Our goal will be to build a model that adds value over this simple baseline method.

(ii) Now, build a logistic regression model that predicts the dependent variable NotFullyPaid using all of the other variables as independent variables. Use the training set as the data for the model. Describe your resulting model. Which of the independent variables are significant in your model?

(iii) Consider two loan applications, which are identical other than the fact that the borrower in Application A has a FICO credit score of 700 while the borrower in Application B has a FICO credit score of 710. Let Logit (A) be the value of the linear logit function of loan A not being paid back in full, according to our logistic regression model, and define Logit (B) similarly of loan B. What is the value of Logit (A) minus Logit (B)?

(iv) Now predict the probability of the test set loans not being paid back in full. What is the accuracy of the logistic regression model on the test set using a threshold of 0.5? How does this compare to the baseline model?

(b) LendingClub assigns the interest rate to a loan based on their estimate of that loan's risk. This variable, IntRate, is an independent variable in our dataset. In this part, we will investigate just using the loan's interest rate as a "Smart baseline" to order the loans according to risk.

(i) Using the training set, build a logistic regression model that predicts the dependent variable NotFullyPaid using IntRate as the only independent variable. Is IntRate significant in this model? Was it significant in the first logistic regression model you built? How would you explain this difference?

(ii) Use the model you just built (with only one independent variable) to make predictions for the observations in the test set. What is the highest predicted probability of a loan not being paid back in full on the test set? How many loans would we predict would not be paid back in full if we used a threshold of 0.5 to make predictions?

Variable	Description
CreditPolicy	1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
Purpose.CC	1 if the purpose of the loan is related to Credit Card, and 0 otherwise.
Purpose.DC	1 if the purpose of the loan is related to Debt Consolidation, and 0 otherwise.
Purpose.Edu	1 if the purpose of the loan is related to Education, and 0 otherwise.
Purpose.MP	1 if the purpose of the loan is related to Major Purchase, and 0 otherwise.
Purpose.SB	1 if the purpose of the loan is related to Small Business, and 0 otherwise.
IntRate	The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Note that borrowers judged by LendingClub to be more risky are assigned higher interests.
Installment	The monthly installments ($) owed by the borrower if the loan is funded.
LogAnnualInc	The natural log of the self-reported annual income of the borrower.
Dti	The debt-to-income ratio of the borrower (amount of debt divided by annual income).
Fico	The FICO credit score of the borrower.
DaysWithCrLine	The number of days the borrower has had a credit line.
RevolBal	The borrower's revolving balance (amount unpaid at the end of the credit billing cycle).
RevolUtil	The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
InqLast6mths	The borrower's number of inquiries by creditors in the last 6 months.
Delinq2yrs	The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
PubRec	The borrower's number of derogatory public records (bankruptcy fillings, tax liens, or judgements).
NotFullyPaid	1 if the loan was not paid back in full, and 0 otherwise.

Table 1: Variables in the dataset Loans.csv.