CREDIT RISK SHINNY APP.

OBJECTIVE

The goal of this application is to check how the quality of the predictive credit risk model changes when some parameters are modified.

INTRODUCTION

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 face the problem of predicting the risk of a borrower being unable to repay a loan.

DATASET

To address this problem, we will use publicly available data from LendingClub.com, a website that connects borrowers and investors over the Internet. This dataset represents 9578 3-year loans that were funded through the LendingClub.com platform between May 2007 and February 2010. The binary dependent variable *not_fully_paid* indicates that the loan was not paid back in full (the borrower either defaulted or the loan was "charged off," meaning the borrower was deemed unlikely to ever pay it back).

To predict this dependent variable, we will use the following independent variables available to the investor when deciding whether to fund a loan:

- *credit.policy*: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
- *purpose*: The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").
- *int.rate*: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
- *installment*: The monthly installments ($) owed by the borrower if the loan is funded.
- *log.annual.inc*: 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.
- *days.with.cr.line*: The number of days the borrower has had a credit line.
- *revol.bal*: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
- *revol.util*: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).

- *inq.last.6mths*: The borrower's number of inquiries by creditors in the last 6 months.
- *delinq.2yrs*: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
- *pub.rec*: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

ALGORITHM

We use logistic regression trained on a training set to predict the dependent variable *not.fully.paid* using all the independent variables. The observations included in the training set depend on which seed to use in R and which proportion of the whole dataset is going to be used to train the algorithm. This two parameters change the ROC curve, a standard estimator of the performance of a binary classifier system as its discrimination threshold is varied. The algorithm output is the likelihood (probability of default in this case) for each observation to be classified as *not.fully.paid*. The third parameter of the application change the threshold we use to classify an observation as *not.fully.paid*. This changes the accuracy of the algorithm, but the ROC curve remains the same. The accuracy of a classification algorithm is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

INPUT PARAMETERS

As we mentioned before, this application has three input parameters:

- **Random seed for sample splitting**: the seed we set before to split the dataset between a training set and a test set.
- **Sample proportion in the training set**: the proportion of the whole dataset included in the training set. The default value is 0.7 (70%).
- **Likelihood threshold to classify as default**: the output probability for an observation to be classified as *not.fully.paid*. The default value is 0.5.

OUTPUT

As a result of the algorithm execution a plot will be shown in the right panel of the web page. In this plot we can see three main elements:

- ROC curve and its comparison with the base model represented as a line with slope equals to 1 (randomly classify any observation as *not.fully.paid*).
- Accuracy of the algorithm as it was defined above.
- AUC (area under the curve) is the area under the blue line in the ROC curve. A perfect classification algorithm would have a AUC equals to 1.

SCREENSHOT

Credit risk. Probability of default by logistic regression

Random seed for sample splitting

25 100 200

Sample proportion in the training set

0.5 0.7 0.9

Likelihood threshold to classify as default

0.25 0.5 0.75

ROC plot for a logistic regression model for loan default