Homework 3, STATS 315A

Stanford University, Winter 2019

Joe Higgins

Question 6

A company in Chile uses crowd-sourcing to fund loans to the public, as a means to offer relief from the high bank interest rates. The data in this challenge consists of historical loan records for a case-control sam- ple of 3000 past customers. The variables characterize some aspects of the loan, such as duration, amount, interest rate and many other more technical features of the loans. There are also some qualitative variables such as reason for the loan, a quality score and so on. One of the variables — our response — is ”default”, a 0/1 variable indicating whether or not the borrower has defaulted on their loan payments.

The company would like to build a default risk score so that they can target high-risk customers early and perhaps preempt the default event, which ends up costly for all involved.

The default rate experienced by this company is 7%. You are provided with a training set `loan_train.csv` which represents a sample of 1000 defaulters, and 2000 non-defaulters, and contains 30 features and the binary outcome ”default” (in the first column). There is also a file `loan_testx.csv` which consists of a random sample of 10000 other customers from the general pool. For these you are provided only the 30 features.

Your job is to build a risk score — probability of default — for each customer in the test set. You may use any of the tools discussed in the lectures in this class. You may not use tools not discussed in this class, such as deep learning, random forests or boosting. You should produce a writeup describing what you did, and how you selected your final model. Give some indication which variables were important in the calculation of your risk score. You will also submit a simple file with 10000 lines, and on each line is your predicted risk estimate for each test customer, in the same order as in `loan_testx.csv.`