Untitled

Niall Trinder - R00088254

3 December 2018

Introduction

For this assignment, we're working on the premise that a manager from a local financial institution, Olivia, has contacted us to help her assess the credit worthiness of future potential customers.

She has provided us with the following data:

- Data set 1; 793 past loan customers with 14 attributes, labelled based on Good/Bad Credit Standing.
- Data set 2; 10 potential loan customers with no labels.

Our goal is to explore and assess the past loan customer data set, chose an appropriate model to train using this data and then use that model to assess the potential customer data set to aid Olivia's financial institution minimize the risk of loaning money to a customer who may default on the loan.

Exploratory Data Analysis (EDA)

ROC Curves (~280 words)

A ROC curve is a plot of the "true positive rate" on the y-axis and the "false positive rate" on the x-axis for every possible classification threshold.

The true positive rate (or sensitivity) is the number of true positive classifications divided by the total occurances of the true class. For instance; if you were to classify a patient has 'having the disease' or 'not having the disease' then the true positive rate would be the number of patients *correctly* classified as having the disease divided by the true total number of patients who have the disease.

The false positive rate (or specificity) is simply 1 - sensitivity, or, the number of false positive classifications divided by the total occurances of the false class.

The ROC courve visualises all possible thresholds were as missclassification rate is the error rate for a single threshold.

A diagonal line from (0, 0) to (1, 1) on the graph would represent a model that does no better than guessing.

You can use the ROC curve to quantify the performance of the classifier by giving a higher rating to better performing models. To do this we evaluate the area under the curve and express it as a percentage of the total area.

Note that most problems in the real world don't have balanced classes and that this does not affect the ROC curve. Also ROC curves are useful een if the predicted values are not properly calibrated.