FinTech Analytics: Data-driven Credit Modeling
Credit Default Modeling Project

Overview

The objective is to develop, individually or in small teams, a default prediction model using a realistic data set. This is not a programming project, though students will find it useful to use the programming language R to estimate the model. Much of the basic com-mand-line code that is required for a basic model is given in the attached project description, along with examples. Successful completion of the project will include presenting and documenting the model in a realistic setting.

Objective

You are being asked to estimate and test a simple model of corporate default. The model should take as input financial statement data on each of the firms being evaluated, and produce as output a one-year probability of default (and any ancillary measures you wish to include) for each firm. You will be provided a development data set and you may use either R or Matlab to estimate and test the model. It is strongly recommended that you use R. The model will be tested on a holdout sample that the instructor maintains. You may work individually or in small groups.

Deliverable Items

1. A PowerPoint or Keynote deck (10-15 slides) describing:
    (a) your data
    (b) the definitions of the variables you included
    (c) the relative importance of the variables in the model
    (d) the functional form of the models you considered
    (e) any data preprocessing you performed
    (f) the details of your final model
    (g) your testing results, and
    (h) a technical appendix (if needed)
2. The source code you used to estimate and test your model
3. Source code that takes as input a data file of the same format as the development sample and produces outputs in the form of probabilities of default for each firm in the holdout sample
4. A file containing PDs for a the validation data (holdout sample) you will be given

Software

You must use the R language to estimate and test your model.

R software downloads: `http://cran.us.r-project.org`
(You will also find a large repository of statistical routines including the caTools package for ROC analysis.)

RStudio R IDE: `http://www.rstudio.com/ide/download/`

This is an integrated development environment, currently also free, which makes loading data, installing packages and overall development generally easier than in native R. I recommend you try this out as I have found it streamlines the model estimation process.

Data

The data-set is titled, bankdata.in.new, has extension .RData and can be found in the Assignments and Resources section of the course web-site. This data-set consists of 117 columns. A data dictionary, Data Dictionary.xlsx, is available in the Assignments and Resources section of the course web-site. Please pay close attention to the bolded fields in the data dictionary. Use your judgment regarding which variables you want to use when building your model.

You may wish to sample the data set down for initial experiments before using larger portions for estima-tion of your final models. It is also strongly recommended that you split your sample into both an estimation and a test-ing sample to allow you to evaluate the robustness of your model before you submit it.

Testing

After you finish your model, you will be given a new data set, in the same format as the first one, but with no default flags. You will use your model to produce PDs for each record in the data set and to then submit this for grading.

Important dates

10/17/2016: Working groups due
Week 9: Final PowerPoint deck and source code due
Week 12: Presentation of selected student models

Hints

Please see the syllabus for a sample submission and for further details regarding the project.