COMP0050 Assignment

Data

Download from moodle the file COMP0050CourseworkData.zip.

This contains two datasets:

1) **bankPortfolios.csv**: The data contain information about assets held by 7783 US commercial banks in their balance sheet in the 4th quarter of 2007. These were collected from the Wharton Research Data Services. Each row of the file is associated with a bank. The first 14 columns represent investments in the following asset classes.

Index	Asset class
1	Loans for construction and land development
2	Loans secured by farmland
3	Loans secured by 1-4 family residential properties
4	Loans secured by multi-family (> 5) residential properties
5	Loans secured by non-farm non-residential properties
6	Agricultural loans
7	Commercial and industrial loans
8	Loans to individuals
9	All other loans (excluding consumer loans)
10	Obligations (other than securities and leases) of states and political subdivision in the U.S.
11	Held-to-maturity securities
12	Available-for-sale securities, total
13	Premises and fixed assets including capitalized lease
14	Cash

Column 15 contains banks debt. Finally, column 16 contains a binary variable corresponding to the output variable y that you need to predict (1 denotes default, 0 no default). Information on defaults comes from the list

of bank failures from the Federal Deposit Insurance Corporation (FBL-FDIC) for the period 1/1/2008 - 7/1/2011.

2) **48_Industry_Portfolios_daily.csv**: this dataset comes from Ken French's website (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html) and contains daily equity returns for 48 industries in the U.S. Note that in the spreadsheet you can find two sets of data that correspond to different ways of computing industrial averages (either weighted equally or by market cap). You are free to select the version you prefer.

Tasks

There will be two tasks corresponding to the two datasets:

1. The task is to build a model to predict whether a bank will default. You should compare the performance of different methods (e.g. logistic regression, classification trees/forests) in terms of their ability to correctly predict defaults. You are free to focus on a subset of the data (e.g. a reduced set of features, or a subset of banks) and to manipulate the data as you like, but you should explain your rationale. You should address in your analysis the issue of unbalanced data.

2. Focus on a subset of the data and perform a clustering analysis on the daily equity return data. Can you find meaningful interpretations of the clusters? Do certain industries tend to cluster together?

For both tasks, justify whether you want to focus only on subsamples of the data. You are also free to explore questions related to the data and the tasks you think are interesting, as long as your analysis includes the development of predictive models of defaults for what concerns task 1 and

clustering for task 2.

Written report

A brief written report (maximum 8 pages, with a maximum 4 pages for each task, font size 11) containing the justification of the approach, the results of your analysis, and a discussion of your results should be **submitted to Moodle before the deadline of Friday 31/03/2023 at 16:00.**

Marking

This assignment is worth **100% of the overall mark (50% for each task)**. The marking will be based on the following criteria (with uniform weights):

1) Clarity of presentation and explanations
2) Validity of results
3) Critical interpretation of the results