Assignment III

Data Analysis 3

MS in Business Analytics,
2023/2024 Winter

Department of Economics and Business
CENTRAL EUROPEAN UNIVERSITY

Goal: Your task is to build the best possible model to predict defaulted firms in the 'Manufacture of computer, electronic and optical products' industry, 2015.

You may work alone or in pairs.

Hold-out sample

- To construct the hold-out sample, follow, the next specification:
- Your definition of default should be the following:
  - Existed in 2014 (sales > 0), but did not exists in 2015 (sales is 0 or missing)
- We are only interested in predicting default for 'ind2 == 26', which is the selected industry, and the firm is a small or medium enterprise (SME). Thus, yearly sales in 2014 was between 1000 EUR and 10 million EUR.
- If you do the sample design properly, you have an overall of 1037 firms. 56 firms defaulted, and 981 stayed alive. The average sales of the firms is 0.4902 million EUR, with the minimum of 0.00107 million EUR and 9.57648 million EUR.
- You should **not use this sample** for modeling, only for your final prediction's evaluation. If you use these data in any (visible) way to estimate a model, you will be penalized with -10 points. You should report your final model of choice's following measures on this hold-out sample:
  - Brier-score
  - ROC curve
  - AUC
  - Accuracy, sensitivity, specificity (for optimal threshold)
  - Expected loss and optimal threshold
    - Expected loss is has the following parameters: $loss(FN) = 15$, $loss(FP) = 3$
  - In addition, report the same descriptive statistics: number of firms, firms defaulted and stayed alive. Mean of sales, minimum and maximum values. This helps our work to evaluate and compare your results.

Getting the data

- Visit the OSF website that contains the data: `https://osf.io/b2ft9/?view_only=`
  - Under 'clean/cs_bisnode_panel.csv' you can find the cleaned but unfiltered sample between 2005-2016

Task

- Build the best prediction to classify the defaults.
- You may do different feature engineering.
- You may make any sample design decisions!
- In each case, document your steps!

- Have at least **3 different models** and compare performance
- Argue for your choice of models
  - One model must be theoretically profound logistic regression.
  - You can use any model you wish, even models that we have not covered in this course.

Documents to submit on ceulearning

- A summary report (pdf), **max 3 pages** including tables and graphs discussing your work. It is targeted at data science team leaders
  - Can use technical language
  - But need to be the point
  - Focus on key decision points, results, interpretation, decision
- Technical report – a markdown / quarto in pdf/html with more technical discussion.
  - May include code snippets
  - May include additional tables and graphs
  - Detail all decisions you made
- Reports should link to code in Git Hub.

Scoring weights

Overall, you can get 30p from this task.

- It is a prediction race. 15 points will be allocated according to model performance compared to your peers.
  - You should aim to get the **lowest expected loss** value.
  - Best gets 15 points; remaining is scaled as a distance from the closest.
- The remaining 15 points can be earned for the following:
  - Data prep, label, and feature engineering (5p)
  - Model building, prediction, and model selection (5p)
  - Discussion of steps, decisions, and results (3p)
  - Quality of the write-up (2p)

Submission deadline: 3rd of March, 2024, 23.59 CET.

Good luck!