DA3 Assignment 3

Finding fast growing firms 2025

Individual or in pairs

The assignment

Your task is to build a model to predict fast growth of firms using the bisnode-firms data we used in class.

- You should design the target (fast growth), it can be measured in any way you like over one (2013 vs 2012) or two years (2014 vs 2012).
- You need to argue for your choice, discussing a few alternatives, maybe 2-3 paragraphs using concepts and ideas from corporate finance.
- Build three different models and pick the one you like the most.
- Should include at least one logit and one random forest.

Data management, sample design

- The dataset is very close to what you saw in seminar.
- But you need to start with the panel for 2010-2015.
- Two pieces of code, sample design and prediction:
    - `ch17-firm-exit-data-prep.`
        - Starts from `cs_bisnode_panel.csv`
        - Yields `bisnode_firms_clean.csv`
    - `ch17-predicting-firm-exit.`
- On the sample, you can make changes, but do not need to.

Features

- You may use similar variables and features we used for exit prediction.
- You may do differently.
- Look at descriptives, lowess, tabulate factors, make decisions.

Tasks 1

PART I: Probability prediction

- Predict probabilities.
- Look at cross-validated performance and pick your favorite model.

PART II: Classification

- Think about the business problem, and define your loss function (like FP=X dollars, FN=Y dollars).
- For each model:
    - Predict probabilities.
    - Look for the optimal classification threshold.
    - Calculate expected loss with your loss function.
    - Pick the model that has the smallest average (over 5 folds) expected loss.

PART III: Discussion of results

- Show a confusion table (on a selected fold or holdout set).
- Discuss results, evaluate how useful your model may be.

Tasks 2

- There are two industry categories in the dataset: manufacturing and services (repair, accommodation, food).
- Define a single loss function, but carry out the exercise for two groups separately.
- Pick a prediction model, carry out classification for manufacturing and then repeat for services.
- Compare the model performance across two samples.

Submit two documents to moodle

1. A summary report (pdf), max 5 pages including tables and graphs discussing your work.
    - It is targeted to data science team leaders and senior managers.
    - Can use technical language but explain briefly.
    - But need to be the point!
    - Focus on key decision points, results, interpretation, decision.
2. Technical report – a markdown/notebook in pdf/html with more technical discussion.
    - May include code snippets (not verbose, avoid iterations, etc.).
    - May include additional tables and graphs.
    - Detail all decisions you made.
    - Reports should link to code in GitHub.

Scoring weights

Task	Weight
Task 1: Project design, introduction	10%

Task 1: Data prep, label and feature engineering	15%
Task 1: Model building and probability prediction and model selection	20%
Task 1: Classification	15%
Task 2: Technical execution and write-up	10%
Discussion of steps, decisions and results	15%
Explain shortly every modeling decision	
Final discussion of findings (2-3 paragraphs)	
Quality of the write-up, prettiness of graphs, etc.	15%

AI use

- You may use AI for any part of the exercise.
- But you are responsible for all submitted work.
- In return, we expect nice and clean code, reproducibility, pretty graphs, and well-written summary texts.