# Assignment III

## Data Analysis 3

Máté Schieszler, Péter Szilvási

2024.03.03

## Introduction

The goal of the project is to build the best possible model to predict company defaults in the ’Manufacture of computer, electronic and optical products’ industry in the year 2015.

## The data

The dataset contains detailed company data from a middle-sized country in the European Union. The database was constructed for from multiple publicly available sources by Bisnode, a business data and analytics company since then acquired by Dun & Bradstreet. The dataset contains 287829 observations on 46412 firms from 2005 to 2016 (the data from 2016 only contains third as many companies as in previous years). The id variables are company identifier (comp\_id) and year (year).

## The race

Half of the maximum 30 points for the assignment are allocated according to model performance. The goal is to achieve the lowest possible loss value on the hold-out sample. The hold-out is 1037 SMEs in 2014 in the chosen industry (‘ind2 = 26’), out of which 56 defaulted and 981 stayed alive. Further info on the firms: average sales is 0.4902 million EUR, with the minimum of 0.00107 million EUR and the maximum 9.57648 million EUR. Using this sample in any way to train the model should be avoided and is penalized by a 10-point reduction.

## Data preparation

The first and most important step of the data preparation is determining which companies defaulted, in other words, the creation of the ‘default’ variable. We define this as companies that had positive sales in a given year and had no sales or are missing from the dataset in the next year. During further preparation we keep the holdout set and its defined description metrics unchanged. The variables with high ratio of missing values are dropped. We create new numerical and categorical variables describing firm characteristics. To keep the holdout set the same, as in the assignment description, we impute important but missing variables, where it only concerns a small share of records. Flags are used for signalling imputation, and possible problems with engineered variables (too low, too high, below zero, where it shouldn’t). Certain columns, such as the growth measured in difference in sales (‘d1\_sales\_mil\_log’ ) are winsorized at the 90th percentile.

## Modelling

We create 4 models to predict defaulting companies and choose the best performing for the final prediction. We compare them on cross validated average loss on the training data.

Training data

OLS

LASSO

Random Forest

GBM