Prediction with Machine Learning for Economists: Assignment 3

Ramzi Chariag

March 10, 2023

Identifying Fast Growing Firms

Intro

In order to support investors and portfolio managers in selecting companies to invest in, by building a predictive model that classifies whether a firm is a fast growing firm or not, I compare many classification models with different thresholds of classification and loss functions.

Data and Modelling decisions

Bisnode-firms dataset – a panel data of firms located in some city of Europe over a period of 2010-2014. The data has 140,100 observations and 47 columns.

For this exercise, I calculate the yearly change in sales, then restrict the data to a 2012 cross section (change in sales is between 2011 and 2012). The data also includes firms that existed in 2012. There are 3751 fast growing firms which is 17.27% of all firms.

Variables, Models and Results

Variables

The original data had financial statement information about firms, as well as some ownership and HR information about management. I generated the target and some extra features from the data. The target, $fast\_growth$ is 1 for firms who has at least 20% growth in sales for each year in a period 2010-2014. Other alternative targets could be employment growth, assets growth, or some financial ratios like ROI or ROE. However, the dataset at hand either lacks these or they have some errors (such as firms with negative assets). Other variables that could predict a firm's growth such as industry category, age of a firm (plus age squared), gender of a CEO dummy, region category, log of sales (plus log squared), return on assets (ROA), return on equity (ROE), return on total assets (ROTA), current ration (CR), and historical growth of sales (diff_ln_sales).
Observations missing key variables were dropped, and when only very few observations were missing, I used mean imputation.

Models

In total, there are 7 models: five logit models of different complexity, one lasso logit, and one random forest.

- **Model 1**
  Logit with variables log of sales, log of sales squared, difference in log of sales (winsorized), profit/loss, and industry dummies.

- **Model 2**
  Logit with Model 1 variables + fixed assets, share equity, current liabilities, age of a firm, foreign management dummy, and flags.
- **Model 3**
  Logit with Model 2 variables + all other financial variables including financial ratios, age squared, new firm dummy, region dummies, and urban dummies.
- **Model 4**
  Logit with Model 3 variables + quadratic terms of profits, income, and equity, firm characteristics, and all flags.
- **Model 5**
  Logit with Model 4 variables + interaction variables.
- **Model 6**
  LASSOLogit with Model 5 variables
- **Model 7**
  Random forest, with all raw variables as dependent variables.

Results

Part I: Probability Prediction
From the following table, I chose Model 4 to use as a benchmark.

Table 1: *Summary of results, no loss function*

Model	*#Predictors*	*CVRMSE*	*CVAUC*
$X1$	11	0.369	0.598
$X2$	18	0.365	0.646
$X3$	39	0.362	0.664
$X4$	83	0.359	0.670
$X5$	164	0.360	0.669
LASSO	131	0.360	0.670

Part II: Classification
As can be seen from the table below, the lowest cross-validated expected loss is achieved by the random forest model

Table 2: *Summary of results, with loss function*

Model	*#Predictors*	*CVRMSE*	*CVAUC*	*CV threshold*	*CV expected loss*
*X*1	11	0.369	0.598	1.130	155.900
*X*2	18	0.365	0.646	1.188	155.888
*X*3	39	0.362	0.664	1.275	155.949
*X*4	83	0.359	0.670	0.637	154.394
*X*5	164	0.360	0.669	0.890	154.801
LASSO	131	0.360	0.670	0.921	155.096
RF	45	0.358	0.674	0.631	152.99

Table 3: *Confusion matrix for Random Forest model*

	Predicted not fast-growing	Predicted fast-growing
Actual not fast-growing	3179	5
Actual fast-growing	605	20

Table 4: *Confusion matrix for Model 4 with 0.3 threshold*

	Predicted not fast-growing	Predicted fast-growing
Actual not fast-growing	3002	182
Actual fast-growing	489	136

Now return to our investor example. The model predicted that there are 25 fast-growing firms. Suppose I invest 1000 EUR to each company. I will spend 25,000 EUR and get in return $5 * 937 + 20 * 1781 = 40,305$ EUR. That is a whooping 61.22% return.

Now suppose that I did not have a loss function. The benchmark model was Model 4, and the threshold = 0.3. Suppose I invest 1000 EUR to each company. I will spend 318,000 EUR and get in return $182 * 937 + 136 * 1781 = 412,750$ EUR. That is 29.8% return, which is much less than in previous case, by a great margin.

This is how relevant the model can be, especially if financing is not abundant, and investors are quite risk averse. What we are doing is that we are essentially increaing the rate of return on investment by being more selective.

Having more domain knowledge of the industries in which these firms operate and of accounting would allow to specify more fitting loss functions. One this that could be done is splitting firm by industry and grouping similar ones together. When I say similar I mean in accounting terms. e.g. firms in retail, food, or any industry that requires a lot of inventory turnover require a high level of working capital as a percentage of total assets, whilst that is the opposite for firms in construction for example. The point of all of this is that the same number on the balance sheet could mean a very different thing depending on industry. While it is true that we could control for industry and that would be the end of that, but it would make more sense for investors to have a model for each industry group as it fits well with the way

that portfolio managers think and it would be a lot easier for them to use and to convince others of their choices.

github repo: **https://github.com/RamziChariag/Prediction-ML**