HGC prediction - Technical documentation

Péter ENDES-NAGY

Introduction

This is a technical documentation of the *Report on HGC prediction.*

Data

Data is from the `bisnode-firms` dataset, collected, maintained, and cleaned by Bisnode, a major European business information company. It covers the entire population of companies between 2005 and 2016 for a few industries in manufacturing (Electronic and optical products, Electrical equipment, Machinery and equipment, Motor vehicles, Other transport equipment, Repair and installation of machinery) and services (Accommodation and Food and beverage service activities) for a given European country.
The report is focusing on a cross-section of companies in 2012, other parts of the dataset is used for calculating high-growth for subsequent years as target variable and some historical growth variables as predictors.
The dataset is available (here).

Data preparation and sample design

Code used for this part: available here)

Data preparation

- Data read directly from https://osf.io/3qyut/
- Variables with too many missing values we dropped and dataset filtered for our panel: 2010-2015
- missing year and company ID combinations were added for calculations (so `lag()` and `lead()` functions work properly)
- `status_alive` variable created for later filtering: sales larger than 0 or missing in a given year
- There are almost 70k observations with missing sales data, filtered out later if not a new company. Log sales and sales in million EUR (both log, both level) were calculated
- a historical 1-year growth variable was calculated (how much the company grew compared to previous year).
- if a company is freshly established, a dummy variable captures it and the historical growth variable was imputed as 0 growth for that given year as well.

Target variable

Our target variable is a dummy, whether a company experienced high-growth in the sebsequent 2 years: average annual growth above 20%.

- Average yearly growth was calculated from log values
- Yearly growth transformed into level
- Dummy variable created with 1.2 cut: above 20% growth, takes value of 1, otherwise 0.

Sample design

We are focusing on a cross-section of companies for the year of 2012.

- Year was filtered for 2012
- Companies not alive (0 sales) and with missing target variable are dropped

According to the HGC literature, it is recommended to drop observations with a base value too low: too easy to grow from a low base, most companies experience a high-growth at least at the beginning of their life-cycle, so classifying them as HGC won't make much sense.
As a good practice, companies under 10 employees (basically the micro segment) shall be filtered out. I had to keep in mind that I need a reasonable sample size, so I was less strict with the filtering on the bottom end. 100k EUR (yearly 10k EUR income per employees, so still very low, not even enough to cover salaries) would have resulted in a sample of 6.000 observations only. 50k EUR yields a sample of 9.100, 10k EUR 15.600 sample.
N = 9.100 is a fair sample size, so companies under 50k EUR sales were dropped. Larger companies above 10m EUR were also dropped as they might behave very differently and they usually aren't prone to grow 20% annually.
Another option would have been filtering by number of employees (from 10 to 249 to keep the SME segment), but that metric is very unreliable in most company datasets. It would have also resulted in a sample size of 277...

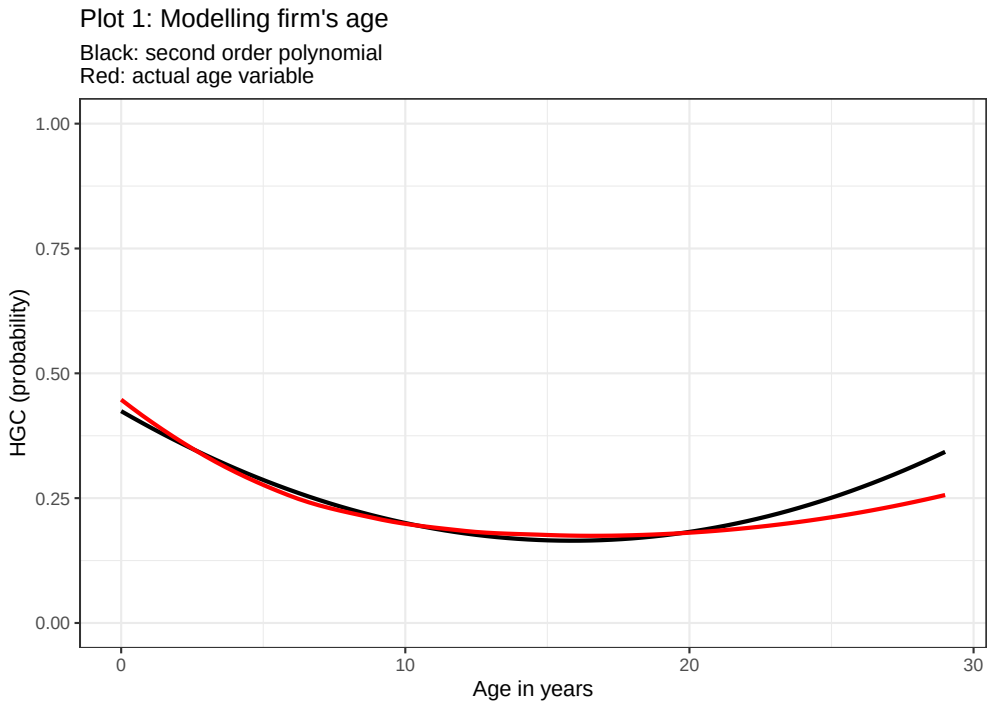
After the sample design, I ended up with a sample size of 9116.

Feature engineering

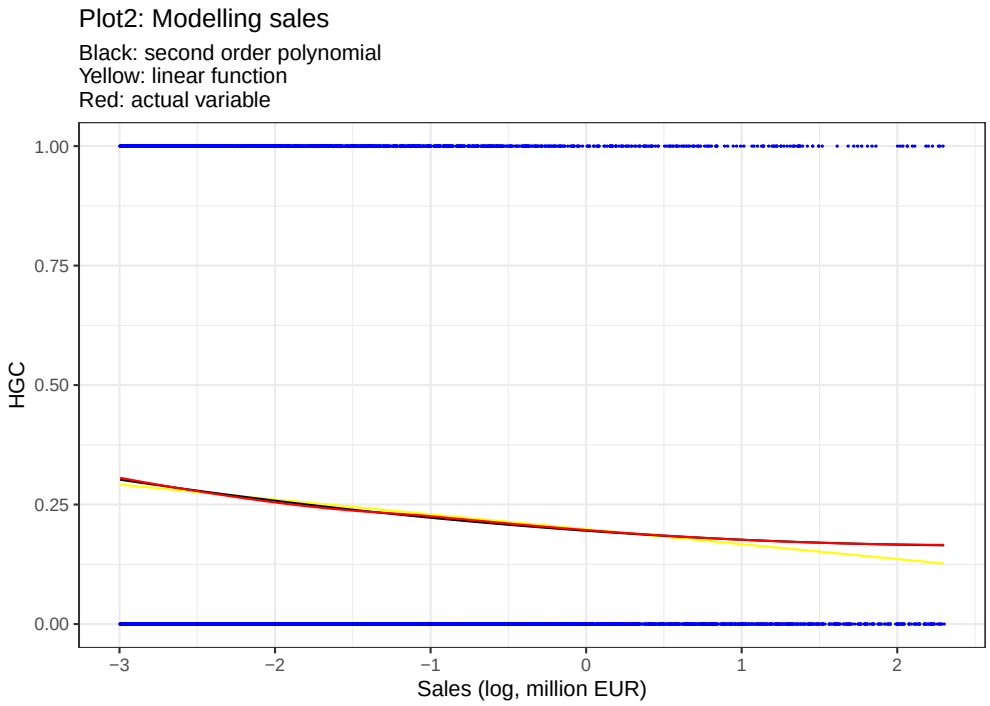
Code used for this part: available here)

Many decisions are based on inspiration from the firm exist case study, as they stem from field knowledge that I do lack in this particular field.

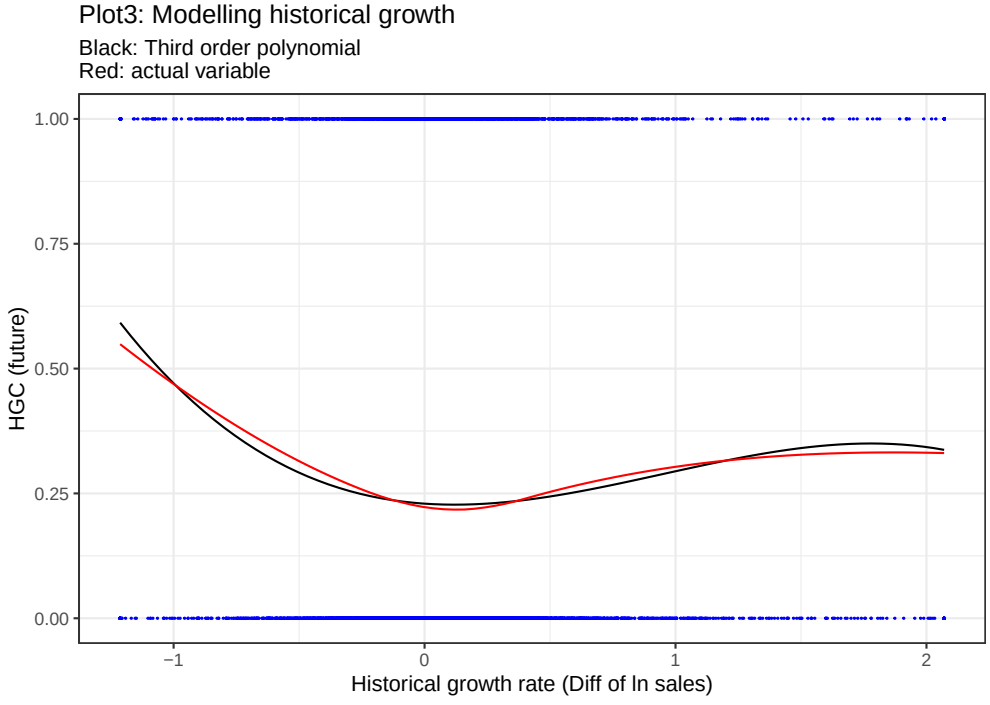
- Industry category codes were regrouped and changed
- Firm characteristics were derived and pooled from existing variables:
  - square of firm age to model non-linearity, a quadratic function captures it really well. See plot1.



- Foreign management dummy if majority is foreign
- Main location and gender of management converted into factors
- As assets can't be negative, the relevant variables were changed to zero and flag variable created if problem with any of the asset classes.
- Financials and ratios:
  - one way would be taking the log values (similarly to sales), another viable option is creating ratios. Deciding on ratios have also additional corporate finance meaning.
  - Loss and Profit account items were divided by sales, the balance sheet items by total assets (total asset was calculated from sub-items).
- For items that can't be negative, flags were created and 0 values imputed.
- Some ratios can be anything, but usually in the [-1,1] range, were winsorized and flags created. Squares were calculated for capturing some sort of non-linearity, although for many of them, higher order polynomials would capture it better, but I'll keep it simple now.
- CEO's age was calculated, tails winsorized under 25 and above 75, flags created. Missing values imputed with mean, dummy created for young CEO's (under 40 years old)
- The number of employees is very noisy (a very unreliable data in most company datasets). Missing values were imputed with mean value and flags created.
- For sales data, polynomials aren't necessary as a simple lm does a fair job, see plot2.



- For the historical growth, winsorizing is necessary, the top and bottom 1% was decided for cuts, flags also created. By looking at the winsorised variable, a third order polynomial captures the non-linear relationship very well, see plot3. It also makes sense, if a company shrunk (a lot) in the past, it is easier to grow a lot, jumping back to the previous level and even surpass it. Companies that didn't grow in the past, might be the boring middle-range, those who successfully grew can grow again, but those who grew a lot, are less likely to be HGC again in the subsequent years.



- Remaning missing values:
  - For key variables, the low number of missing values were dropped: liquid assets ratio, foreign management, industry, material cost ratio, main region.
  - For unnecessary variable with too many missing, the variables were dropped: D(?) , birth and exit year, exit date.
- Finally, unused factor levels and flags without variation were dropped.

After the feature engineering and dropping a few observations, the final dataset contains 8203 observations.

Model building

Code used for this part: available here)

4 models were built in total: 2 logit models, a logit model with LASSO and a random-forest model.

Predictors

The varibales were grouped into the following categories:
- `rawvars`: main firm variables without tranformations (log, winsorizing, etc.) - `qualityvars` : data quality, as of problem with data and length or years that the balance sheet covers. - `engvar`: financial ratios - `engvar2`: quadratic forms of some ratios incl. in `engvar` - `engvar3`: flag variables related to financial ratios - `d1`: historical growth variable, quadratic and cubic forms, flags (winsorisation) - `hr`: human capital related variables - `firm`: firm's history related variables as of age (quadratic incl.), new, region, industry, location - `interactions1`: industry group interactions with a some variables - `interactions1`: sales (log) interactions with a some variables

Models: - X1: sales (log), financial ratios, firm, historical growth - X2: sales (log), financial ratios + their quadratic forms + flags , firm, human capital, historical growth, data quality - LASSO: X2 and interactions - RF: untransformed financials, historical growth (not winsorised), hr, firm, human capital and data quality

The number of variables, `nvars` was saved for each model. Quadratic forms, interactions, flags weren't included, in the summary tables there are different columns for the number of variables and number of coefficients.

Prediction

- Sample divided into working and hold out sets with 1:3 ratio.

- 5-fold cross-validation was defined, the summary function uses RMSE criteria
- For the random-forest model, 'gini index' was chosen to decide split rule. Multiple tuning parameters were tried, [5:8] for `mtry` and [10,15,20,25,30] for minimum node size. 8 and 25 was choosen by the machine as best tuning parameters.
- Models and their results were saved into a list
- RMSE and AUC was calculated for each folds (5 in total), then averaged and saved into lists.

	N.vars	N.coeffs	CV.RMSE	CV.AUC
X1	22	35	0.4209	0.6501
X2	29	75	0.4201	0.6514
LASSO	29	25	0.4201	0.5117
RF	29	43	0.4185	0.6540

Classification

For the loss function, we can imagine a bank that wants to offer premium services to potential HGC's. They expect extra revenue from the HGC's as they are going to use more services of the bank in the future as they grow, let's say 5.000 EUR per real HGC. The premium service comes with higher cost for the bank, as it includes free consulting services for the potential HGC's, let's say 1.000 EUR per company.

Therefore the false negative case (classified as non HGC but becomes a HGC) is the lost extra revenue minus consultation cost: 4.000 EUR. The false positive case (calssified as HGC but not HGC) comes with a loss of the consultation cost, 1.000 EUR.

For each model, expected loss and best threshold was calculated for each fold, then averaged and saved into a list. For the best threshold, the Youden index was maximized.

	N.vars	N.coeffs	CV.RMSE	CV.AUC	CV.tresholds	CV.expected.loss
X1	22	35	0.4209	0.6501	0.1875	0.6475
X2	29	75	0.4201	0.6514	0.1721	0.6483
LASSO	29	25	0.4201	0.5117	0.2465	0.7482
RF	29	43	0.4185	0.6540	0.1999	0.6384

Hold-out set

The above mentioned 2 steps (prediction and classification) were calculated on the hold-out set as well for each models.

	N.vars	N.coeffs	HO.RMSE	HO.AUC	HO.tresholds	HO.expected.loss
X1	22	35	0.4169	0.6571	0.1711	0.6532
X2	29	75	0.4155	0.6563	0.1723	0.6488
LASSO	29	25	0.4182	0.6562	0.1938	0.6576
RF	29	43	0.4144	0.6704	0.1715	0.6468

Furthermore, the confusion table was calculated for each model on the hold-out set (best CV threshold used in the calculation), displayed for the best performing RF model.