Finding fast growing firms

Maeva Braeckevelt and Brúnó Helmeczy

Executive Summary

This analysis aimed at predicting the fast growing of a firm.

Introduction

This analysis serves to predict fast growth of firms between 2012 and 2014. This analysis could help an investor to decide to invest in a certain company or not. The data was originally created by Bisnode, a major European business information company. The original dataset, bisnode-firms, considers companies between 2005 & 2016 in both Manufacturing (Electrical equipment, Motor vehicles, etc) & Services (Accommodation, and Food & Beverage services activities).

Data

Label Engineering

Firstly, we chose to define fast growth in terms of sales, more specifically, the compound annual growth rate of sales between 2012 & 2014 (Please see the formula specification below). We defined any firm with a % CAGR of at least 25% in the coming 2 years as a 'Fast Growing' firm, implying their sales would grow ca. 60% two years into the future. One reason for this, is the time-consuming administrative process of actually investing into firms.

$$CAGR = (\frac{Vfinal}{Vbegin})^{\frac{1}{t}} - 1$$

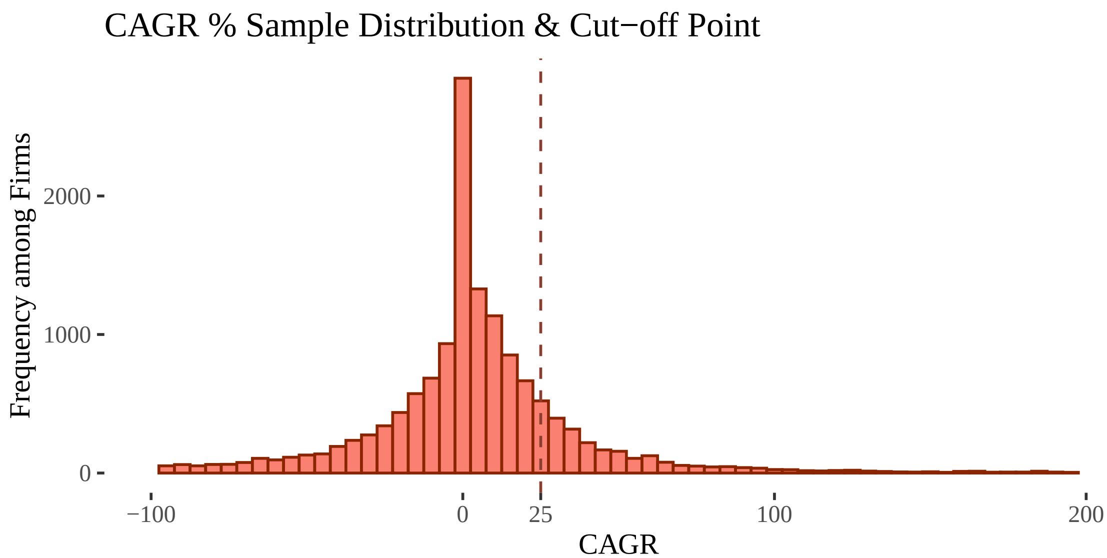
- **t** being the time in years, we defined it as 2.
- **Vbegin** is the Sales (in euro) in the year ongoing
- **Vfinal** is the Sales (in euro) the year +2

Sample Design

As business context, we consider our client to be a silent investor group as of 2012, who is looking for suitable companies to invest in, with expected returns on their investments in 2 years. As such, we considered companies at least 1 year old, since such companies have already accumulated financial data over multiple years, yet many are young, with higher chances of gaining further market share. We only considered businesses generating annual sales between 100 thousand, & 10 million Euros, boasting sufficiently large sales for the business venture to be considered 'serious', yet small enough for it's size not to be considered a disadvantage. After having design our sample, we end up with 14896 companies and 2413 of them (16%) are fast growing.

```
CAGR_plot <- ggplot(data,aes(x = CAGR)) +
  geom_histogram(binwidth = 5, color = "orangered4", fill = "salmon") +
  geom_vline( xintercept = 25, color = "coral4", linetype = "dashed") +
  scale_x_continuous(limits = c(-100,200), breaks = c(seq(-100,200,100),25)) +
  theme_tufte() + labs(title = "CAGR % Sample Distribution & Cut-off Point ",
                       y = "Frequency among Firms", y = "CAGR % ")

grid.arrange(CAGR_plot,ncol = 1)
```



Feature Engineering

Now, we need to select, clean & possibly transform our x variables. Some variables are financial accounts (e.g. inventories, current liabilities etc.) so they cannot be negative, thus we replaced the negative values to 0 & included flagging dummy variables to indicate our inputation. We created ratios, easier for interpretation & spotting extreme values. We also created new variables, e.g. change between the variable now & a year ago, we did it for the log of Sales (in million), inventories, total assets & amortization. Finally, we used a method called 'winsorization' on the change in sales & in total assets, to avoid extreme values. Afterwards, we grouped our variables, for easier interpretation, & better overview. Please see them below:

- **Firm**, *5 variables* : Age of firm, squared age, a dummy if newly established, industry categories, location regions for its headquarters, and dummy if located in a big city
- **Financial 1**, *16 variables* : Winsorized financial variables : sales, fixed, liquid, current, intangible assets, curret liabilities, inventories, equity shares, subscribed capital, sales revenues, income before tax, extra income, material personal and extra expenditure, extra profit, EBDTA, amortization and tangible assets.
- **Financial 2** : Flags (extrem, low, high, zero - when applicable) and polynomials : quadratic term are created for profit and loss, extra profit and loss, income before tax, and share equity.
- **Financial 3** : total assets, fixed assets divided by total assets, liquid assets divided by total assets,current assets divided by total assets, share equity divided by total assets, subcribed capital divided by total assets, intangible assets divided by total assets, extra expense divided by total sales,

extra income divided by total sales, extra profit and loss divided by total sales, income before tax divided by total sales, inventories divided by total sales, material personal and extra expenditure divided by total sales, profit and loss divided by total sales, personal expense divided by total sales, working capital divided by total sales, EDITDA divided by total sales.

- **Growth**, *X variable* : Sales growth is captured by a winsorized growth variable, its quadratic term and flags for extreme low and high values.
- **HR**, *5 variable* : 5 variables : For the CEO: Female dummy, winsorized age and flags, flags for missing information; foreign management dummy; labor cost, and flag for missing labor cost information.
- **Data variables**, *3 variables*: Variables related to the data quality of the financial information, flag for a problem, and the length of the year that the balance sheet covers.

We chose to include some interactions as well, that we defined by common knwoledge.

- **Interaction** : Interactions with the sales growth, firm size, and industry

Probability prediction and model selection

We decided to use three different models for the prediction of fast growing firm : Logit, Lasso and Random Forest. Our 5 Logistic regression models gradually incorporate more & more of the variable described groups above, where our final model, similarly to LASSO, incorporates all available variables. Finally, our Random Forest model considers all variable groups except interactions. We used a 5-fold cross-validation to estimate models and then we selected the best model based on expected loss. To calculate the expected loss we defined the cost of false negative and false positive errors.

Define loss function

We first calculated the average CAGR amongst Hyper growth, & non-hyper growth companies. We found that companies labeled as hyper growth, on average have a CAGR of 64%, meaning they triple their size in terms of sales in 2 years. We assumed this implies our investors triple their money as well, giving us the base, true-positive scenario. Our False Negative case is when we do not decide to invest, but should have, implying the investors missed out on tripling the capital they would have invested, i.e. we set our False negative = -2.

The False Positive case is when we invest in a firm, but should not have. To estimate our costs in this case, we calculated average CAGR among non-hyper growth firms whom did not go out of business, & found that on average they stagnated (CAGR = 0.45%). However, ca. 18% of non-hyper growth firms defaulted in the 2 year horizon, implying investors losing their capital. Furthermore, by investing in the wrong firm, investors obviously missed out on the expectation of tripling their money by investinng in the right firms. On the other hand, we believe mistakes breed learning & growth. Thus, we set our false positive error as 20% * -1 (Lossing capital due to default) -2 (not tripling investment) +1 (learning from the experience) = -1.2. After rounding the whole numbers for simplicity, our False Positive Error = 3 & False Negative Error = 5

Modeling

The predictors of the first two models were handpicked and then we gradually add more categories at each models.

- **X1** : Log of sales in Millions, the square of the Log of sales in Millions , Winsorized value of the change of the log of Sales versus last year, profit and loss by total sales, Industries categories

- **X2** : X1, fixed assets divided by total assets, share equity divided by total assets,current liability divided by total assets, hight flags for current liability divided by total assets, flag error for current liability divided by total assets,age, foreign_management
- **X3** : Log of sales in Millions, the square of the Log of sales in Millions, Firm, Financial 1, Growth
- **X4** : Log of sales in Millions, the square of the Log of sales in Millions, Firm, Financial 1, Growth, *Financial 2*, *HR*, *Data quality*
- **X5** : Log of sales in Millions, the square of the Log of sales in Millions, Firm, Financial 1, Growth , Financial 2, HR, Data quality, *Interactions*
- **Lasso** : Log of sales in Millions, the square of the Log of sales in Millions, Firm, Financial 1, Growth , Financial 2, HR, Data quality, Interactions
- **Random Forest** : Log of sales in Millions, the square of the Log of sales in Millions, Firm, Financial 1, Growth , Financial 2, HR, Data quality

Model	Number.of.predictors	CV.RMSE	CV.AUC	CV.threshold	CV.expected.Loss
X1	11	0.366	0.58	0.306	0.805
X2	18	0.364	0.608	0.354	0.802
X3	61	0.361	0.646	0.469	0.795
X4	114	0.362	0.648	0.47	0.793
X5	245	0.366	0.644	0.549	0.798
LASSO	26	0.361	0.625	0.488	0.8
RandForest	114	0.361	0.651	0.45	0.795

Please see all models summary statistics above. Observed the lowest Root Mean Square Error is 36.1 %, given by models X3, Lasso & Random forest. However all models were very close, the highest RMSE being 36.6%. The highest 'Area Under the Curve' is given by the Random Forest model, with AUC = 65,1%. I.e. Using this model, we are 15% better off than by simple random guessing, & 7% better off vs the base model, X1. However, like for the RMSE, the model X4 is very close to the Random forest with a AUC of 64,8%.

EXPECTED LOSS We base our final model selection on minimized expected loss, for which we first selected optimal classification threshold values of each model, thereby comparing the best possible expected loss for each model. The smaller expected loss is 0.793 is the model X4. Considering that it is a very theoretical concept, we only interpret it in relative terms: it is the lowest of all models, though our models again show little variation. Thus, our final model choice is Model X4. It had the lowest expected loss (the most important criteria), while the AUC & RMSE were very close to being the best.

Ex-Sample Testing & Model Diagnostics

prediction
no_Hyp.Growth
Hyp.Growth

on	no_Hyp.GrowthFinal_Model	Hyp.GrowthFinal_Model	no_Hyp.GrowthBenchmark_Model	Hyp.GrowthBenchmark_
o.Growth	2471	466	2480	
owth	21	21	12	

HyperGrowth_f	AvgAnnGrowth	Firms_Selected	Investment_per_Firm	NetEarnings
no_Hyp.Growth	-29.3	21	1e+04	-1.05e+05
Hyp.Growth	85.9	21	1e+04	5.16e+05

[1] 0.977865

```
OutcomePerInd
```

HyperGrowth_f	ind2	AvgAnnGrowth	Nr_Investments
no_Hyp.Growth	26	-26.1	3
no_Hyp.Growth	27	-46.3	3
no_Hyp.Growth	28	-56.1	2
no_Hyp.Growth	29	-25.1	3
no_Hyp.Growth	33	-100	1
no_Hyp.Growth	55	-21.6	1
no_Hyp.Growth	56	-11.2	8
Hyp.Growth	26	62.1	1
Hyp.Growth	27	31.5	1
Hyp.Growth	28	34.3	2
Hyp.Growth	29	155	4
Hyp.Growth	33	119	1
Hyp.Growth	55	100	2
Hyp.Growth	56	70.1	10

Conclusion / Summary