Assignment 3 - Finding fast growing firms

Course: ECBS6067 - Prediction with Machine Learning for Economists

Anja Hahn, Teresa Huebel & Anne Valder

2023-12-14

Summary

The goal of our analysis is to classify fast growing firms in order to identify lucrative investment opportunities. This report includes three different models to predict fast growing firms. It is based on bisnode panel data. Firm growth is defined by average annual percentage growth in sales over two years. We use data from 2012 to predict sales growth two years ahead. The models are logit, random forest and logit lasso. The models' predictive power is evaluated once without and once with the use of a loss function. After the introduction of a loss function, we identified random forest as strongest model. This model is also used on the subsets manufacturing and services. In comparison, the random forest performs better in the services sector. Further results and interactive graphs can be found in the shiny app linked to this report.

Sample Design

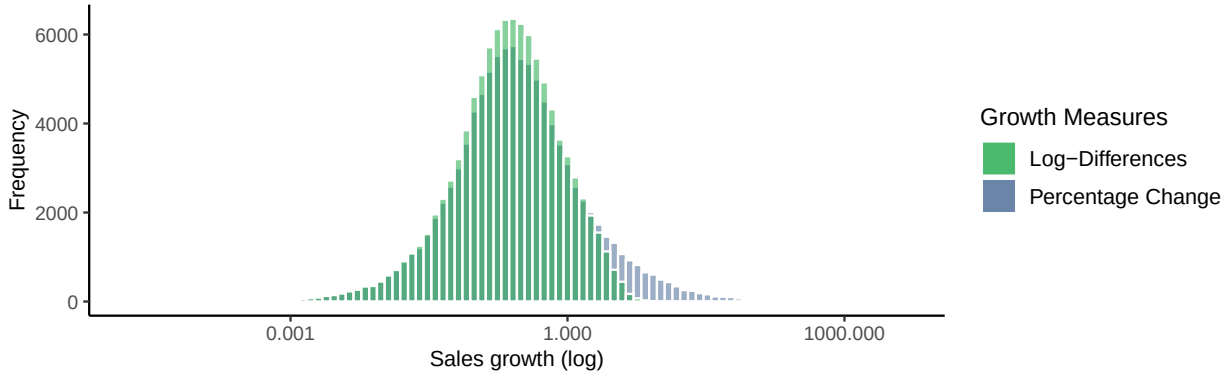
We exclude some variables that have extremely high rates of missing values (above 90%) and only look at firms that are active in 2012, that is, they have some sales. What is more, our analysis is only based on firms with annual sales between 1,000 and 10 million euros. We believe this subset presents the most attractive investment opportunities. First, very small firms with sales below 1,000 EUR in sales are likely not to seek investment and high growth events might happen at random more easily. Second, very large firms are well represented in various indices and any expected growth is likely to be already incorporated in the stock price. Hence, our sample selection represents a business decision to focus on a segment that is more predictable and more promising for investors.

Label Engineering

We are exploring multiple alternatives as target variable to represent high growth firms. In our selection, we pay attention to the following dimensions: i) the indicator to choose, ii) the measurement, iii) the time horizon, and iv) the threshold. The literature does not yield one consensus over a definition of high growth firms, it rather points out that any decision along these dimensions depends on the goal of the analysis.[1] We decided for relative average annual sales growth over two years for the following reasons: First, sales provided a more relevant and reliable target than number of employees. Sales data is missing only in 2.6% of cases whereas number of employees is missing in 50.9% of cases. Also, sales growth is more closely linked to higher profits. What is more, 85% of firms had less than 1 employee on average. Second, we decided to use relative growth instead of absolute growth. This decision will favor smaller firms since high relative growth is harder to achieve for larger firms. We believe relative growth to be more relevant since it is more closely linked to our expected return on investment. To be more specific, we calculated percentage change and did not rely on log differences. Even though this decision does not affect the ranking of firms, it does affect the

[1] Alex Coad, Sven-Olov Daunfeldt, Werner Hölzl, Dan Johansson, Paul Nightingale, High-growth firms: introduction to the special section, Industrial and Corporate Change, Volume 23, Issue 1, February 2014, Pages 91–112, https://doi.org/10.1093/icc/dtt052

magnitude of the growth rate in the subset of high growth firms. The difference is visualized in a plot that shows the distribution of growth rates for both measures, where log-differences are right-skewed.



We favored percentage change primarily because of its more intuitive interpretation. Third, we decided for a two year time horizon instead of a one year horizon to relate more closely to the OECD definition of high growth firms that is defined over a three year horizon. Fourth, we decided for a threshold of 30% annual growth. In our sample this is equivalent to classifying roughly the fastest growing 20% of firms as HGF. In the literature, both thresholds on a certain growth rate and a certain percentile of firms are used. We decided for a threshold based on a specific growth rate since we want to analyze different sectors (manufacturing and services) with a fixed benchmark.

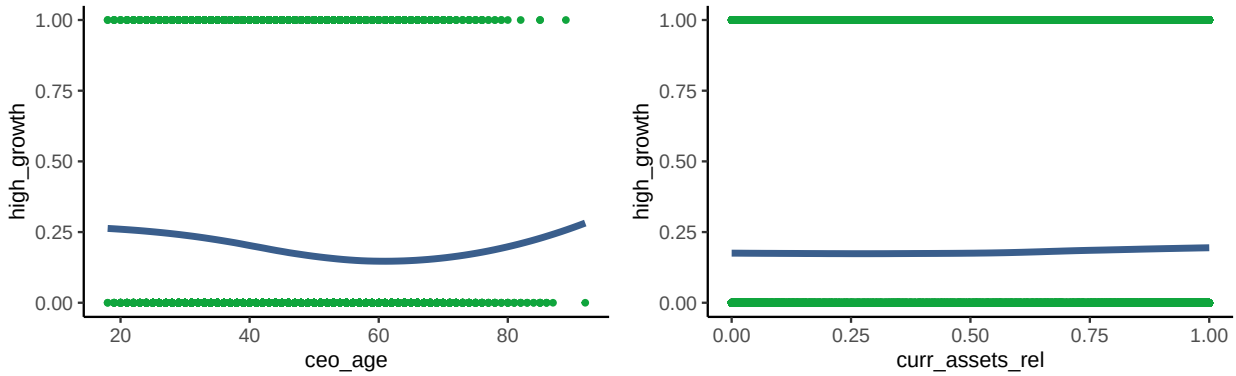
Feature Engineering

The feature engineering process includes plausibility checks: As some of our financial variables (e.g. *inventories*, *fixed_assets* etc.) must not be negative, we flag them and set negative values to zero. Other implausible variables, such as *ceo_age* below 18 years, we set to NA.

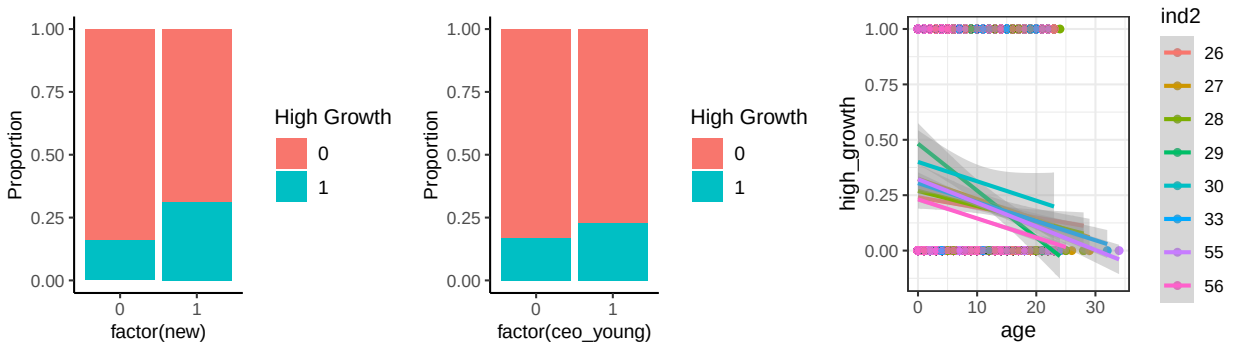
In the process of analysing missing values, we impute some important ones (*ceo_age*, *labor_avg*) and add flags for missing values. For other, very important variables (mainly assets, CEO and industry information), we do not impute but rather drop NAs. Moreover, we intend to use the variable on previous firm growth as predictor for (future) fast growth – very much in the same fashion as we know from time trend models. Thus, we include the lagged growth in sales (2011 to 2012). As this results in some NAs (due to some missing sales values in 2011), we again flag the observations concerned and impute them by replacing them with 0. This means we assume that said observations did not experience any growth. Although this solution is not ideal, it is based on the assumption that we compare a firm's (missing) sales value in 2011 to the value of its "nearest neighbour" – the firm's sales in 2012.

Moreover, we compute some new variables: *age* of the company, a dummy for whether it is a *new* company, has *multiple_ceos* or a *foreign_management*. We classify the observations based on their two-digit NACE codes into "manufacturing" vs. "services", generate a variable for *total_assets* and compute ratios of some financial variables (with either *total_assets* or *sales* as reference). Lastly, we windsorize tails of numeric variables and add flags accordingly.

Next, we look at relationships between predictors and the binary response. We use the LOESS (locally estimated scatterplot smoothing) method to fit a smooth curve to the data points. In case the curve shows some clear non-linear behaviour, we add quadratic terms of predictors concerned. These decisions are reassured by the estimation of simple GLMs, where we explain *high_growth* with the respective numeric variable and its quadratic – whenever we detect a statistically significant relationship between a quadratic term and the binary response, we add that variable as quadratic predictor as well. From visual inspection of the plots, we can say that in general, asset variables do not show a lot of variation in the response i.e. do not seem to affect *fast_growth* as much (see examples below).



Moreover, we looked at the relationship between the binary response and binary/categorical predictors. In the plot below, we see, for example, that *high_growth* seems to be more prevalent with a young (defined as < 40 years) CEO. Also, a firm being very young (< 1 age) seems to play a role. As regards possible interactions, the below right plot hints towards a relationship between industry category (NACE 2 digit classification) and age of the company, as can be seen by the different slopes of the fitted line per category.



Variable Selection

Logit Model For the logit, we try out two specifications. Both of them include the sales variable (past growth), engineered variables (i.e. ratios of financial variables, windsorized), the detected (possible) quadratic terms and the flags we created for extreme, missing and potentially erroneous data. Also, both specifications include quality variables (balance sheet information) and information on HR (e.g. CEO information) and the firm itself (e.g. regarding the location). The difference lies in interactions included: Whilst the first specification does not include any interactions, the second one interacts selected variables with the industry variable.

Probability Forest Since the probability forest can deal with complicated relationships between variables, we do not include any quadratics, interacted, modified and/or windsorized variables. Thus, the raw financial, quality, sales, HR and firm variables are added. Also, potentially highly correlated variables are considered together. Categorical variables are encoded as factors.

LASSO Logit For the LASSO logit, we consider the same variables as in the second logit specification (including one additional highly correlated variable).

Model Estimation

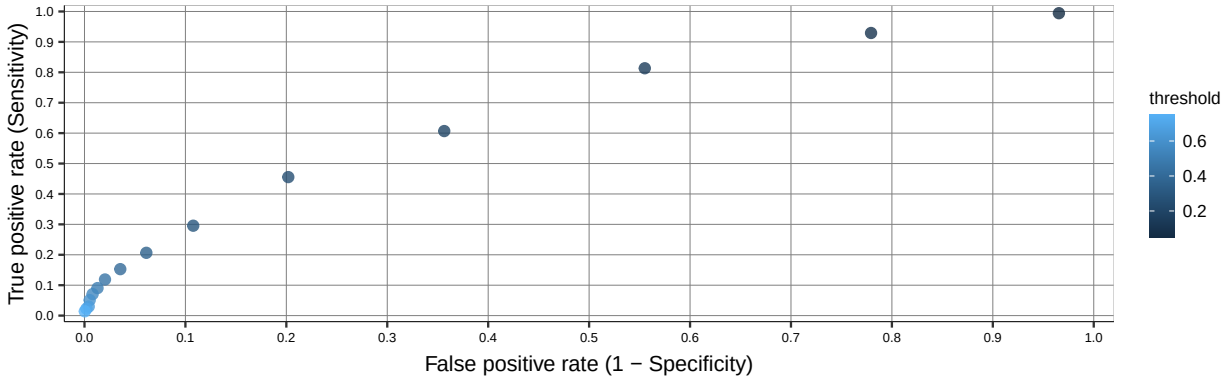
We divide our data set into a work (80 percent, 15,229 observations) and holdout set (20 percent, 3,807 observations). The workout set will be used to train our models. We use 5-fold cross validation to evaluate performance of our models.

Probability Prediction The below table shows the results for our different model specifications. We would choose the first simple logit model over the second, since it has fewer coefficients and even better (lower) RMSE and (higher) AUC. The logit LASSO does not perform better, either (worse AUC and similar RMSE). Based on the CV results, we would therefore go for the random forest which performs slightly better in both the RMSE and AUC.

Table 1: Summary of Prediction Models (no loss function)

	Model	Preds	Coeffs	RMSE	AUC
1	Logit w/o interactions	67	91	0.375	0.672
2	Logit w interactions	67	175	0.376	0.670
3	Logit LASSO	68	176	0.375	0.642
4	Random Forest	36	NA	0.372	0.682

We use our best model to predict on the holdout set and receive a RMSE of **0.372**. Also, we produce the below (discrete) ROC plot which shows the trade off between making false positive and false negative errors.



The confusion matrix for default classification (using 0.5 as threshold) is shown below.

When comparing the above confusion matrix (default classification of 0.5 as threshold) to the one below – where we set the threshold to the mean of predicted values (approx. 0.2) – we see, as expected, how a lower threshold would lead to a higher probability of predicting high growth. Also, we nicely see the trade-off between false positives (going up) and false negatives (going down) when lowering the threshold. However, these choices of thresholds are still not based on a proper loss function, to which we will come next.

Introduction of Loss Function We introduce a loss function specifically tailored for our business case. In selecting the best model, our primary objective is to minimize false positives. This is crucial because we aim to avoid investing in firms that do not exhibit strong growth potential. However, there's a challenge: only 20% of firms in our dataset are categorized as fast-growing. Overemphasizing the cost of false positives could inadvertently lead the model to classify all observations as low growth. Such an approach would undermine our goal of identifying high-potential business opportunities.

To address this, we have calibrated the cost associated with false positives (FP) and false negatives (FN) in a balanced manner. We have set the cost of a false positive at 150% of that of a false negative. This ratio reflects a strategic decision: while avoiding unfruitful investments is important, we also do not want to miss out on promising firms. Additionally, we have incorporated the prevalence of high-growth firms into the model's weighting system. This adjustment ensures that our model remains sensitive to the relatively scarce yet valuable high-growth opportunities in our dataset, thereby aligning the model's predictions more closely with our business objectives.

Model Comparison with Loss Function After the introduction of our loss function, we can now compare the models based on their average expected loss over all folds. The model with the lowest expected loss is still the random forest. The simple logistic regression without interactions has the second lowest expected loss.

Confusion Table The confusion table shows that we have an accuracy of 82.03% on the holdout set. The model only predicts 1.65% of models to have high growth whereas the prevalence of high growth firms in the data is 18.57%. This low sensitivity (6%) is necessary in order to arrive at the high specificity (99.3%) we desire. This is a trade off we are willing to engage in, in order to avoid investing in firms that will not outperform the market or might even go bankrupt. If we apply our model to similar datasets in multiple countries we will still arrive at a reasonably large number of predicted high growth firms such that our portfolio will be diversified.

Of course, we realize that we could not test the external validity of our model and applying it in other countries might lead to worse performance. To that end, we conduct another exercise and apply our model to different market segments, that is, manufacturing and services.

Comparison of models in manufacturing and services The random forest performs better in the services sector. It exhibits a lower RMSE, a higher AUC and a lower expected loss.

Table 2: Comparing Manufacturing and Services Models

	Model	CV RMSE	CV AUC	CV threshold	CV expected Loss
1	Manufacturing	0.395	0.638	0.532	0.199
2	Services	0.363	0.692	0.526	0.169

More differences can be seen in the respective confusion tables. First, manufacturing exhibits a higher prevalence of high growth firms (20.6% versus 17.6%). Nevertheless, the model performs slightly better in predicting high growth firms in services. In this sector it exhibits a lower false positive rate which is our main goal. (However, it has a higher false negative rate). Notably, the false positive rate in our holdout set is basically at zero, meaning we classified (almost) no firm in the services sector as fast growing that did not turn out to become a high growth firm. This finding points to the services industry as an attractive market albeit its lower average growth rates. It seems to be the case that the characteristics of firms in services make their future growth quite predictable.

Table 3: Confusion Matrix Manufacturing

	Actual Low Growth	Actual High Growth	Total
Predicted Low Growth	79.34%	6.23%	85.57%
Predicted High Growth	0.05%	14.38%	14.43%
Total	79.39%	20.61%	100.00%

Table 4: Confusion Matrix Services

	Actual Low Growth	Actual High Growth	Total
Predicted Low Growth	82.36%	11.19%	93.55%
Predicted High Growth	0.03%	6.43%	6.45%
Total	82.38%	17.62%	100.00%

Variable Importance

As a last task, we looked at some methods to interpret our random forest probability model. Therefore, we set up a dashboard, which can be found in our dashboard (see github folder, incl. README).

Overall Summary

We estimated three models to predict high growth firms (logit, logit LASSO, random forest). Random forest turned out to be the best model, both, when evaluated based on RMSE or AUC as well as expected loss.

Our definition of high growth is based on average annual sales growth over two years ahead above 30%. In our loss function we assigned 50% higher costs to false positives, reflecting our desire to avoid bad investments. This decision lead to a high specificity of our classification but at the same time to a low sensitivity. We prefer this trade-off, since we do not need to identify all high growth firms but rather just find some of them with enough confidence to make an investment.

We also created some visualizations that help understand the influence of important variables of our best model. These visualizations including some discussions can be found in a dashboard that is linked to this report.

In a next step, we applied our model to the two segments of manufacturing and services. The model performs better in the services sector leading to lower false positives and lower false negatives.

We are confident that our model points us to some of the firms with the highest growth potential and believe it to be a valuable tool for predicting high growth firms. At the same time, we realize, that our high specificity comes at the cost of a low sensitivity. This means that we will miss many other high growth firms.