Finding fast growing firms

Anton Shestakov
Syuzanna Poghosyan

April 2025

Target variable

To define a target variable, the study turns to corporate finance framework and accounting theory. Given the assignment setup and data availability, including various balance sheet and P&L indicators, the firms' pace can be determined by custom methods. In this chapter, the discussion comes down to three different ways of fast growing firm identification. In particular, through total assets in BS, net profit/loss in P&L, and total sales.

The method of determination firms' growth rate through evaluation of total assets is considered to be one of the most trivial. It ranks right up there with sales, revenue and workflow. The total assets demonstrate a nice snapshot in a definite period of time. It allows to evaluate the general firm's well-being through analysis the capital structure. However, the latter varies a lot across the type of companies's activities. Specifically, the study in this paper includes manufacturing and service firms which have non-similar BS structure and assets' liquidity capability. For example, manufacturing companies have likely a greater share of long-term assets (property, plants or machines) than service firms. In turn, non-current assets are stable and not subject to drastic quantitative change within 1 year compared to current assets, which probably has a greater impact on service companies. Thus, a manufacturing company can show a great growing performance in 1-2 years without a considerable change in its total assets at the same time. There are such factors as optimization of plants usage, learning workers and etc. that might increase firms' market capitalization, revenue or net profit. And this fast growing can be not fully shown in change of total assets.

The second method turns to estimation through net profit/loss line in P&L statement. As a matter of fact, this method is intersected with sales and conceptually close to it, so initially the second option should've been the number of labor force index. However, it was decided not to embed the latter due to a considerable number of missing values. Getting back to net profit, this indicator is the most accounting-driven. It can be exposed to different account shenanigans for the sake of tax payment cutting. Companies are used to practice artificial understatement of net profit by re-investing, increasing expenditures and etc. As a result, it can reflect not a real growth rate of the firms. Of all the options available, this one is the least reliable.

Finally, the last option under consideration is sales which basically reflect firm's revenue. This study will stick to this value as a measure of fast growing firms determination. Sales are a strong indicator for identifying fast growth companies because it directly measures a company's ability to generate revenue and capture market demand. Moreover, sales are balanced indicator that is more difficult to manipulate with. Unlike net income, which can be distorted by accounting tricks with taxes or timing of cost structures, sales provide a clear picture of firm's performance. In addition, for many companies it is a KPI regardless of the type of activity. Specifically, manufacturing and service spheres can be estimated by this value. Thus, versatility and reliability are the reasons for choosing this variable as the target one for this work.

Data

Detailed data pre-processing and cleaning with technical report in markdown is presented in folder **'Assignment 3'**: https://github.com/Anton21a/Machine-Learning

The study measures growth rate of companies in the span of two years (2014 vs 2012). The choice is based on the idea of capturing non-1year-short-term changes in firms' sales but rather determining more structural growth over a longer 2 years period. Beyond the deepening in technical part, the dataset involves growth rates of different firm's financial indicator from both BS and P&L statements. The values were calculated through logarithmic difference method. The log growth rate results are close to the standard way of estimation but the skewness of the data is also taken into account, adjusting for the asymmetry of the values. The description statistics on some numeric variables is shown in Table 1. Later, the study embeds categorical 'flag' variables that catch conditional difference within different variables.

The outcome variable is defined as a binary variable "fast_growth_f" showing whether a firm increased its sales by more than 71.8% in a span of two years between 2012 and 2014.

Variable	**Obs**	**Mis**	**Mean**	**SD**	**Min**	**Median**	**Max**
comp_id	16184	0	-	-	-	-	-
sales	13892	0	301218.1	872978.2	1000.0	61942.6	9963926.0
founded_year	32	0	2002.7	7.0	1951.0	2003.0	2014.0
ceo_count	7	0	1.3	0.5	1.0	1.0	7.0
foreign	10	0	0.1	0.3	0.0	0.0	1.0
female	10	0	0.3	0.4	0.0	0.0	1.0
age	32	0	11.3	7.0	0.0	11.0	63.0
age2	32	0	177.9	181.0	0.0	121.0	3969.0
sales_2012	13625	0	251101.8	735154.4	1000.0	54022.2	9786907.0
growth_ln_sales	16100	0	0.1	0.9	-6.3	0.1	8.5
growth_ln_curr_assets	16157	0	0.1	1.2	-10.1	0.1	11.4
growth_ln_curr_liab	16086	0	0.2	1.7	-14.3	0.1	13.2
growth_ln_extra_inc	1912	0	0.6	2.7	-15.0	0.0	14.4
growth_ln_fixed_assets	13595	0	0.6	2.7	-15.6	0.0	13.3
growth_ln_intang_assets	2547	0	0.0	1.9	-12.3	0.0	13.9
diff_profit_loss_year	15919	0	0.4	3.9	-91.3	0.0	126.5
diff_share_eq	16001	0	-0.2	5.0	-94.9	0.0	97.7
growth_rate	16100	0	1.8	43.9	-1.0	0.1	5091.1

Table 1: Summary Statistics of Numeric Variables

Models & Results

The first step of the probability prediction analysis is running logit model using glm method with 5 fold cross-validation. The model includes 5 specifications with different sets of variables.

The second method for prob prediction in the study is LASSO logit regression. The model employs the set of variables used in the fourth specification of logit model with glm. In contrast, LASSO includes (λ) coefficient which is a regularization strength over a grid of 10 values placed between log 10^{-1} and 10^{-4}. It applies the same idea of 5 fold cross-validation through which the best RMSE performance is selected with optimal (λ).

Figure 1 shows intermediate comparison between five logit specification and LASSO. It can be observed that the latter, on average, outperforms each logit specification except of the result

Table 2: Model Specifications for Fast-Growth Prediction

Specification	Included Variables
Equation 1	Firm age (linear and squared), foreign management, female ownership, municipality size, industry type
Equation 2	All in Equation 1 + Growth variables (e.g., sales, assets, expenses)
Equation 3	All in Equation 2 + Dummies for >2× growth (dummy_gt2x_*)
Equation 4	All in Equation 3 + Dummies for > -50% growth (dummy_drop50_*)
Equation 5	All in Equation 4 + Interaction terms (industry × growth, female × growth)

in the third fold. The difference in RMSE results are not so varied across other folds, but the third that might have an impact on the future model estimation in the framework of loss function introduction.

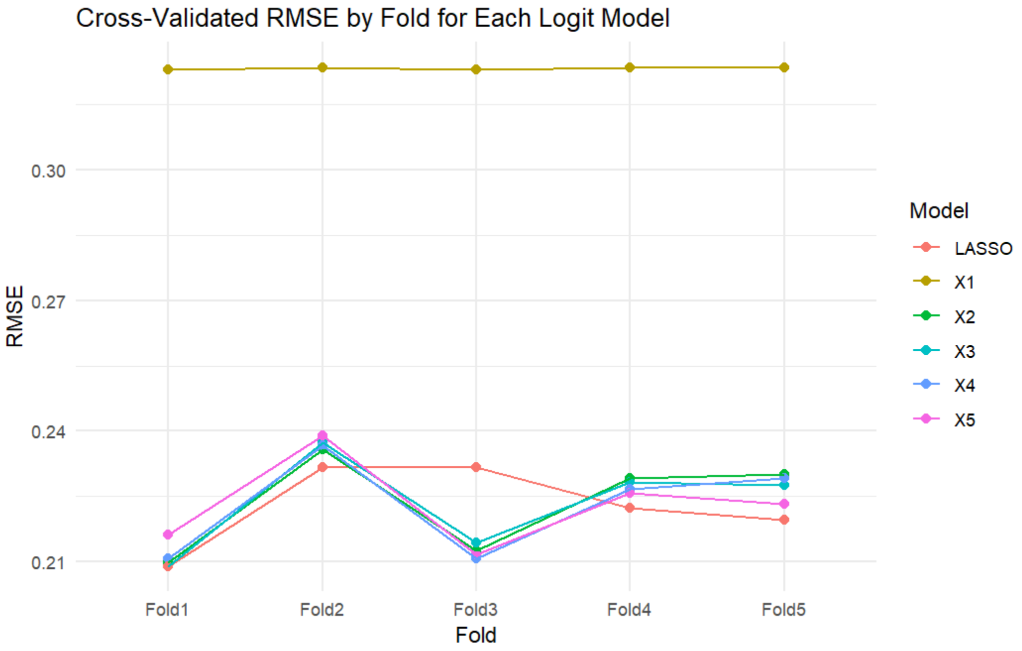

Figure 1:

The last method used in the study is Random Forest. To optimize model performance, the analysis performs a grid search over three hyperparameters: the number of variables randomly selected at each split (mtry = 5, 10, 15), the minimum node size (min.node.size = 5, 10), and the splitting parameter which is the Gini impurity index. The cross-validation procedure identifies the combination of hyperparameters that outputs the best predictive performance.

The Figure 2 demonstrates the total results by RMSE across the best specifications out of logit model, RF and LASSO. For each fold RF outperforms other models with results between 0.19 and 0.212.

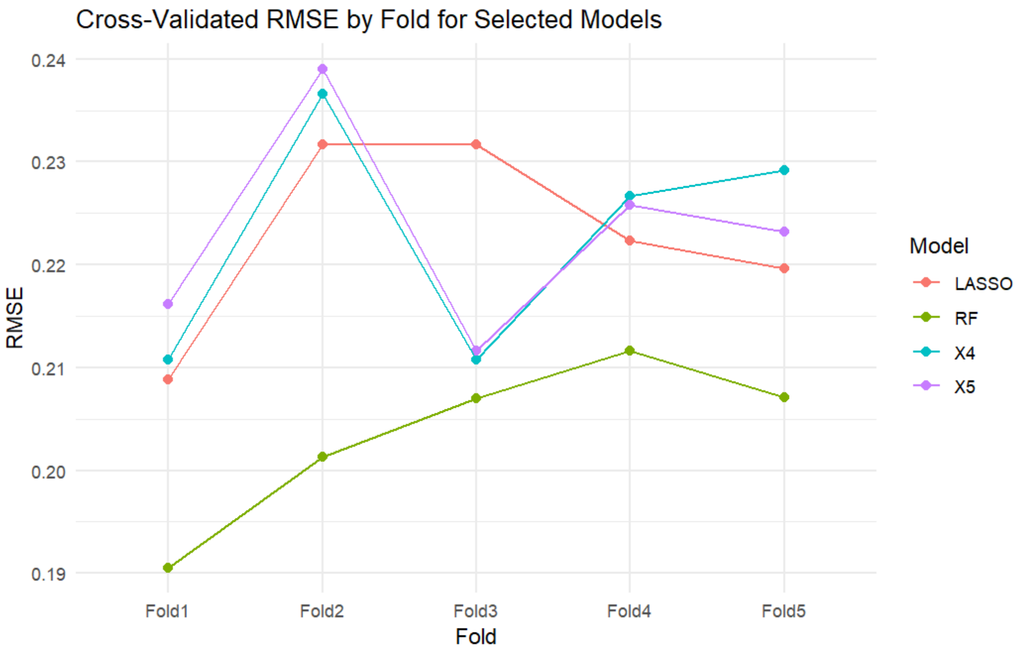

Figure 2:

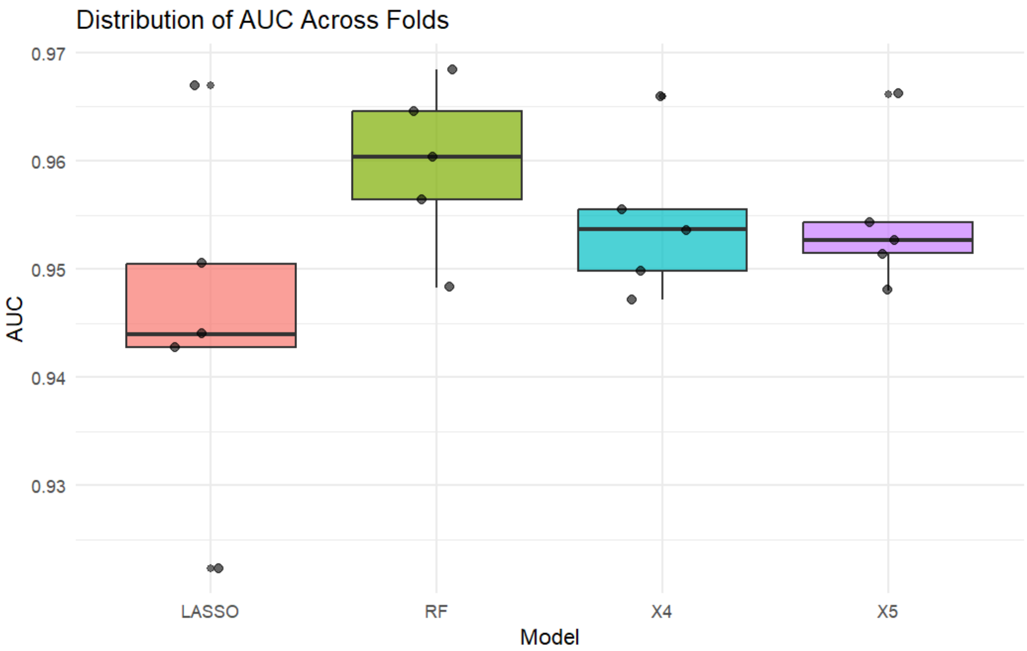

Figure 3:

Figure 3 shows boxplot chart with AUC distribution across the same models with 5cv method. Area under the Curve measures a model's ability to distinguish between firms that have fast growth rates and those that do not. The results depict that Random Forest model has the best ability to determine whether a firm is fast growing or not.

Within each model the probabilities in the range of 0 and 1, whether a firm will face with fast growing, are calculated. Random Forest method runs classification for holdout data catching the binary option whether firm will get into a fast growth phase or not.

Loss Function

To define a loss function, the study makes up some assumptions. Firstly, the case relates to investing company which is not risk-averse. Secondly, average return on initial investment is 3x. In other words, if our investing firm doesn't invest in a company but it is actually fast growing one, then lose opportunity is defined as 30.000$ (FN). If the investing firm invests funds but a company actually is not fast growing, then it loses 10.000$ (FP). The detailed information is shown in confusion matrix (Table 3).

Actual / Predicted	**Fast Growth (Yes)**	**No Growth (No)**
Fast Growth (Yes)	True Positive (TP) We invest in a firm, and it will quickly grow	False Negative (FN) We miss a fast-growing firm, and lose opportunity
No Growth (No)	False Positive (FP) We invest in a firm, but it doesn't grow fast	True Negative (TN) We don't invest in a firm, and it doesn't grow fast

Table 3: Confusion Matrix for Predicting Fast-Growing Firms

Analysis is separated on two parts. It includes finding average expected loss and avg optimal threshold by model across industry type: manufacturing and service. Table 4 shows results on manufacturing companies. Random Forest demonstrates the least value of avg expected loss and the highest optimal threshold among other models. So, using this method for making-decision whether company is fast growing or not, and with condition that each FP costs 10k $ while each FN costs 30k $, the average loss per fold is 859.62 dollars. Figure 4 in Appendix gives a visual results on avg exp. loss investing in the manufacturing firms.

Model	Avg. Expected Loss	Avg. Optimal Threshold
X4	980.299	0.126
X5	1142.511	0.082
LASSO	1051.741	0.120
RF	859.621	0.169

Table 4: Average Expected Loss and Optimal Threshold for manufacturing firms

The same calculations were undertaken for service firms. The sample size differs between the manufacturing and service firm databases: the service firm dataset includes almost 12500 observations, while the manufacturing firm one includes about 3000. This might partially explain lower variation in avg. expected loss by models in Table 5. Nevertheless, RF model demonstrates the best performance with the value of 938.11.

Model	Avg. Expected Loss	Avg. Optimal Threshold
X4	966.96	0.1552
X5	964.94	0.1584
LASSO	1036.02	0.1502
RF	938.11	0.1938

Table 5: Average Expected Loss and Optimal Threshold by Model for Service firms

In general, comparing two results on firms' type of activity, it can be inferred that RF holds the bar for the best model while LASSO - for the worst. Two specifications of logit model show mixed relative results. Most likely, it is connected with the size of the dataset.

Discussion

Beyond defining the best model's performance and technical part of the prediction, the obtained results align with the specified settings and design. All models in both cases demonstrate the values of average optimal threshold less than 0.2. That means that for our investing company prioritizes the risk of losing money from False Positive error rather than lose the opportunity and commit the error of False Negative.

Another important mention from the analysis is the robustness of the Random Forest model across different samples by type of firms' activity. RF shows consistently better RMSE and AUC values, outperforming other methods used in the work.

Appendix

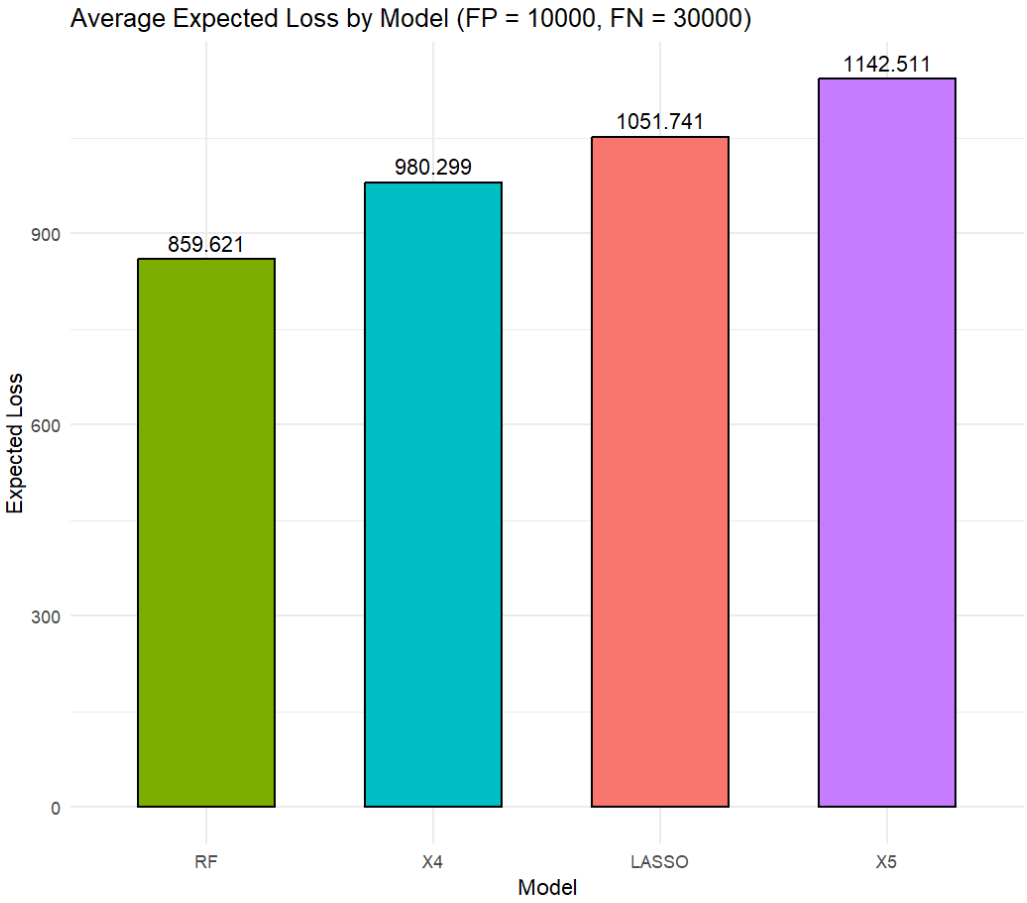

Figure 4: