Summary report

In this report, I discuss how I predicted the fast-growing firms and model outcomes based on available firm level characteristics using bisnodes-firms data. Codes and outcomes can be found at: https://github.com/Dilnovoz/Data-Analysis-3/tree/main/Assignment_3

Sample are restricted to the firms who are present in the market once they entered to it or in other words, defaulting firms are removed from the analyses. I aimed to have quite a homogenous sample of firms and among those, I measure and predict their growth rates: Defaulting firms are more sensitive portion of the sample that pooling them with “not fast-growing firms” in one category against fast growing firms may drive the outcomes. As an investor in the first step, you try to be sure that the firm will be in the market as you invest and then in the second step, you choose among these firms to which to invest to get the highest return. Further extremely small and large firms are also removed from the sample.

Firm growth level is measured as percentage change in the sales volume from 2012 to 2013 and firm characteristics in 2012 are applied to predict the firm sales growth in 1 year period. Top 15 percent of the firms that had highest growth in sales volume are *categorized* as fast-growing firms which had growth over 38.7 percent at minimum. Growth rate of the firms can be measured in various ways. For example, based on their employment changes over time or return on investment over time as well as opening new branches, entering to new markets. In this dataset employment is quite noisy variable so I decided not to use it to measure the growth rate of the firms. Moreover, I could measure growth rate not only 1 year period difference but also in two-year period or constantly growing in several years. But I believe measuring several year period sales growth rates is more proper to measure sustainable growth over time rather than fast growing firms. Hence, I used above mentioned way to measure fast growing firms in the market.

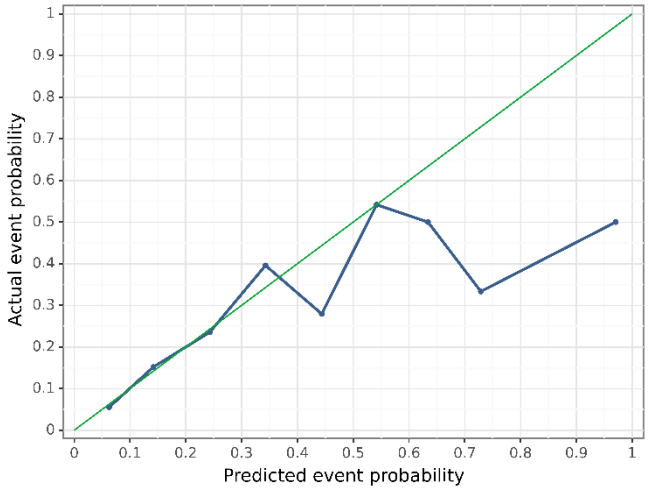
Features that are applied to the model can be grouped into several categories: Firm related factors (age, location, and industry of the firm), financial indicators (different types of assets, liabilities, and equity related information taken from balance sheet and P&L statements), human capital (age gender of the CEO, foreign management, labor force), and historical factors are controlled using sales growth from 2011 to 2012.

Fast growing firms are measured as dummy variable, I applied logit regression estimation with different set of features as well as Random Forest model to find the optimal model for my prediction

and classification analyses. Based on cross validated performance of the models, I chose logit model with 76 variables (M4) as RMSE score is quite small for this model and even logit model with interactions (M5), Logit with Lasso and Random Forest are using quite complicated ways and requiring more time, they're not improving the quality of the model significantly.

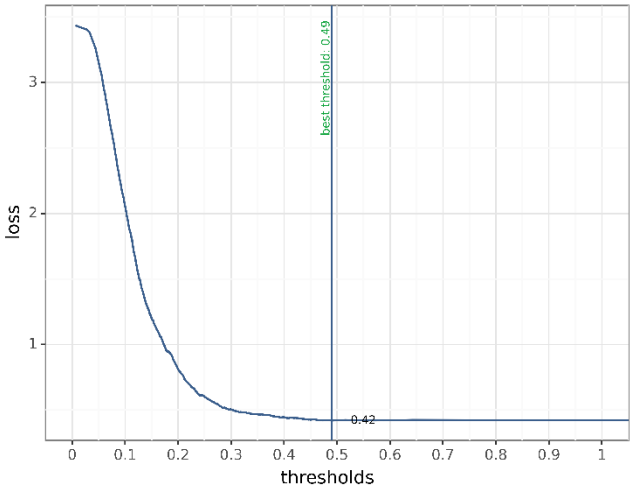
Table 1

	Number of Coef	CV RMSE	CV AUC	CV threshold	CV expected Loss
M1	12	0.340415	0.649101	0.665775	0.41755
M2	19	0.339111	0.667022	0.746741	0.418368
M3	33	0.339367	0.664183	0.948112	0.418293
M4	*76*	*0.333711*	*0.701493*	*0.502432*	*0.412423*
M5	143	0.33317	0.706208	0.600458	0.412348
LASSO	68	0.332403	0.709634	0.513219	0.413017
RF	n.a.	0.332119	0.717473	0.953849	0.415693



Further, based on Model 4, I predicted probability of being fast growing firms on holdout set and as it is visible from the graph, our model is doing good job for the lower actual probability values, however it is overpredicting for the higher actual probability values and deviating from the ideal prediction line.

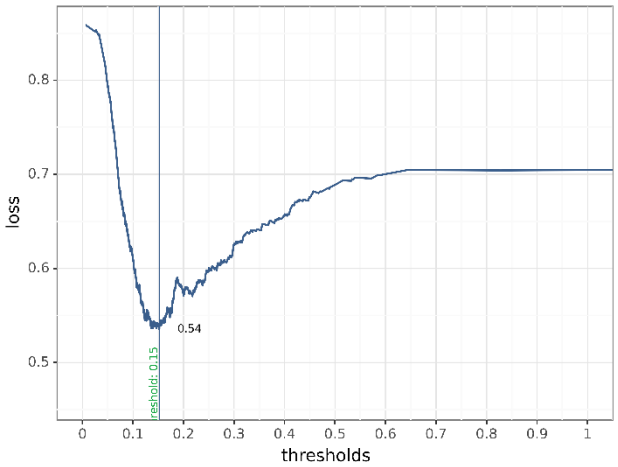
There are two separate cases can be studied for this classification 1) which is most likely: if investor plans to invest into the firm, it brings more profit to find fast growing firms. Let's assume if investor invests in fast growing firms, her return is 4 euros per 1 euro investment while the rest of the firms will bring 3 euros per 1 euro invested. In this scenario, false positive (wrongly assuming not fast-growing firm as fast growing firm) costs investor 4 euros while false negative (vice versa) costs her 3 euros. As we can see from last column of table 1, the minimum cost is achievable with Model 4 and threshold of 0.5 (which is interestingly coinciding with default classifying threshold of the model but with proper reason behind 😊). In this scenario, on the holdout set, model is predicting 21 fast growing firms and 2887 not fast growing firms correctly and the investor is getting what she expected, while for the rest of the outcome, model incorrectly classified

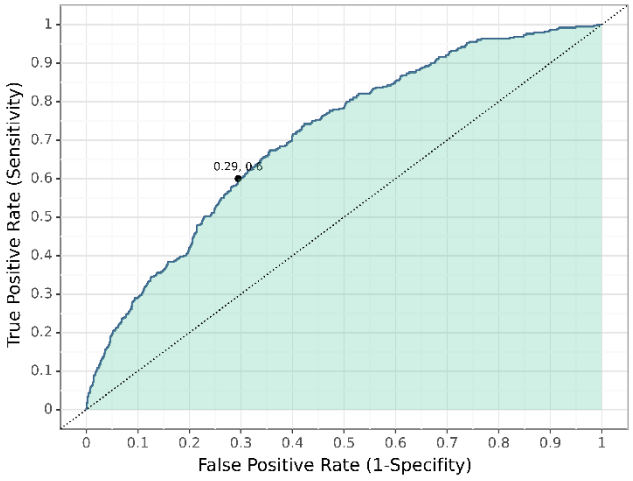


	Predicted no High_growth	Predicted High_growth
Actul no High_growth	2887	20
Actual High_growth	437	21

On the other case if governmental organization that are aiming to help slow growing firms in the market, then it is better to find nongrowing firms and help as it is more valuable to this firms than fast growing firms. Assume helping 1 slow growing firms create 5 new employees in the market while helping fast growing firm will create only 1 extra job in the market. In this scenario optimal classifying threshold is 0.15 and the wrongly predicting fast growing firms increases noticeably.

New classification:





	Predicted no High_growth	Predicted High_growth
Actul no High_growth	2233	674
Actual High_growth	227	231