Assignment 3 - Finding fast growing firms

Balint Parragi, 2023-03-10

Preliminary data inspection, cleaning and data munging

The business task is to predict fast-growing firms within the sample data (bisnode) using all possible covariates and domain knowledge. The raw bisnode data consists of 287829 observations and contains data regarding European firm-level data over the period of 2006-2016, with many dimensions (sales, profits, employees, CEO characteristics, balance sheet attributes, etc.). As some variables are not well presented (many NA-s or inadequate values), and the panel is unbalanced as well, we need to do some preliminary cleaning and computation.

Some of the most important steps:

- Non-NA and positive `Sales`
- Non-zero `Labor avg`
- If Labor_avg is missing, then impute it with the median with respect to each year
- Create newly-founded indicator
- Regroup industry indicator based on NACE categories
- Calculate cost, expenses, profits in percent of sales, calculate various asset categories in percent of total assets
- Remove the very end of distribution for some variables and add a flag that those values are modified (thresholds are selected that very a few observations were modified)
- Finally, restrict the sample to year 2012, and design target variables - rapid growth.

Defining the target

Rapid growth of a firm can be measured in various ways, but both words have to be cleared: how long is a "fast" period and what should grow? For the latter, it can be the increase of sales, profits, number of employees or turnover. Data is not available for the last, and unfortunately, for the number of employees (`labor_avg`) a large share (50.9%) is missing in the raw data.

However, I will try out 3 different metrics with 4 different variables. The metrics are:

- 1-period change (from 2012 to 2013)
- 2-period change (from 2012 to 2014)
- 3-year moving average of 1-period changes (averaging the yearly changes from 2012 until 2015)

and the variables are:

1. Sales (in logs)
2. Number of employees (annual average, in logs)
3. Modified number of employees (annual average, in logs, missing data replaced with median)
4. Profits-to-sales ratio (in levels)

and one extra measure (according to the OECD's definition, available here):

"All enterprises with average annualised growth greater than 20% per annum, over a three year period should be considered as high-growth enterprises. Growth can be measured by the number of employees or by turnover. [...] A provisional size threshold has been suggested as at least 10 employees at the beginning of the growth period." — OECD, 2007

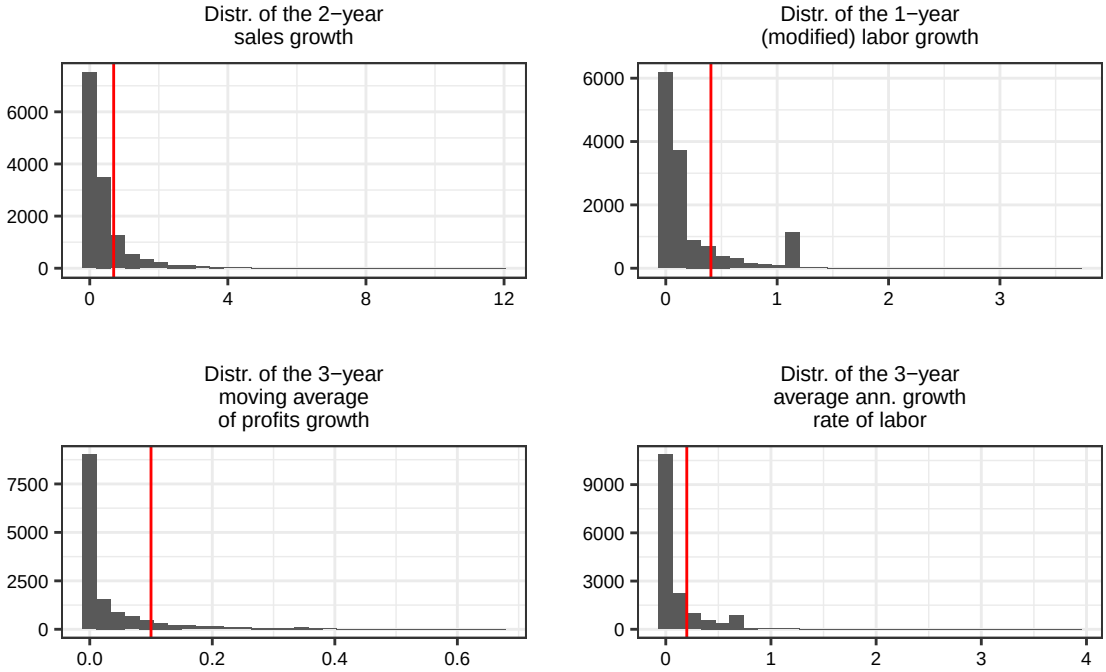
These various measures can be helpful in uncovering many aspects, though computationally it would expensive to conduct parallel analysis for all. Sales is potentially the best measure to illustrate the expansion of a company, but do all firms do sales

activities as their main profile? Profits can be similar, even though they can hide the fact that a fast-growing firm might create losses in some periods while it is growing - and it can produce higher profit margins later. Number of employees (according to the OECD definition) would be the most sophisticated measure.

The selection of the change metrics is also difficult: should we only focus on very short-term, immediate boom (1-year change), or on more periods? The rolling mean is the most conservative as it considers a 3-year window, and potentially a worse year with a small increase in output can result in a higher metrics value. The 2-year metric is the middle ground: ephemeral success is not enough, moderately permanent growth is needed.

One last important part is to select the thresholds for these targets, above which a firm is considered as *fast-growing*, otherwise not. As apart from the OECD definition there is no clear-cut choice, my decision depends on some distributional inspection.

A winsorized measure could have been used, but my aim was to have reasonable and easy-to-interpret thresholds while have a sample size of fast-growing firms around 10% of the sample. The thresholds for sales and profits are generally larger than for labor, as it is much more feasible to increase these as employ more and more people on the short run. Also, 2-year change is assigned a higher threshold as continuous growth would imply that from the same starting point a higher growth can be achieved in 2 years compared to 1. But the 3-year moving average is again assigned a much lower threshold as it is very difficult to maintain such a large average growth over a 3-year span. One simple example: 1-year change in sales is linked to a threshold of 0.56, which is $e^{0.56} \approx 1.75$, so it means a 75% increase in sales over 1 year. The same for the 2-year change is 0.693 ($e^{0.69} \approx 2.0$ - doubling sales), and for the moving average it is 0.405 (so $e^{0.4} \approx 1.5$), achieving 50% increase for three years on average). For the OECD measure, the sample size is extremely low (7.92‰Ğ), so it might not be a good measure in this dataset.



Correlation table

	sales d1	sales d2	sales rm3	labor d1	labor d2	labor rm3	labor mod d1	labor mod d2	labor mod rm3	profit d1	profit d2	profit rm
sales d1	1											
sales d2	0.472	1										
sales rm3	0.363	0.578	1									
labor d1	0.155	0.142	0.085	1								
labor d2	0.089	0.182	0.13	0.155	1							
labor rm3	0.147	0.25	0.243	0.414	0.441	1						
labor mod d1	0.158	0.131	0.108	0.434	0.014	0.213	1					
labor mod d2	0.085	0.176	0.131	0.116	0.79	0.377	-0.004	1				
labor mod rm3	0.123	0.196	0.21	0.255	0.25	0.493	0.44	0.303	1			
profit d1	0.235	0.197	0.161	0.017	0.011	0.011	0.084	0.04	0.047	1		
profit d2	0.166	0.282	0.217	0.014	0.032	0.037	0.067	0.075	0.085	0.496	1	
profit rm3	0.127	0.218	0.259	0.02	0.042	0.059	0.057	0.071	0.115	0.389	0.557	1
labor oecd	0.017	0.054	0.03	0.104	0.104	0.225	0.04	0.085	0.097	-0.019	-0.015	-0.017

The correlation table of the targets suggests that the targets based on labor and modified labor are quite similar with higher positive correlation, and also sales and profits are quite similar within their respective group (1-year, 2-year difference; 3-year-

mean). But across the different targets, there are not much similarity, which indicates that the selection of either one can potentially yield distinct results in the end. However, because of *completeness* and *reliability*, I select `Sales` as the basis of the analysis. Unfortunately, computationally and because of time constraints it is note possible to run the whole analysis with a different target, but I believe there might not be large differences along model selections (but can be different in terms of prediction accuracy and loss value).

Among the covariates, one could be genuinely interested in `new`, as it is plausible that newly-founded firms grow the fastest (because of large marginal returns and not-yet diminishing marginal production).

Also note that the sample period (firstly reduced to 2010-2015, then to 2012 with growth measures computed using data from 2013-2015) covers largely a boom period in the EU, but not everywhere, as there was the Eurozone recession in 2012, years after the GFC. Controlling for such time effects is out of the scope of this analysis.

Part I: probability prediction and model selection

Models defined

The target is then whether the 2-year change of log sales is above 0.7, so the sales of a company should be doubling to qualify as fast-growing. As there are many different explanatory variables, I refer to some larger categories they are within.

1. Model 1: Logit model; $log\,million\,sales + log\,million\,sales^2 + labor + profits + industry\,category_2$
2. Model 2: Logit model; $Model_1 + Fixed\,assets + Shareholder\,equity + Current\,liabilities + Age + Foreign\,management + Flag\,variables$
3. Model 3: Logit model; $Model_2 + firm\,related\,variables + others\,based\,on\,equity\,and\,sales$
4. Model 4: Logit model; $Model_3 + HR\,variables + other\,flags + quadratic\,terms$
5. Model 5: Logit model; $Model_4 + interactions$
6. Model 6: Logit LASSO, based on $Model_5$
7. Model 7: Random Forest with basic tuning, no interactions or quadratic terms

The interactions included are between:

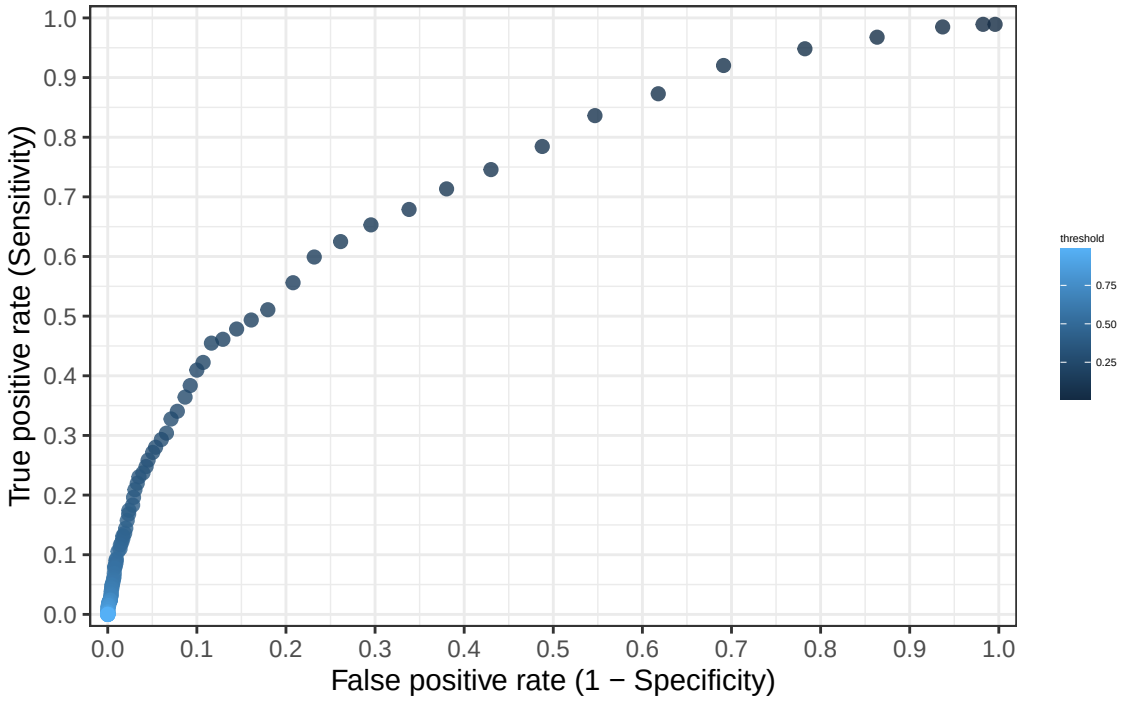
- industry categories and HR features
- sales and profits

	Number of predictors	CV RMSE	CV AUC
X1	11	0.326	0.686
X2	18	0.320	0.724
X3	31	0.320	0.732
X4	75	0.319	0.734
X5	135	0.320	0.734
LASSO	109	0.336	0.711

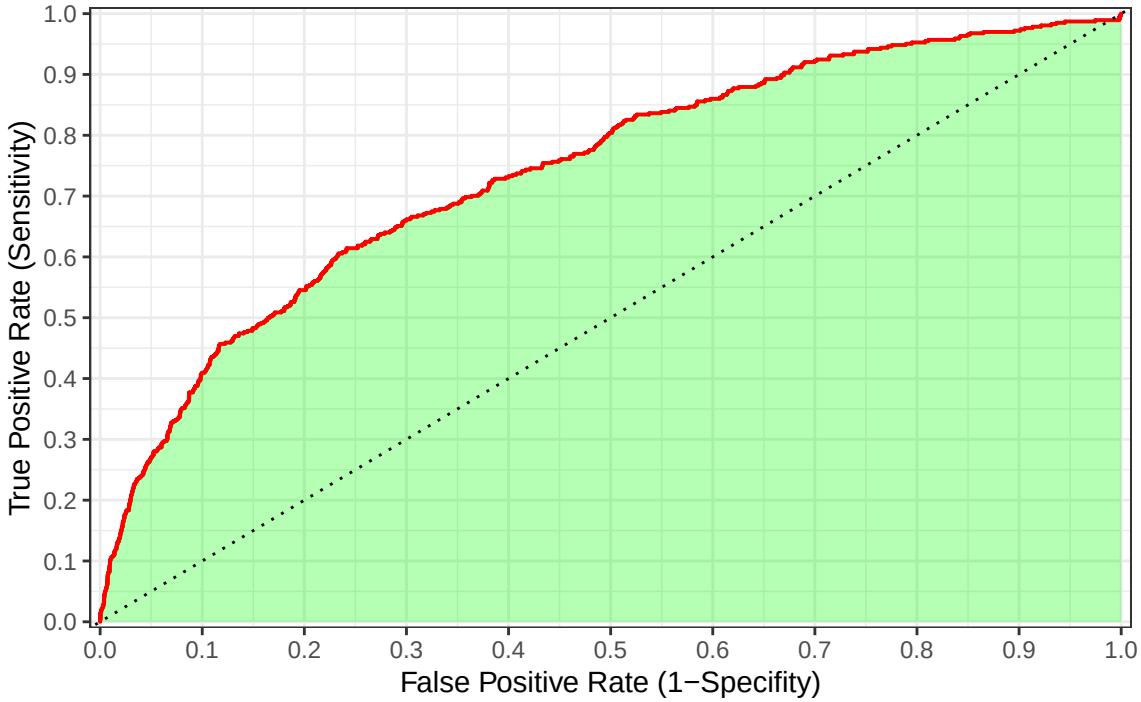
Some models are very close across the performance measures, but probably $Model_4$ is among the ones with the lowest RMSE and the highest AUC values, so this is the preferred Logit model. LASSO interestingly does not perform well. The features with non-zero coefficients are very educative with their sign and size:

- Sales has a relatively large negative coefficient, supporting the theory of diminishing returns of growth/production
- Inventories, quadratic profit (which is quite tricky as loss also becomes positive...), shareholder equity, some indicator categories all have a positive coefficient, as well as *new* (quite large:)
- Age of the company and the CEO have negative coefficient which is again in line with theory
- Only one flag is included: labor's flag as it contained a lot of missing data

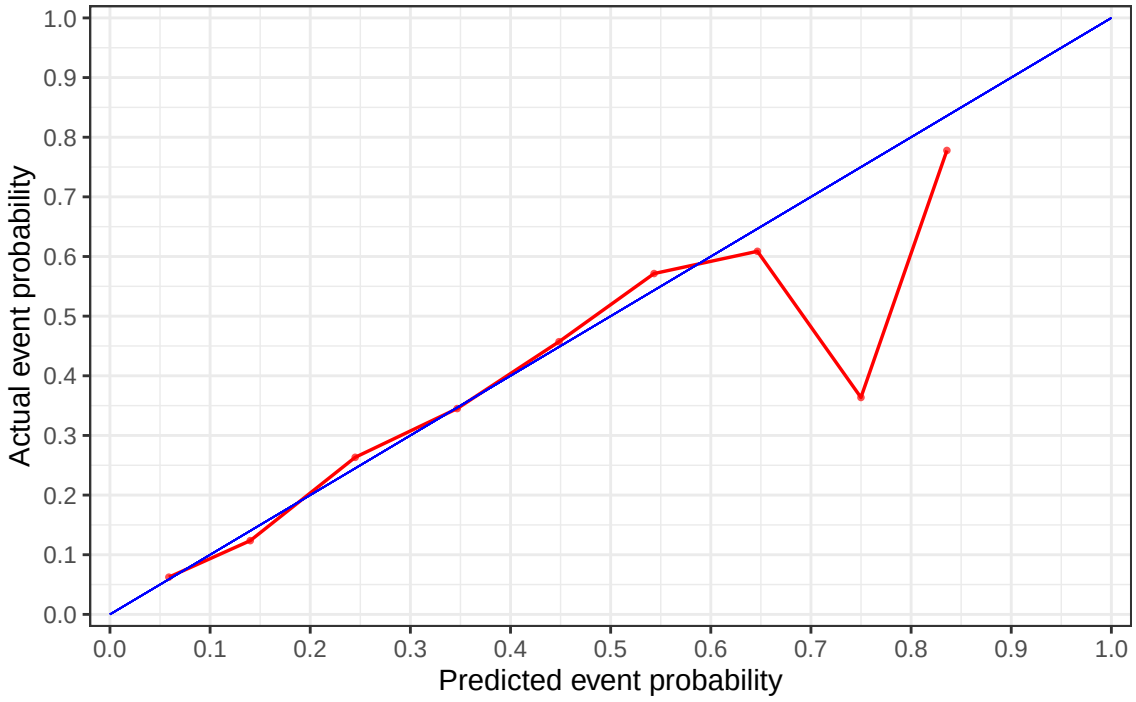
The plot with $Model_4$ shows the trade-off between the True and False positive rates for many thresholds. It is somewhat clear that a smaller threshold is needed to get better-off, which is not surprising as it mirrors the share of fast-growths within the sample (which is small).



The same is visible for a continuous set of thresholds.



Looking at the next plot of the estimated and actual event probabilities, their relationship shows that it is well align on the 45-degree line until the event probability is high, because then it cannot predict them at all, by severely underestimating the probabilities.



Part II: classification

Defining the Loss-function

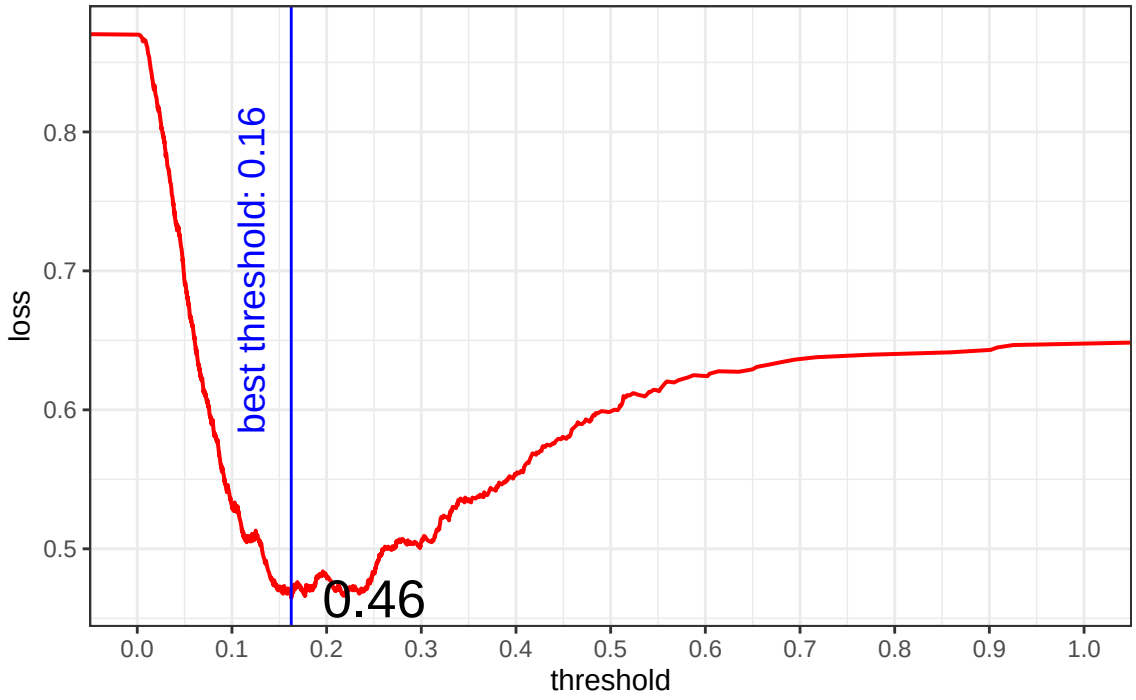
The statistics question is that whether we want to minimize FN or FP instead. As a business related question, this can be translated to the following. What is more costly to me (as an analyst, competitor, investor, etc.): to incorrectly classify some non-fastgrowing firms as fast-growing (False Positive - FP), or to label some indeed fast-growing firms as non-fastgrowing (False Negative - FN). FP error can result in some insufficient or suboptimal financing or investing in non-fastgrowing firms, but those can be still growing at a level that creates reasonable returns (and of course, when these firms are actually financed, it can change the situation that they produce a better growth in the subsequent years). But if we make FN errors and miss out on fast-growing firms, then it can amount to high losses (lost possibility of high revenues) - and other competitors or investors can get it sooner, which also undermines our relative position. Even though FN errors do not consequently lead to financial loss and costs, in missed opportunity cost they can be much higher, so we want to take care of that. Thus, creating our Loss-function should weigh FN predictions more. The ratio depends on the market and our preferences, I set them accordingly (1-to-5 ratio), because of lost investment, stronger competitors (both the actual fast-growing firm and their new investors) and lower potential returns.

$Cost\ function: \frac{FN}{FP} = 5$

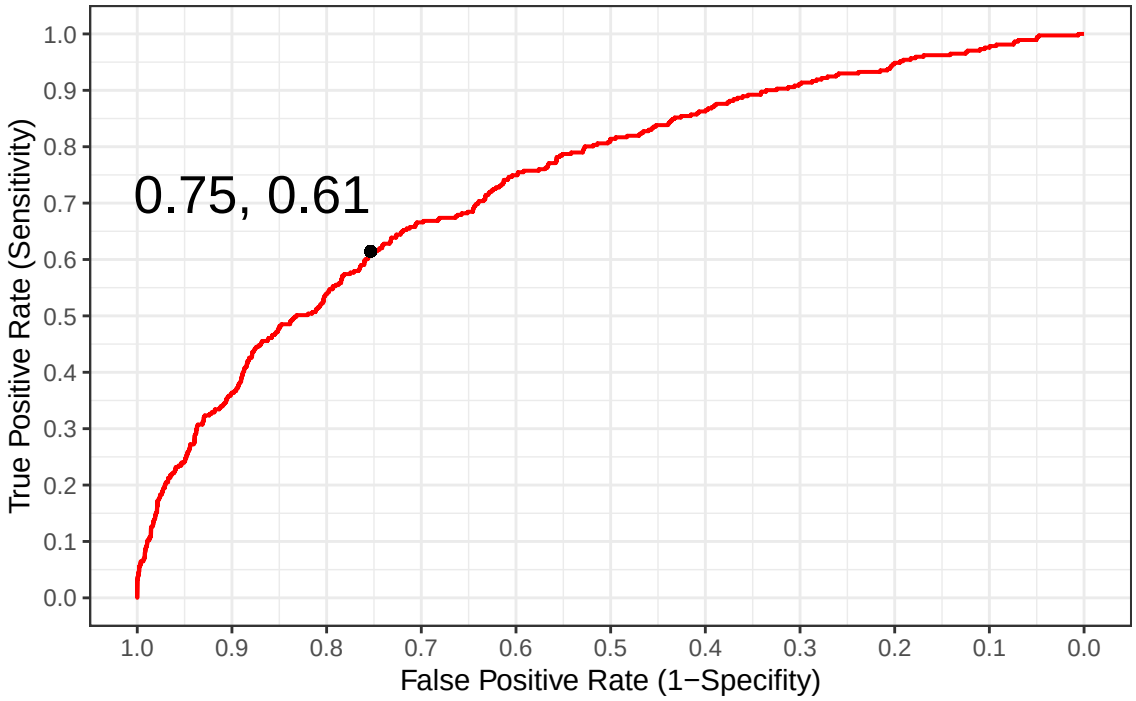
	Avg of optimal thresholds	Threshold for fold #5	Avg expected loss	Expected loss for fold #5
X1	0.156	0.141	0.514	0.541
X2	0.171	0.175	0.474	0.454
X3	0.182	0.190	0.468	0.461
X4	0.172	0.172	0.465	0.479
X5	0.182	0.175	0.470	0.480
LASSO	0.160	0.175	0.495	0.496

Again, $Model_4$ produces the smallest expected average loss. The optimal threshold in this case is very close to our inverse cost ratio (0.2).

For Fold-5, the best threshold is 0.163, which is close the inverse cost ratio and to the sample frequency of fastgrowths.



The next plot shows the ROC-curve's frontier, at the previously shown best threshold's position.



Part III: Discussion

	Number of predictors	CV RMSE	CV AUC	CV threshold	CV expected Loss
Logit X1	11	0.326	0.686	0.156	0.514
Logit X4	75	0.319	0.734	0.172	0.465
Logit LASSO	109	0.336	0.711	0.160	0.495
RF probability	34	0.316	0.742	0.201	0.463

The results of the table above are striking and clear: again, our Random Forest model outperforms every other model specification, but Logit $Model_4$ is very close in terms of every metric (RMSE, AUC, Loss). Indeed, it has more than twice as many features which makes it more complicated, moreover it required model specification as well. But it is useful to see, that something based on our domain knowledge and theoretical reasoning can perform head-to-head with the black box-wise RF method.

External validity and performance

$Model_4$ and the Random Forest gives very similar results on the holdout set. It again emphasizes that these are interchangeable.

```
## [1] "RMSE on holdout with Model 4: 0.316"
```

```
## [1] "RMSE on holdout with Random Forest: 0.314"
```

Confusion matrix with $Model_4$ on the holdout set

	no_fastgrowth	fastgrowth
no_fastgrowth	2564	228
fastgrowth	547	236