Finding fast-growing firms

1. Purpose of the analysis and data

In the following sections, I present a selection of models (logits, logit with LASSO, and random forest) that would help to **predict the fast-growing firms**. Moreover, I underline the main results and considerations that can be drawn from these models.

2. Data and feature engineering

The dataset comes from the *bisnode-firms* data we used in the class, and it was cleaned by the *Data Preparation_Yu.ipynb*. There is a total of 112654 observations from 2010 to 2015 in the end.

a) Label engineering

The **target variable (fast_growth)** is designed from the log difference over one year of the scaled **sales variable**[1]. When the sales growth rate is larger than **0.1**, then the firm is fast growing, which means the target variable equals 1.

b) Sample design

I focus on the small and medium enterprise (SME) sector, captured by firms' sales. Therefore, only firms below 10 million euros and above 1000 euros of annual sales are kept. And the analysis dataset is **panel data from 2010 to 2015**.

c) Feature engineering

The growth rates of firms are related to firm characteristics, financial information, and human resources. Note that I assume that **sales variables can't be used as predictors** since I design the target variable based on sale variables.

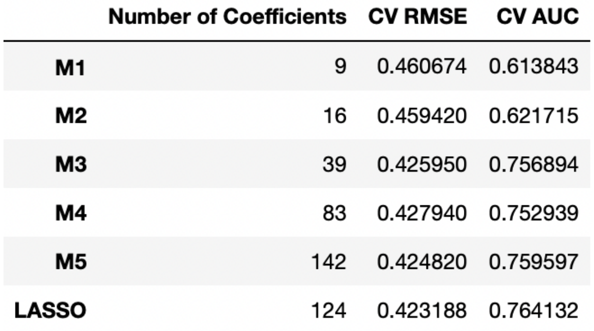
I **winsorize** most of the variables, especially financial variables to capture extreme values and attach the **flag variables** with them. And I also add quadratic terms to capture **nonlinearity** for some variables and **interactions**.

3. Probability prediction and model selection

I start my analysis with **5 logit models and a logit LASSO model**, from the simplest to the most complex (the exact variables contained in these models can be found in the code file). On the training sample, I perform a **5-fold cross-validation** for model selection: the aim is to avoid overfitting (better fit to original data, but worse with live data).

The **logit M5** and the **logit LASSO** models seem to work comparably well, thus I consider the logit M5 model as the benchmark model, which contains almost every variable.

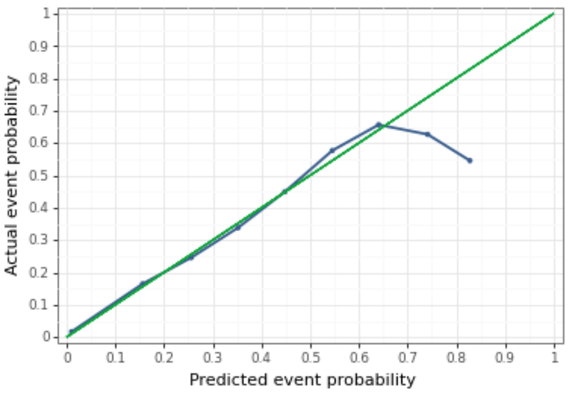
The right figure shows the calibration curve for the predictions. I group predicted probabilities from the logit



	Number of Coefficients	CV RMSE	CV AUC
M1	9	0.460674	0.613843
M2	16	0.459420	0.621715
M3	39	0.425950	0.756894
M4	83	0.427940	0.752939
M5	142	0.424820	0.759597
LASSO	124	0.423188	0.764132

[1] The average of the sales growth rate is 0.043, so I use 0.1 as the threshold to classify whether the firm is fast growing or not. Of course, it is also reasonable to use profits growth rate to recognize the fast-growing firms.

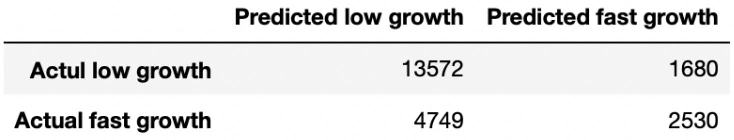
M5 model into 10 bins, and plotted the average predicted probability for each bin with the proportion of y = 1 case for those observations. The calibration curve shows that the model is **quite well-calibrated for the low-level probability** but it **doesn't perform well at the high-level probability**.



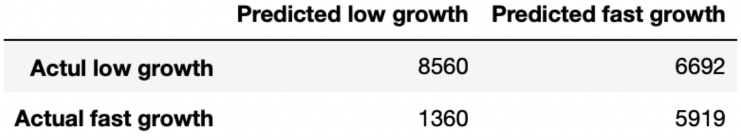
4. ROC curve and classification diagnostics by the confusion table

a) Confusion table with different thresholds

There are confusion tables for two possible thresholds: **0.5 and 0.323** (mean of predicted probabilities). The above one is the threshold of 0.5 while the below one is the threshold of 0.323. Having a **higher threshold** leads to **fewer predicted fast-growth firms**. With a higher threshold, there are also **fewer false positives** but **more false negatives**.



	Predicted low growth	Predicted fast growth
Actul low growth	13572	1680
Actual fast growth	4749	2530

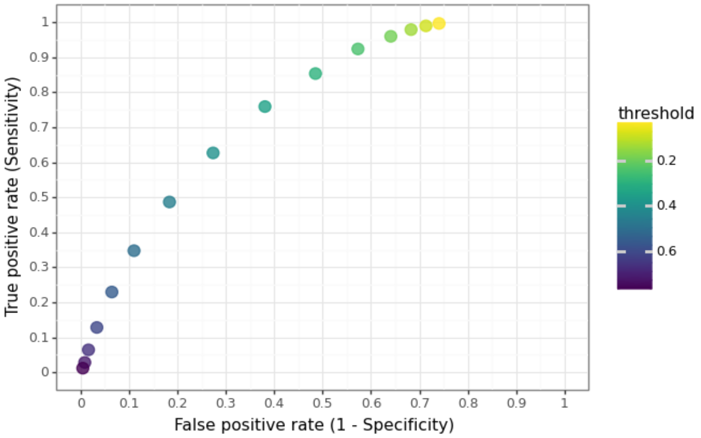


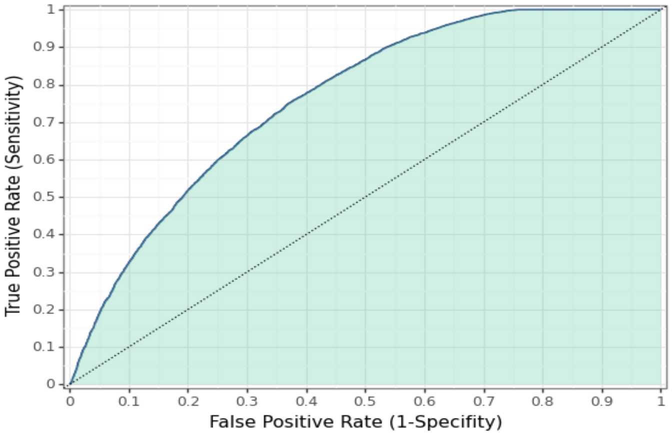
	Predicted low growth	Predicted fast growth
Actul low growth	8560	6692
Actual fast growth	1360	5919

b) ROC visualization (with thresholds in steps) on holdout

The first figure shows values of the **ROC curve** for selected threshold values, between 0.05 and 0.75, by steps of 0.05. It also uses color coding to denote the approximate values of the corresponding thresholds, which are not shown directly on the ROC curve. The second figure is similar to how the ROC curve is usually shown. It fills in for threshold values in between, but it has no reference to the corresponding threshold values.

The logit M5 model's **AUC** is 0.76 which is close to that of the logit LASSO model. And the order suggested by RMSE and AUC is very similar but not the same. That's a rather typical outcome.

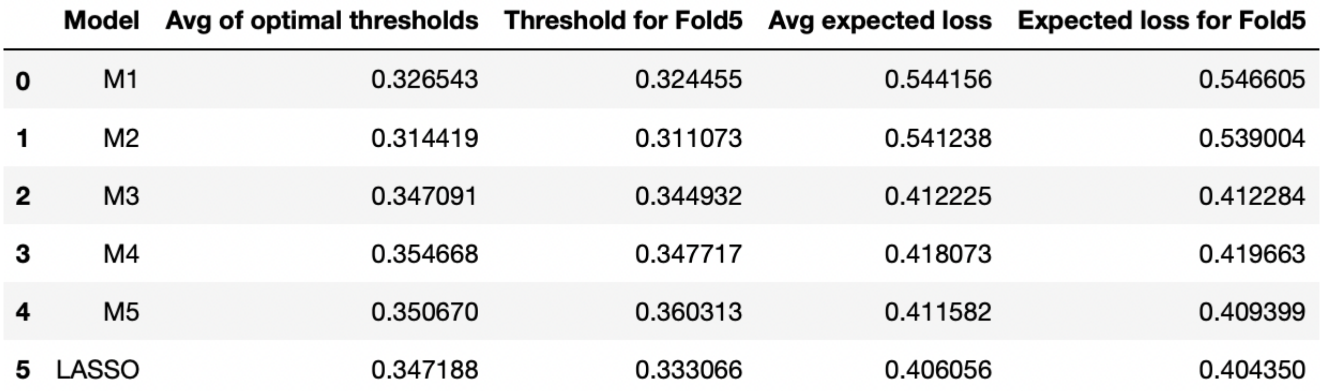




5. Classification from predicted probabilities

Let’s assume that one investor decides to invest 1000 euros in one firm based on the model’s recommendation. When he/she invests 1000 euros in the fast-growing firm, 1000 more euros will be awarded. But when he/she invests 1000 euros in the low-growing firm, only 500 more euros will be awarded. If one firm is predicted as a fast-growing firm, the investor will make an investment in it; but if that firm turns out to be a low-growing firm, he/ she will lose the expected 500 euros. And if one firm is predicted as a low fast-growing firm, the investor will not invest; he/she will lose the expected 1000 euros if that firm is actually a fast-growing firm. Therefore, the **loss function is FP = 1 and FN = 2 (relative ratio)**.

I use an algorithm that searches for that **optimal classification threshold** by maximizing the **cost-sensitive Youden index**. And I repeat this for all five folds, and take the average of these five estimated thresholds and the five **expected losses**. The optimal classification threshold turns out to be similar, but not exactly the same, for the six models. The logit M5 still has a similar result as the logit LASSO model. So the variation illustrates why optimal threshold selection may help find the best probability model when our goal is classification.



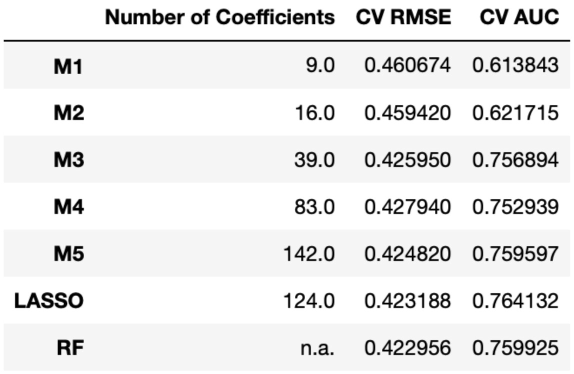
	Model	Avg of optimal thresholds	Threshold for Fold5	Avg expected loss	Expected loss for Fold5
0	M1	0.326543	0.324455	0.544156	0.546605
1	M2	0.314419	0.311073	0.541238	0.539004
2	M3	0.347091	0.344932	0.412225	0.412284
3	M4	0.354668	0.347717	0.418073	0.419663
4	M5	0.350670	0.360313	0.411582	0.409399
5	LASSO	0.347188	0.333066	0.406056	0.404350

6. Probability prediction and classification with random forest

I build a **probability random forest model** to predict probabilities and use the same variables as in model M5. But I don’t enter polynomials, add flag variables to extreme values, or winsorize values above or below certain thresholds. The main advantage of the tree-based models is that they are supposed to find good ways to approximate the most important nonlinearities and interactions.

The random forest **outperforms** the simple logit models and performs similarly to the logit LASSO model in terms of predicting probabilities.

In general, **the best model** is the **random forest model**, both for probability prediction and classification. Then it is **followed by the M5 and LASSO logit models**, which are very similar in all fit measures, with somewhat different rankings within them.



	Number of Coefficients	CV RMSE	CV AUC
M1	9.0	0.460674	0.613843
M2	16.0	0.459420	0.621715
M3	39.0	0.425950	0.756894
M4	83.0	0.427940	0.752939
M5	142.0	0.424820	0.759597
LASSO	124.0	0.423188	0.764132
RF	n.a.	0.422956	0.759925