Assignment 3: Final Report

Timur Naushirvanov

11/8/2021

Introduction

In this report, I am presenting the results of my predictive analysis of firms that can be called fast-growing (business motivation: my client wants to invest in fast-growing firms). The **dependent variable** is a binary variable showing whether a firm is fast-growing or not. After doing some label and feature engineering, I am building several models (including logit, LASSO, and random forest) and comparing their performances, in order to propose the best predictive model to my client.[1]

Sample Design

The task requires to work with data from 2010 to 2015. I also decided to focus on small and medium size enterprises (up to 50 million euro sales according to the European Commission's definition), because very large firms may have different potential for growth (and very large sale numbers - a variable that I use for constructing my dependent variable for this analysis, but more information is presented in the next section).

Label Engineering

The main dependent variable is a fast growth of firms. It can be defined in different ways: as changes in incomes, changes in current assets, total sales. I am more inclined to use a total sales indicator, because it is an economic variables that allows easier understanding of the firms' development: if it is successful and growing, then firms will sale more of its production/services. Incomes do not always allow to reflect the firms' growth: it can sell its assets/ fire workers and receive more incomes / reduce expenditures this way, but it is usually happenning not because of the company's growth.

For my analysis, I am taking a percentage difference in total sales (percentage allows to take into consideration different firms sizes). The timeframe is two years: a percentage difference in sales between 2012 and 2014, in order to reduce possible one-year fluctiations and look at a wider time range. A disadvantage of sales as a dependent variable is that many other exogenous confounders can affect it (for example, economic or political crisis leads to lower incomes and reduced purchasing power, which directly affects sales); however, let's assume that non of similar significant events happened between 2012 and 2014 in the United States (which is quite true in reality, I believe).

After imputing some NAs in the sales growth indicator and filtering those observations which cannot be imputed (14041 firms), I created a fast growth dummy as the main dependent variable of my analysis. Initially, I wanted to use a 3 quartile to define a fast growth company, which is 60. However, I believe that 40 or 50 percent increase in sales over two years can also be defined as fast. Thus, I am using a 65 percentile instead of a 3 quartile (= 35% growth in sales). It allows me to distribute data in a decently equal way (65% to 35%), and from the empirical perspective, I think that 35% growth in sales over two years can be classified as fast. Overall, 7432 firms are classified as fast-growing, 12230 as not fast-growing. The total number of observations is 102399.

Feature Engineering

I divided my independent variables in several categories.

1. Firm's Performance: "ln_amort", "amort_flag", "ln_curr_assets", "ln_fixed_assets", "ln_intang_assets", "ln_liq_assets", "ln_tang_assets", "flag_asset_problem", "ln_curr_liab", "ln_extra_exp", "ln_extra_inc", "inc_bef_tax_st", "ln_inventories", "ln_material_exp", "ln_personnel_exp", "profit_loss_year_st", "ln_subscribed_cap", "flag_error"

[1] Additional Technical Report and code can be accessed through this link.

2. Human Resources: “ceo_count”, “ceo_count_flag”, “foreign”, “female”, “inoffice_days”, “gender”, “origin”
3. Geography: “urban_m”, “region_m”
4. Fixed Effects and other relevant characteristics: “year”, firm’s industry code (modified “ind2”), imputed DV (flag)

In many cases, I use natural logarithms instead of real values, because it helps to neutralise the effect of large values and make distributions look more Gaussian. Also, some financial indicators (for example, assets) cannot have values below 0, that is why I imputed them and flagged this imputation. Furthemore, I standartised and winsorised two variables (income before taxes and profit_loss_year). All NAs (which number usually was small for all variables) were imputed. I also added some polynomials and interactions for logit and LASSO.

Model Building

I built 6 basic models.

- X1: fast_growth ~ financial_perform
- X2: fast_growth ~ financial_perform + hr
- X3: fast_growth ~ financial_perform + hr + geography
- X4: fast_growth ~ financial_perform + hr + geography + fe (used for random forest as well)
- X5: fast_growth ~ financial_perform + hr + geography + fe + polynomials
- X6: fast_growth ~ financial_perform + hr + geography + fe + polynomials + interactions1 + interactions2 (used for LASSO as well)

For probability prediction, I divided dataset into two sets (80% / 81920 obs as training and 20% / 20479 obs as holdout). Based on the models’ calculated cross-validated performance, I think it is better to choose model X6. Even though the difference between model 4 and model 6 in terms of CV RMSE not so huge, AUC is still quite significantly better. Interestingly, LASSO shows the same result for CV RMSE, however, it is much worse in CV AUC.

Model	Number of predictors	CV RMSE	CV AUC
X1	18	0.475	0.554
X2	30	0.470	0.608
X3	34	0.470	0.609
X4	52	0.466	0.623
X5	67	0.464	0.634
X6	115	0.463	0.636
LASSO	102	0.463	0.616

Analysing Performance

For classification, I need a loss function. My business problem sounds in the following way: I need to know whether a firm is going to be fast growing, because my client wants to invest only in fast growing firms. In terms of risk aversity, I know that my client is very risk averse and does not want to lose any of his money or have low incomes from these investments (he has good other options for investments). At the same time, he wants to multiply wealth, that is why not capturing successful companies will make him a bit dissapointed. Bearing this in mind, I would specify my loss function in the following way:

False Positive (predicting that the firm is fast-growing when in fact it is not) = 9 False Negative (predicting that the firm is not fast-growing when in fact it is) = 3

When calculating average of optimal thresholds, sometimes I received Inf. I think the main reason is that in such folders, all predicted probabilities were classified as Negatives (since my loss function punishes FP quite severely), that is why it was not possible to find an optimal threshold in such cases. For this task, I imputed Inf with NAs (in order to calculate averages).

Based on an average expected loss, I should also choose model X6.

Model	Avg of optimal thresholds	Threshold for fold #5	Avg expected loss	Expected loss for fold #5
X1	0.549	0.536	1.051	1.051
X2	0.598	Inf	1.052	1.052
X3	0.587	Inf	1.052	1.052
X4	0.706	0.696	1.051	1.051
X5	0.730	0.737	1.050	1.051
X6	0.721	0.663	1.049	1.048
LASSO	0.722	0.666	1.051	1.052

Comparing Performances

Finally, I also build a random forest model. While comparing performances, we can notice that RF model is better from different perspectives: it shows lower CV RMSE and expected loss together with a much higher AUC.

Model	Number of predictors	CV RMSE	CV AUC	CV threshold	CV expected Loss
Logit X1	18	0.475	0.554	0.549	1.051
Logit X6	115	0.463	0.636	0.721	1.049
Logit LASSO	102	0.463	0.616	0.722	1.051
RF probability	30	0.419	0.818	0.550	0.985

The following tables show the confusion matrix (with percentages) of Model X6 and Random Forest on a holdout set with optimal thresholds. It can be clearly seen that RF predicts much more firms as fast-growing.

Table 1: Model X6

X	X0	X1
X0	64.83	34.76
X1	0.10	0.31

Table 2: Random Forest

X	X0	X1
X0	63.09	27.04
X1	1.84	8.03

Discussion

Based on these results, my random forest model seems to be the best in probability prediction and classification. Then goes model X6 and LASSO, which are quite similar in their fit measures. The worst model is model X1. Interesing to note that the later model is not very different in terms of loss function compared to LASSO or model X6.

Since the loss function was defined as 9 in case of false positive and 3 in case of false negative, when multiplied by 1000 euros (for example), we see that the random forest model will help my client to lose less money (-985 euros per firm against - 1050 euros). At least for this reason, this model can be useful for predictions.

To conclude, this analysis helped to build a predictive model which predicts whether firms are going to be fast-growing or not. With a given loss function, random forest is the best model for predicting the probabilities and classifying fast-growing firms, while all other models (logit and LASSO) are not so much different (in terms of predicting losses).