https://github.com/pilipentseva/predictions

Predicting fast growth of firms

Goal: In my project I try to predict fast growth of the firm. I consider the growth for the firm between years 2012 and 2013. The growth I count in sales revenue. Although there are other different metrics which can be used to account for the growth, I chose sales since they show the expansion in the market (even though they don't allow to account whether this growth is profitable or not).

Data: For my project I use data set from a set of firms in a certain European country (bisnode-firms dataset) created by Bisnode company. Initial data set contains data on firms from 2005 to 2016 excluding larger companies with annual revenues above 100 million euros. The data set used in my project is a panel for 2010-2015 years cleaned and prepared by code used in *Data Analysis for Business, Economics and Policy* textbook (Ch 17).

Data Cleaning and Preparation: I discovered that there were many missing variables in the data. I completely dropped columns with many missing variables (many I consider to be more than 30 %). I replaced most of the other NA-s with the means of the columns. I took the decision not to replace missing values with 0-s for the following reason. Although some variables could be thought of having 0-s (assets, inventories), I saw in the created table of missing values that for several such columns the amount of these missing values is the same. Most likely a company in the data set couldn't exist without assets, liabilities, inventories, etc at all. So, I assumed these patterns of missing values in such columns are rather due to systematic reasons of some companies not reporting it rather than to the fact that the company could have had 0-s in these values. Some of the variables didn't make sense to replace with mean since it would confuse the information (you can't make sense of what 1.2 is ceo and top executives count number, for instance), so I replaced them with modes. At this stage I considered categorical variables (gender, origin etc) to be also part of this group and replaced them with mode.

I dropped not needed columns: nace_main and ind (industry code) since ind2 depicted almost the same information. I checked duplicates and found none. I also converted columns which needed to be numeric but were recorded as strings to numeric values.

The final cleaned data set before feature engineering contains 287829 observations.

Feature engineering and sample design: I change some of the variables to numeric. Then I proceed with changing a format of sales variable to use it to filter the firms.I create a variable to know whether the firm is new, since I expect it to affect the rate of growth a firm can achieve. I want to focus on firms which are small and medium since among big firms there may be different tendencies too different from other firms in the sample. I filter and only leave firms with sales from 1000 euros to 10m. I generate variables for the firm being alive and default to account for active firms in the industry. I make a square variable for the age of the firm and encode some variables as categorical. One of the most crucial parts in this section of work is that I account for the fact that financial variables (liabilities, assets etc can not be negative. Moreover, I follow the case study in the book and create ratios: balance

sheet variables I express as ratio to the size of balance sheet (total assets), while in the same logic I scale items related to profits and loss by total sales. I proceed by creating some categorical variables.

Then I use the winzorization - identifying for each variable a threshold variable and then replacing value outside of threshold with the threshold value and adding a flag variable. It helps to capture extreme values.

To define a target variable (y) which is a fast growth of the firm between, I limit the data to 2012 and 2013 years and group by company id. Then I calculate the growth by change in sales revenue between 2013 and 2012 divided by sales revenue in 2012. I consider growth to be fast if it is equal or higher than 10% in sales revenue (I account for the fact it should be positive growth as well). I create a dummy variable for fast growth.

Probability prediction: I take 3 simple logit models and perform probability prediction by logit. In the 1st model I include variables for sales, profit and loss. In the second I add to the 1st model variables additional financial variables and age. And 3rd model contains the most variables. All variables in these models I select manually and based on domain knowledge. Moreover, I include one more model (the 4th one) with logit LASSO, where predictors are chosen by the LASSO algorithm. I allow LASSO to make a choice from predictors I hand-picked for the 3rd model. It considers some less important, so compared to 3th has 20 predictors instead of 24. I start by exploring these models through marginal differences and coefficients. Some of the variables are less correlated with a firm's fast growth, some more (for example, age is shown to be correlated which makes sense) some less.

Then I carry out a probability prediction/seletion by logit and 1 LASSO:

Variable/Name	N of predictors	RMSE
X1	4	0.3526
X2	10	0.3519
X3	24	0.3384
X4 (LASSO)	20	0.3225

I can see that RMSE is quite close to each other. However, the model with LASSO has the smallest one, so I would choose model number 4.

Loss function and classification: To determine a loss function, I need to define a cost of False Positive (FP) and False Negative Classification. I can imagine a situation where investors or angels are investing their money in the firm. Then the “positive” event is that the firm grows fast and “negative” if it grows slow. Investors tend to invest if my analysis tells them the positive event is going to happen and tend to restrain from investments if negative is going to happen. My assumption is that in the case of FP when investors invested, but the company grows slowly investors lose their money - 10 thousand euro. And in the case of FN

when they didn't invest and lost money they could have gained, the loss is 2 thousand euro. Then the cost would be FN/FP. I also run an algorithm to find an optimal threshold with 5-fold cross-validation. The table with optimal thresholds is shown below. The best classification can be defined by the model and its corresponding optimal threshold that produces the less expected loss. In this case it is the LASSO model and its optimal threshold. However, X3 gives almost the same average expected loss. For convenience, in later estimation I used X3 instead of LASSO. Overall, it turned out that the model with the smallest RMSE is also the model with the smallest average expected loss.

	Model	Avg of optimal thresholds	Threshold for Fold5	Avg expected loss	Expected loss for Fold5
0	X1	1.655565	1.869946	0.363994	0.364399
1	X2	1.475819	1.875989	0.363821	0.364399
2	X3	1.388289	0.765278	0.363763	0.363531
3	LASSO	1.036614	0.830466	0.363647	0.364110

One of the biggest concerns is the confusion table I got in the end. The result does not seem to be realistic to believe the model predictions.

	Predicted no FG	Predicted FG
Actul no fast growth	7057	0
Actual fast growth	1580	0