Assignment 2 - Fast growing firms prediction

Tamas Stahl

11/02/2021

Aim of the analysis

As our second assignment for Data analysis 3 we had to build models to predict fast growth of firms using the bisnode-firms data. At least three different models were to be built and then the best chosen from those. The result of this analysis could be financially beneficial as by using the correct variables we could predict the companies that will grow fast, therefore, investment opportunities might rise.

Data

Our assignment was based on the case study presented to us in chapter 17, however, our target variable changed from whether the company would default in a given time frame to whether the company will grow fast.

The project uses the bisnode-firms dataset, which was collected, maintained, and cleaned by Bisnode, a major European business information company. The authors of the mentioned book cleaned and combined this dataset into one single work file which is an xt panel at the company-year level. The dataset contained information about firms between 2005 and 2016, from which I excluded the observations for 2016 as they had many missing values.

Data preparation

We were provided with an R code for data preparation and prediction as well on class. In the data preparation I used the same variables, interactions, marker flags that we used in class. In my opinion those variables are useful in the question of fast growth of firms, not only for the defaulting of firms.

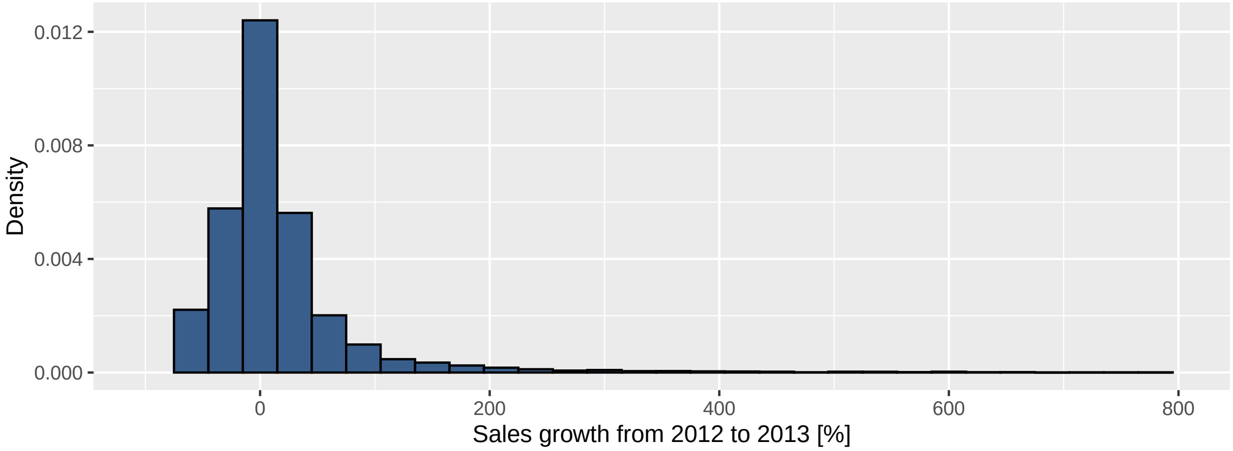
My task was to predict the probability of fast growth companies. For which we needed a target variable to use. First of all, for determining the growth rate of company I used the well-known compound annual growth rate (CAGR) formula. CAGR is one of the most accurate ways to calculate and determine returns for anything that can rise or fall in value over time.

So after calculating the CAGR I determined 30% as fast growth rate for a year, meaning that the 2013 sales figures exceeded the 2012 sales figures by at least 30%. I chose to look at only the period between 2012 and 2013, although, it might be beneficial to test our model for two or maybe three years as well as the pattern of growth might be captured differently. **But for the sake of our task I chose 2013 vs 2012 and established 30% and above as the fast growth rate.**

After establishing the growth rate I created a variable called fast_growing which is a binary variable which determines whether the CAGR is above our 30% rate or below.

As mentioned above I included the year 2012 and the companies that are still alive, meaning that their sales is above 0. On top of that I determined that only those companies matter for my analysis that have a **sales number for 2012 between 10 million euros and 10 thousand euros.** In class we used 1,000 as the minimum value for the sales. However, from an economic point of view I increased that limit to 10 thousand euros as 1,000 euros are way too low for an annual sales number.

The following histogram shows the sales growth percentage (in other names CAGR) from 2012 to 2013. We could see that the distribution is skewed right and that there are significantly large number of negative values, meaning the companies declined in the 2012-2013 period from a sales point of view.



Finally, my work sample consists of 12,556 observations (2,442 are fast growing) and my holdout set consists of 3,138 companies (610 are fast growing).

Model building

As part of the analysis I built 7 different models:

1. Model logit X1: Handpicked
2. Model logit X2: Model logit X1 + "Firm" predictor variables
3. Model logit X3: "Firm" + "Financial 1" + "Growth" predictor variables
4. Model logit X4: Model logit X3 + "Financial 2" + "HR" + "Data quality" predictor variables
5. Model logit X5: Model logit X4 + Interactions
6. Logit LASSO
7. Random forest

As mentioned above I created variables and interactions for model 5 and LASSO, used a method called winsorization, in which method I identified threshold values for each variable and replaced the values outside the threshold with the threshold value itself and added a flag variable. "Financial 2" group of variables include all the flags mentioned.

For the random forest model I used the same variables as in model Logit X4, except I did not do any feature engineering, meaning that there are no polynomials, flags for extreme values or winsorized values. I used the same method for random forest just like in class, meaning that the default option of growing 500 trees, simple tuning with 15 observations as the minimum per terminal node, 5 variables for each split and 5-fold cross-validation were used.

Finally, for training the models I used 5 fold cross-validation on the train data set with the same folds.

Propability prediction

As per Table 1 we could conclude that the random forest model performs the best with the lowest RMSE (0.381) and the highest AUC (0.675). However, both models Logit X4 and X5 are comparably performing the same with a bit higher RMSE (0.384) and bit lower AUC (0.669).

In order to interpret the results a black box model like random forest would not be beneficial. But purely from a probability prediction point of view that is the best performing model. During my analysis I used Logit X4 and random forest.

Table 1: RMSE and AUC for models

	Number of predictors	CV RMSE	CV AUC
Logit X1	11	0.393	0.595
Logit X2	18	0.389	0.635
Logit X3	35	0.385	0.663
Logit X4	79	0.384	0.669
Logit X5	153	0.384	0.669
Logit LASSO	106	0.383	0.646
RF probability	36	0.381	0.675

Classification

For classification I am determining a loss function and finding the optimal threshold that gets the lowest expected loss. In our case the false negative case is when we predicted a slow growth but it actually grew fast. False positive case is when I predicted fast growth but eventually it grew slowly.

In our case the more costly mistake would be the false negative, as we would loose out on profit from investment into a fast growing company. As defined at the beginning of the analysis, fast groth would mean at least 30% sales increase from 2012 to 2013. Therefore, the penalty for the false negative scenario (FN) would be 3. Meanwhile, the false positive (FP) penalty would be 1 as in case of false positive scenario, we would not loose money, only our firm will not earn as much profit. **We are penalizing the missed opportunity more compared to the not so good taken opportunity, with FN=3 and FP=1.**

According to the learned formula the optimal classification threshold would be around 1/4 = 0.250. To find the optimal threshold I also used a search algorithm, where I considered the random forest model, its predicted probabilities and the loss function detailed above. I ran the algorithm on the work set with 5-fold cross-validation, which resulted in 0.28 as the optimal threshold and 0.49 as the expected loss.

Summary

To evaluate the performance of our model (random forest) I took the holdout set and looked at the classification results in the appropriate confusion matrix.

This prediction could be beneficial if we were to help investment teams by predicting the probability of fast growing firms. Let's say that we have 10 thousand euros to invest.

With the help of the determined loss function of FN=3 and FP=1, we could calculate the expected loss. We say that we penalize false native more than false positive, as I would say that the opportunity not taken is more costly than an opportunity taken but without the anticipated high profits.

As a result of false negative decision we would miss out on profitable investments. So the 0.492 expected loss would translate to 492 euros loss per classification on the live data. Table 2 shows the confusion matrix for random forest classification, with thw aisles being our prediction and the columns being the actual values.

Table 2: Confusion matrix for random forest classification

	no_fast_growing	fast_growing
no_fast_growing	2161	404
fast_growing	367	206

Table 3 shows us the summary of the model performance measures. The best model is the random forest model for classification as well as probability prediction. All measures below come from 5-fold cross-validation on the work set.

The cross-validated expected loss (0.492) was smaller than any other models, however, the difference is marginal as model logit X5 has an expected loss of 0.493. Furthermore, random forest model performs the best with the lowest RMSE (0.381) and the highest AUC (0.675).

We can see that the range of the expected loss is between 0.558 (model logit X1) and 0.492 (random forest). If we count that we miss out on 1,000 profitable investments due to the selected model, we could loose 66,000 euros if we use the worst model instead of the best one.

Table 3: Summary of model performance measure

	Number of predictors	CV RMSE	CV AUC	CV threshold	Expected Loss
Logit X1	11	0.393	0.595	0.223	0.558
Logit X2	18	0.389	0.635	0.245	0.521
Logit X3	35	0.385	0.663	0.259	0.496
Logit X4	79	0.384	0.669	0.270	0.498
Logit X5	153	0.384	0.669	0.255	0.493
Logit LASSO	106	0.383	0.646	0.225	0.518
RF probability	36	0.381	0.675	0.289	0.492

External validity

External validity is always an interesting point to include. The starting point of the analysis was a single cross-selection. I took financial variable values for 2012 and predicted the fast growth of companies in a year. We could check external validity easily if we just use our model for a different period, such as 2013 vs 2014. As an educated guess I would say that the two periods should not differ significantly, but if anybody is interested the variables could be changed in the code called "DA3_assignment2_fast growth_prepare.R" found in the GitHub repo of this assignment.

Furthermore, I would encourage anybody to come up with determining fast growth better or argue for other loss function.

Conclusion

To predict the probability of fast growing firms I used a complex data set and 7 different models. After that I specified a decision situation with losses due the the mistakes of our model or bad decisions.

First of all, we could conclude that three models performed quite similar, namely logit X4, X5 and random forest. This is good for us, as if we need any interpretation we could easily use the logit models instead of the black-box random forest.

Furthermore, it is important to highlight that one of the most important point of our prediction is the determination of the loss function. To have a meaningful prediction we would need a good loss function, for that we need domain knowledge as well as expertise in the field. **I would like to highlight it again, that the decision makers need to decide on these points at the beginning of a project.**

Another important part of the prediction was the target variable classification. Some could argue that I did not use the best formula for the fast growth of a company. With resources and expertise it could be further improved.

To sum it up, I found that random forest preformed the best from the seven models with the best prediction of the probability of fast growing firms and classifying firms that are likely to grow fast.