Data Analysis 3 – Machine learning

Assignment 3

Gabriella Zsiros

Introduction

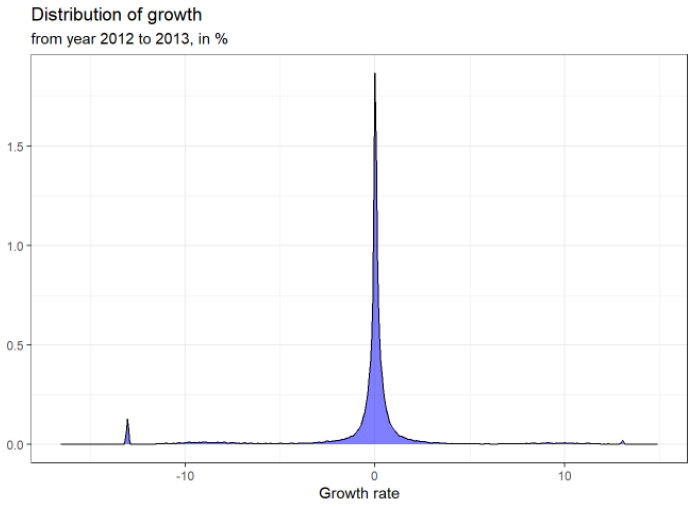
The target is to predict the growth of firms in `cs_bisnode_panel.csv` dataset, which contains several variables about firms, including balance sheet, income statement, HQ and CEO information.

In this exercise, the data is taken from 2010 to 2015. During feature engineering, binary variables were factorized, and variables, where are more than 10% is missing, are dropped.

In the remaining data, missing variables are imputed through the mean of numerical values, and the mode of binary variables. Sales data had been transformed to natural logarithmic, as well as quadratic terms and it was expressed in millions, instead of single dollars.

Income statement and balance sheet variables are grouped and normalized by sales and total asset value, variables without variance are also flagged, then eliminated.

A summary of sales data shows the 95th percentile around 1.1 million, while the 99th and max are 10 and 22 million dollars respectively. Considering that sales data can be a significant indicator in the target variable, fast growth, 99% of the data will be kept. Quadratic term is also added to the variables.



Sales growth shows a normal distribution, with 0 average.

Based on distribution of the sales growth data, growth rate above 35% will be considered as fast growth

PART I: Probability prediction

Five combinations of variables are considered, each being more complex than the previous one. The starting point is simple financial data. The second level includes quadratic term, as well as more details about the firm.

The third level add balance sheet and income statement variables. The fourth level includes variables of the CEO, and the fifth is completed by interaction terms.

The dataset is split into a holdout (20%, n = 5207) and training (80%, n = 20830) subset.

Logit probability prediction is calculated and cross-validated through 5 folds.

Linear

The logit probability model is run across 5 folds, testing each variables combination X1 through X5

By calculating the average ROC across folds, model X4 has the highest score: 0.72.

- *Model: X1, Mean CV ROC: 0.628148785116198*
- *Model: X2, Mean CV ROC: 0.701091407330525*
- *Model: X3, Mean CV ROC: 0.728170486641647*
- *Model: X4, Mean CV ROC: 0.728308887066775*
- *Model: X5, Mean CV ROC: 0.727309995552693*

Lasso

For Lasso Parameters, the lambda was set as 10-based logarithmic set of parameters, from -1.

Comparing their means ROC, the fourth variable model and LASSO has similarly good value, with LASSO performing stronger, with average ROC value of 0.7293.

- *Model: X4, Mean CV ROC: 0.728308887066775*
- *Model: LASSO, Mean CV ROC: 0.728232170099543*

Random Forest

With Random Forest turning, the model iterates through 6, 8 and 10 tree numbers, which results in 6 variables being the most optimal set-up.

AUC values are then calculated with ROC function.

While the number of predictors is increasing with model complexity, LASSO has better ROC values with fewer predictor than the two most complicated linear variables model. This doesn't hold, however when examining the AUC, where the fourth model has quite good score.

Random Forest perform the best in both metrics, with 0.804 ROC and 0.804 AUC scores.

	Number of predictors	ROC	AUC
X1	4	0.6281488	0.6281488
X2	10	0.7010914	0.7010914
X3	23	0.7281705	0.7281705
X4	31	0.7283089	0.7283089
X5	40	0.7273100	0.7273100
LASSO	29	0.7282322	0.7074976
RF	40	0.8057376	0.8047589

The best logit result without considering a loss function is the random forest model. When testing on the holdout dataset, it has a very good RMSE value: 0.3611022.

PART II: Classification

Loss function

When defining the loss function, it must be considered, that false positive mean we might invest into a firm which will not have large ROI, but this also has an opportunity cost of not investing that money to a firm that would have been growing fast. False negative means that we missed out on an opportunity to invest, Therefore, the ratio is determined as 2:1 (FP/FN)

Classification threshold & Average expected loss

During classification, the most optimal threshold is determined through 5-fold cross validation.

Two factors are playing an important role: the prevalence and the cost. Prevalence shows the proportion of the positive (TP+FP) in the dataset, while the cost reflects the ratio of False Positive and False Negative. In this case, the Random Forest model determines the threshold at 0.569, very similar to the first linear model.

It is visible that the threshold is much higher if we use more complex linear models. Random Forest gives similar result to the first, simple variable model. The average expected loss is produced with random forest, with 0.193, while the rest of the models perform around 0.21

	X1	X2	X3	X4	X5	LASSO	RFP
Threshold	0.5691508	0.7128088	0.7797533	0.7819325	0.7811263	0.7401033	0.5692295
Expected loss	0.2120236	0.2118314	0.2119756	0.2119756	0.2119756	0.2119996	0.1934259

PART III Discussion of results

Evaluation

Testing the models on the holdout set, the best optimal threshold is set by random forest at 0. 569, since this results in the lowest expected loss. Using this, the model proves to be a good fit on the holdout set, the expected loss is 0.203.

Confusion table

	Reference	
Prediction	not_fast_growth	fast_growth
not_fast_growth	3984	852
fast_growth	103	263

Calibrated by the random forest model, the confusion table shows 4071 + 233 correctly predicted Positives and negatives, there the classification of fast-growing company was properly determined. 839 + 64 are false, with False Positive being a considerably lower value. This corresponds to the pre-determined loss function which penalizes the false positives more.