Assignment 3 - Predicting firms with high investment potential

Ádám Vig & Szilvia Flanek
09 November 2021

In this assignment we address the following business problem: Based on available firm level data, **which firm should we invest for a one year period**? We take the Bisnode database and use a 2010 and 2011 cross section to predict which firms have the highest earnings potential in 2012. We build 7 models, evaluate their predictive performance and choose the best one based on the minimum average loss induced by our loss function. We test this model on "live data" by calculating the profit we would make in 2015 by investing in 2014, based on the model's prediction. All code can be found in this Github repository.

Sample design

The baseline sample contains 33,359 firms, and the "live data" contains 32,365 firms. We assumed that we would only invest in existing firms, so we restricted our sample to firms that still existed at the time of the hypothetical investment. First we dropped firms that did not report in 2010 or in 2011 (2013 and 2014 in "live data") based on the fact that the beginning of balance sheet value was missing. For similar reasons we dropped firms that stopped reporting before the end of the year – their end of balance sheet value is earlier than the 31th of December. In addition, we dropped firms that had zero (or missing) sales in 2010 or in 2011 (2013 and 2014 in "live data"), because we do not consider these firms alive. Therefore our baseline sample has 19,147 firms and the "live data" has 19,740 firms.

Choosing a target variable

The high earnings potential of firms can be defined in many ways. We might want to invest in firms and then sell our shares a year later, so we look for growth in size, which can be measured by the change in the company's total sales. The other option is to look for firms with high dividend potential, which can be measured by the return on equity (ROE). In both cases we invest for one year, thus our target variable is the observation of sales growth or ROE in 2012. First we looked at models with sales growth as a target variable. We define sales growth as the difference in the logarithm of sales in 2012 and in 2011. We say that a firm grew fast if the firm's sales grew by more than 30% in one year. With this specification, around 14% of firms had fast growth in 2012. We first created models with sales growth as target variable, however, these models failed to predict firms that had extremely high growth. A possible explanation could be that firms that have extremely high growth in size in one year might be very similar to firms that have extremely low growth. These are the high risk - high return firms and the models fail to differentiate between them.
Therefore we turn to our second choice of target variable, which is the firms' return on equity. We calculate this as the share of the company's income before tax and shareholder's equity. To get reasonable results, we impute this variable with the minimum value in the given year if either both the numerator and the denominator are negative or if the denominator is zero. We want to invest in firms with high earnings potential, which we define as higher than 30% ROE. With this specification, around 16% of firms have high earning potential in 2012.

Feature engineering

We pooled the industry category codes into bigger groups, created additional firm characteristics like squared age, foreign management as dummy, gender of manager and region as category. Since we are trying to predict growth, we created some additional financial variables like total assets and ratios (variables as the ratio of sales or total assets). We flagged firms that seem to have an asset problem such as one of the asset variables is either zero or missing. We imputed asset variables with zero where it was negative. We flagged and winsorized the tails of any other accounting features that cannot be negative if they took up a negative value or that should be between -1 and 1 if they are outside of this range. We created a variable for the age of the CEO and flagged the firm if the CEO is less than 25 years old or more than 75 years old. We also added a dummy variable for young CEO-s (less than 40 years old). We imputed the number of employees feature with mean values if missing and flagged the imputed observations. We dropped all observations where a key variable (liquid assets, foreign CEO, industry, age, material expenditure, and region) is missing. We also use the sales of the company as a feature when predicting the ROE.

To improve the prediction we added some more financial variables. The EBIDTA as the ratio of the sum of income before taxes and amortization by the total assets, the ROA as the ratio of income before tax and the total assets, the ratio of sales and total assets, the liquidity rate as the ratio of current assets and current liabilities, and finally the own equity rate as the ratio of the shareholder equity and the total assets. We imputed these variables with means, minimum, maximum and median values to correct for unreasonable values.

We only keep the observation for 2011, which also includes some features that were calculated based on the previous year's values. We drop firms with less than 1000 euro or more than 10 million euro revenues, since these are either too small or too big firms for our scope now. Finally, we only use transformed features in our logit and LASSO logit models, and use the initial features in the Random Forest model.

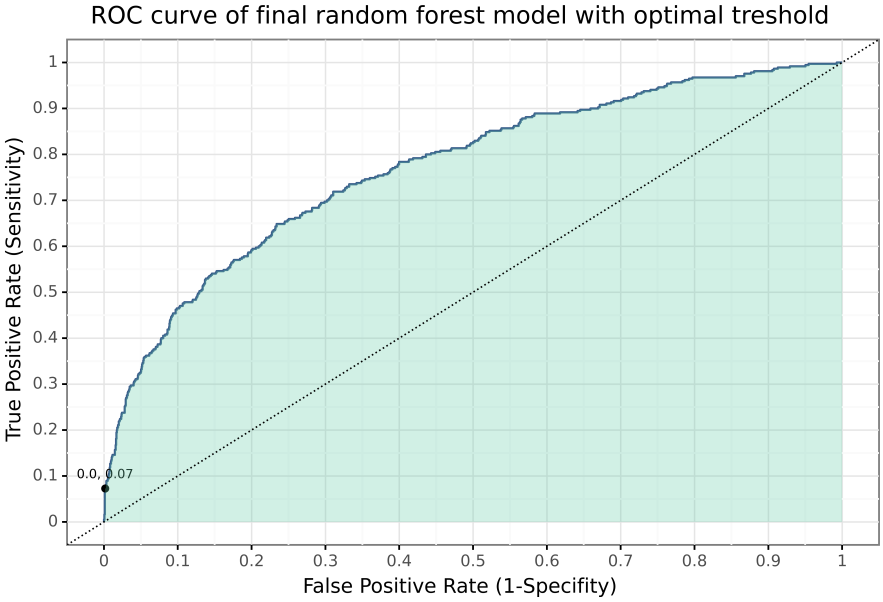
Probability prediction

We used a 70-30 train test split on the baseline dataset and trained 5 logit models (with different features), a Lasso logit model, and a Random Forest model on a set of reasonable hyperparameters. We calculated the 5-fold cross-validated RMSE and AUC to compare model fit. Based on these, the best model on the work set is the Random Forest with 0.336 RMSE and 0.778 AUC. We evaluated the Random Forest model on the hold-out set, which gave us a 0.333 RMSE and a 0.788 AUC which suggest that we avioded overfitting. The ROC curve is far from the 45 degree line (as suggested by the high AUC values) and the calibration curve is close to the 45 degree line, which shows that the model is well calibrated.

Loss function and classification

On the one hand, if we invest in a firm that do not have a high ROE, we might be loosing a lot of money. On the other hand, if we do not invest in a firm with high ROE we just lose opportunity cost. Here, we want to avoid false positives a lot more than false negatives, thus we chose the ratio of the cost of false positive and false negative to be 2.

Based on the above defined loss function we calculate the expected loss with the optimal cutoff formula $\frac{2}{3}$ and a cutoff search algorithm. The cross-validated expected loss is the lowest (0.153) for the Random



Forest Model, here the cutoff search results in a 0.533 threshold. Our best logit, the LASSO logit and the Random Forest performances are displayed in the following table.

Model	N features	CV RMSE	CV AUC	CV treshold	CV loss
logit	80	0.342	0.751	0.574	0.156
logit Lasso	86	0.430	0.757	0.610	0.154
Random Forest	43	0.336	0.778	0.533	0.153

After re-estimating the best model on the whole train dataset, we calculate the final confusion-matrix on the holdout set. Based on our model we would invest in around 3% of the firms, and we would get a false positive for less than 1 percent of the firms, while we would get a false negative for 14% percent of the firms. This comes from twice as large a loss on false positives than the loss of false negatives. The Random Forest model generates an expected loss of 0.155 on the hold-out set.

	Predict low ROE	Predict high ROE
Actual low ROE	4033	36
Actual high ROE	685	98

Making profits on ”live” data

Finally, we take our best model and test it on the “live data” to see how our investment portfolio of firms would perform. Based on the best model’s classification in 2014 we would invest evenly in 521 firms (2.6% of ”live data”) , with 78% of them actually having more than 30% ROE in 2015. **The average ROE of the firms we invest is 64%**. This results in an expected maximum profit of 164,262$ in 2015 if we invested 100,000$ in the portfolio in 2014 – the actual payoff depends on the firms’ dividend policy.

Comparing manufacturing and services

Separating the sample

The manufacturing industry firms are under the industry code 1 and 2 (5,012 firms in total), while the service industry firms are under the industry code 3 (11,161 firms in total). The share of high ROE firms is 12,5% for the manufacturing industry and 24% for the services industry. We used the same features and label and predicted the same models as above with these restricted samples.

Probability prediction

The best fitted model is again, the Random Forest model for both the Manufacturing and the Services industry. In the Manufacturing sample it produces a fit with 0.382 cross-validated RMSE and 0.785 AUC on the train set and 0.385 cross-validated RMSE and 0.769 AUC on the holdout set. In the Services sample it has a 0.312 cross-validated RMSE and 0.744 AUC on the train set and 0.314 cross-validated RMSE and 0.76 AUC on the holdout set. For both samples the ROC curve is far from the 45 degree line (as suggested by the high AUC values) and the calibration curve is close to the 45 degree line, which shows that the model is well calibrated.

Classification

The cross-validated expected loss is the lowest (0.218) for the Random Forest Model for the Manufacturing sample, while in the Services industry all of the models have very similar expected loss (0.122 for the Lasso Logit and Random Forest and 0.123 for all the others). For both samples the model chooses very similar optimal threshold values: 0.51 for Manufacturing and 0.57 for Services. Confusion matrices for the holdout sets are in the Technical Report, here we report results in "live data" in the next section.

Profit

In the live data average ROE of the predicted firms is higher for the Manufacturing sample (66.6%) than for the Services sample (42.8%). This means that if we invested 100,000$ in 2014 we would have a maximum of 166,588$ in the Manufacturing industry and 142787$ in the Services industry, but the actual payoff depends on the firms' dividend policy.
The following table displays **confusion matrices in "live data"**. We bolded the number of firms we would have invested in. The model definitely performs worse in Services, we only would have invested in 130 firms, loosing 2181 that actually had greater than 30% ROE-s. On the other hand we perform quite great in Manufacturing: The model predicts that we should invested in 674 firms which would have resulted in 637 firms that actually had high ROE-s in 2015 and we would have lost only 566 firms with high ROE.

	Services		Manufacturing	
	Predict low ROE	Predict high ROE	Predict low ROE	Predict high ROE
Actual low ROE	10021	**42**	3772	**37**
Actual high ROE	2181	**88**	566	**637**