ASSIGNMENT 3

Business problem: As an investor bank, the client wants to know which firms have high growth potential to consider them as an investment option.

I build 5 linear regression models, one LASSO and one Random Forest model to predict which firms will be a high-growth firm in 2014. The base year is 2013. I did not take 2-year growth prediction as it will not as accurate as a shorter term prediction. The best model selection is based on the minimum loss that the loss function predicts.

Data preparation. We start with loading the bisnode panel data. I keep the data for years 2012-2015 to use for predictions. Clean the dataset from the firms that went bankrupt in the potential investment period and has not started half-way through. Label engineering, sample design, feature engineering. Variable groups.

Target variable. As a measure of growth, I use sales and return on equity because these two variables are well-known and trusted metrics used by investors and institutions to decide between competing investment options. Return on equity (ROE) measures a corporation's profitability in relation to stockholders' equity. Whereas sales growth will contribute to the company expansion, so it will be easy to get rid of our stocks in case we have second thoughts about our investment choice.

The threshold to be considered as high growing firm is 30% annually. It might be pretty high, but as we are talking about rapid growth, 30% seems to be reasonable. On average, 20% of 13647 firms grew fast in 2013. The observations which had negative value or the denominator was 0 were imputed with the minimum value.

Feature engineering. This step was quite similar to what we did in class: combined industry category codes into bigger groups, created additional firm characteristics (squared age, foreign management as dummy, gender of manager and region as category). Also, additional financial variables like total assets and ratios were created. Zero or missing values were flagged, negative values were imputed with zero where having negative value did not make sense.

One important note in feature engineering when predicting ROE is that sales variable was used as a feature. Additional financial variables like EBITDA, ROA, liquidity rate and others were created to improve the prediction performance of our models.

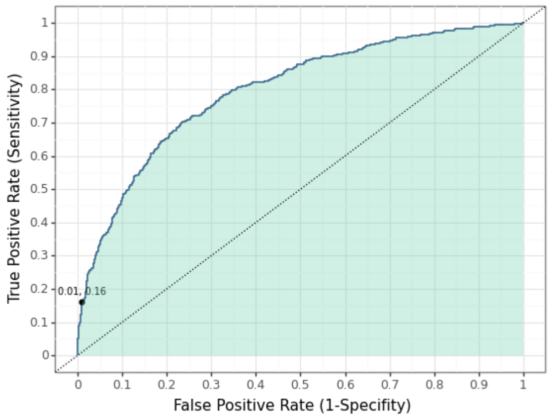
The focus of the analysis were the SME firms whose revenue was within the range of 1000-10 million.

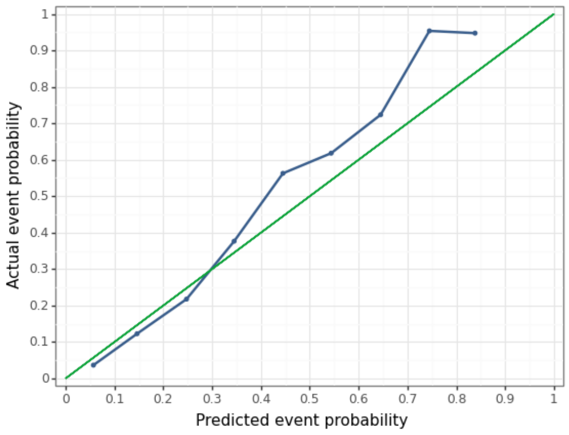
Probability prediction. Train and test split was made at 70x30 ratio. Seven models mentioned above were employed to predict. I calculated 5-fold cross validated RMSEs, AUCs to compare the models and choose the best. As usual, RF model performed best compared to the logit and LASSO models when predicting ROE growth, but performed similar or slightly worse than simple logit model at predicting sales growth. The reason why all models performed badly at sales growth prediction is that some firms grew at an incredibly rate.

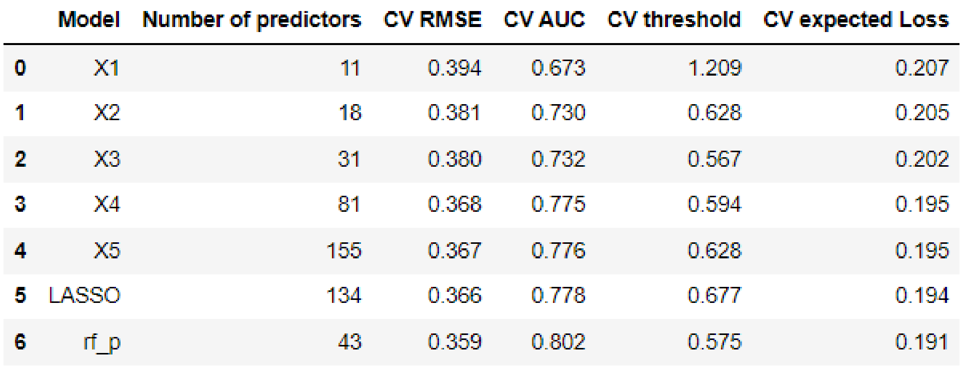
Loss Function. We need a loss function to find the balance between taking investment risk and missing the opportunity to invest in a successful firm. Since loss in case of firm bankruptcy is much higher than the opportunity cost, I take the ratio of 2 (FP/FN) because I would want to avoid false positives as much as possible.

RF model has the lowest cross-validated expected loss of 0.191. CV threshold is 0.575. The table shows that the best logit model X4 and LASSO models use much more features but perform relatively worse.

The ROC, calibration curves and model comparison tables of ROE prediction are below:

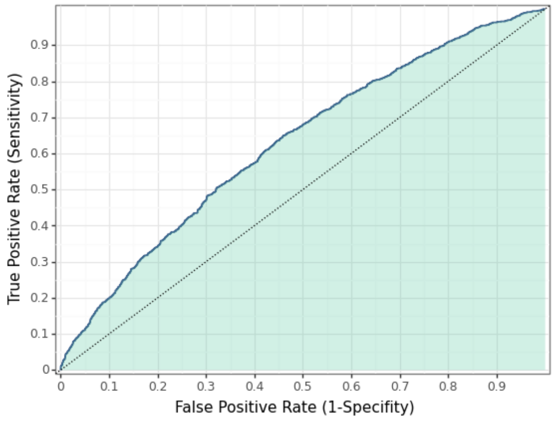


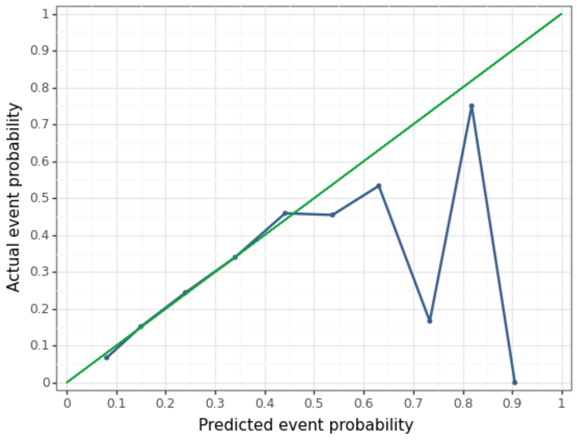




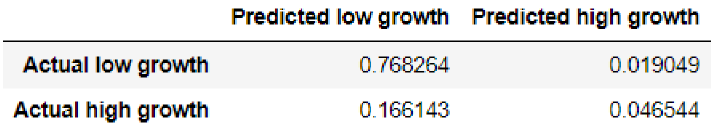
	Model	Number of predictors	CV RMSE	CV AUC	CV threshold	CV expected Loss
0	X1	11	0.394	0.673	1.209	0.207
1	X2	18	0.381	0.730	0.628	0.205
2	X3	31	0.380	0.732	0.567	0.202
3	X4	81	0.368	0.775	0.594	0.195
4	X5	155	0.367	0.776	0.628	0.195
5	LASSO	134	0.366	0.778	0.677	0.194
6	rf_p	43	0.359	0.802	0.575	0.191

The following graphs can show how badly the sales prediction models were performing because of too high and too low growth firms which are difficult to predict:





Therefore, I recommend relying on ROE prediction results to make an investment decision in this specific case. After selecting the best model (RF), I refit it on the whole train set and estimate its performance on the holdout set. It gave me the following confusion matrix:

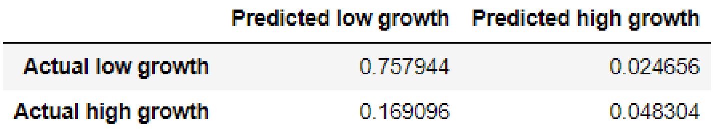


	Predicted low growth	Predicted high growth
Actual low growth	0.768264	0.019049
Actual high growth	0.166143	0.046544

According to the confusion matrix, the investors would invest funds in 4.6% of the firms who will have high growth and invest in 1.9% of the false positive low-growth firms.

Expected loss on the holdout set is 0.2.

Eventually, I tried to predict the profits on the live data, 2014, to give the investors tangible results (“If you invest this much, you would get that much”). Prediction performance is pretty similar on the live data:



	Predicted low growth	Predicted high growth
Actual low growth	0.757944	0.024656
Actual high growth	0.169096	0.048304

Based on the confusion matrix, we would invest in 1086 firms and receive the profit of 144 728 EUR in one year if we invest 100 000 EUR in 2013. Easy cash 😊