Assignment 3 - Technical Report Fast Sales Growth

A. BOGNAR

Introduction

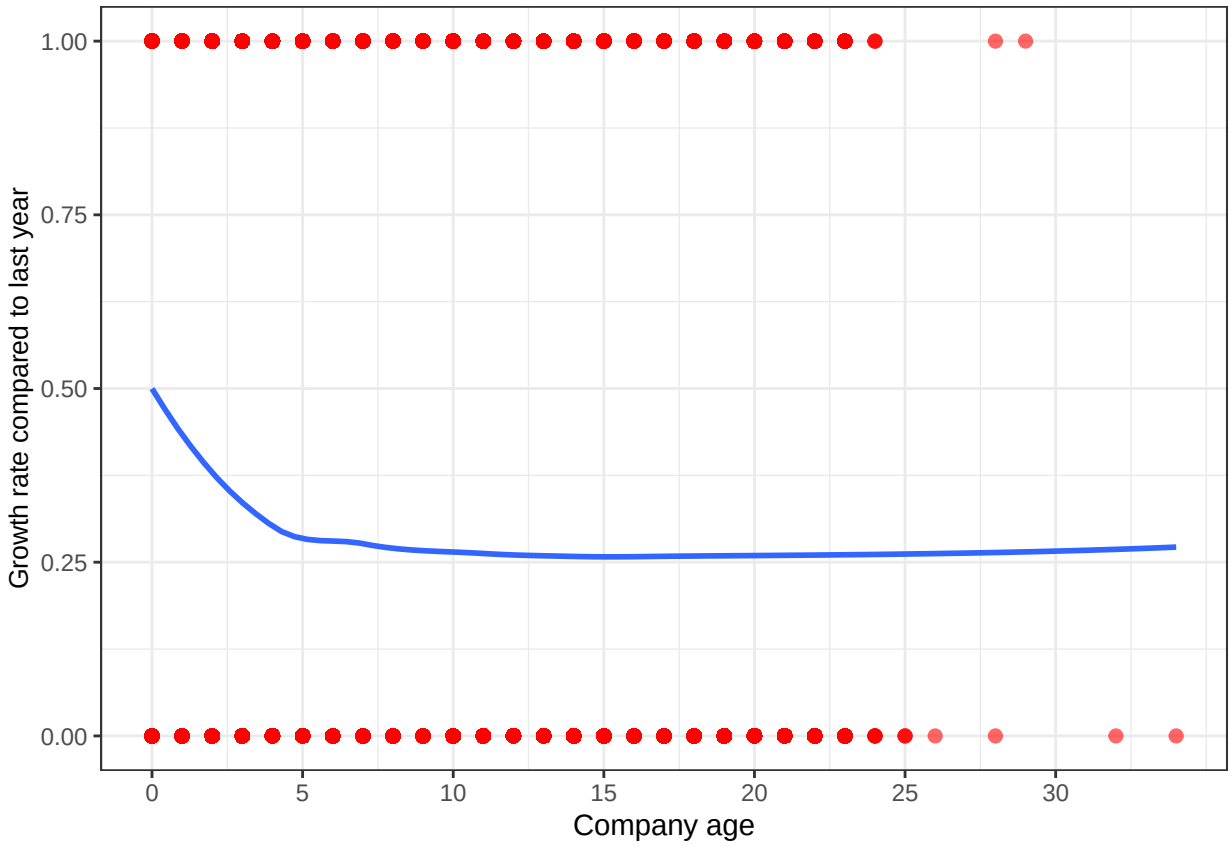
This is a technical paper to *Predicting Companies' Fast Sales Growth* explaining some of the decisions in the analysis and providing further background on the findings.

EDA/Feature engineering

Features, sample design largely followed the decisions made in the original case study. With the significant difference of constructing the target variable. For that:

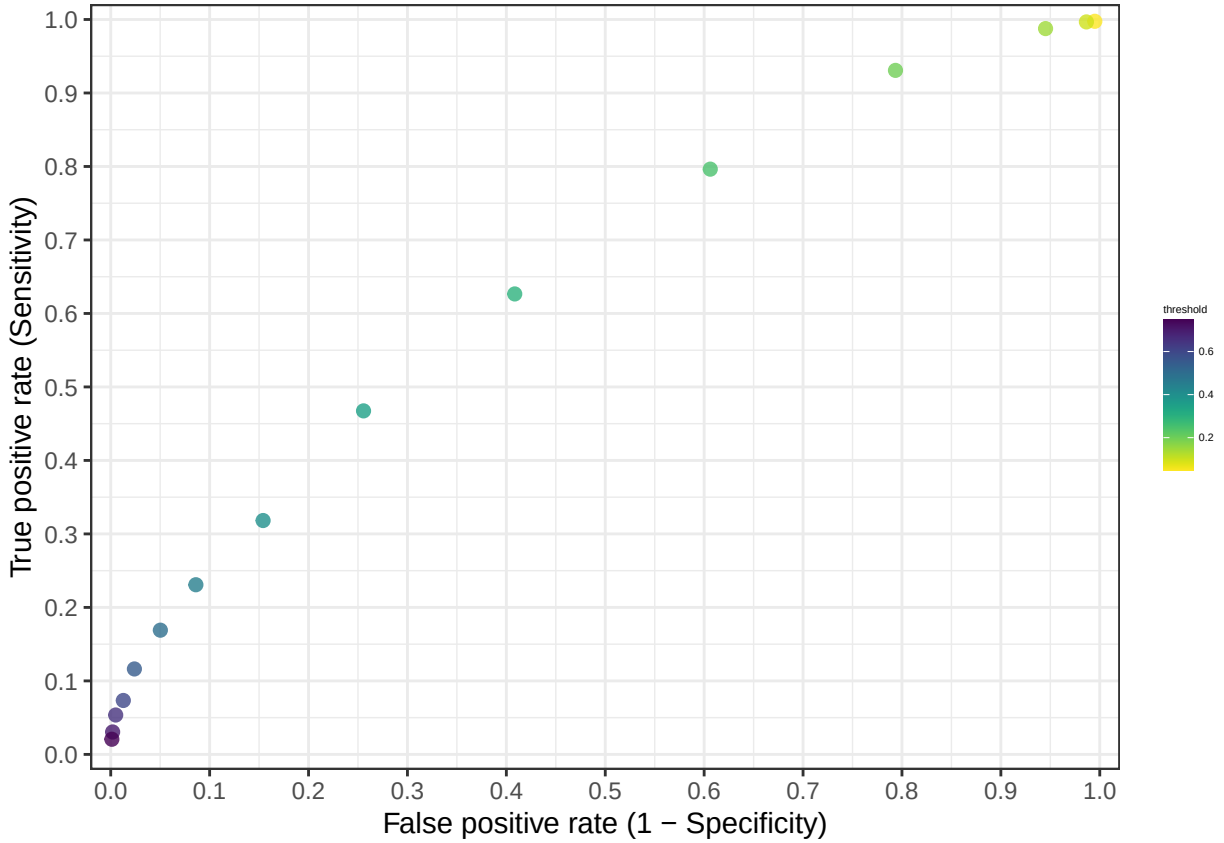
- First, during sample design, gather the company sales in the 2 years prior and 2 years following the year of analysis (2012). Specifically code lines 86-92 in the cleaning file.
- Based on that, calculate yearly changes. Prior years sales' used as predictors, change to upcoming year (2013) from current year (2012) used as target variable.

Of the many predictors listed, I would highlight below the age-sales growth relation. We can clearly see that rapid growth tends to take place in companies younger than 5 years. Afterwards, the chance of rapid growth is stagnant.

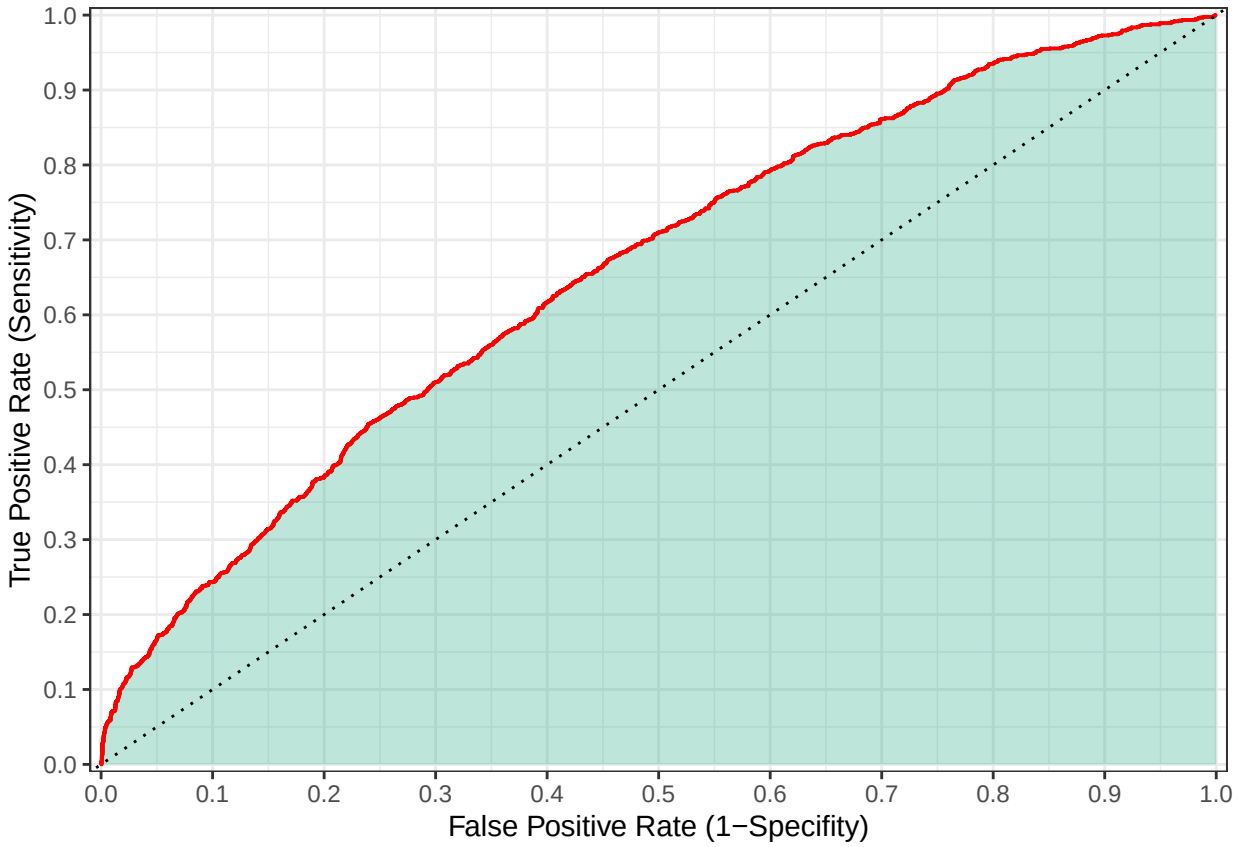


Analysis background

The ROC curve shows that threshold needs to be relatively high to avoid false positives but in turn few real positive results are found as seen in the main analysis.



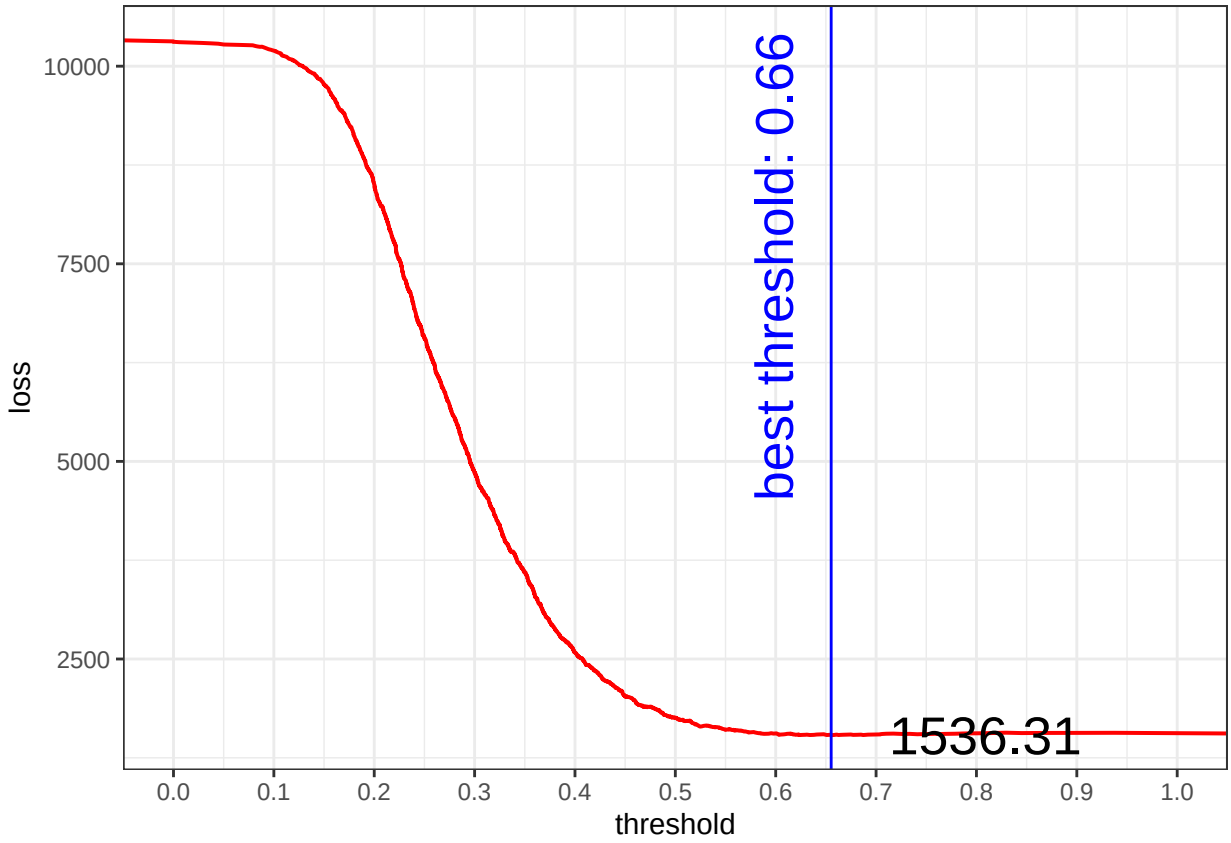
The Area Under Curve Plot of logit Model 4 shows that the avg. predictions are significantly better than random choice would be, but still far from perfect decision as the confusion matrix will later show.

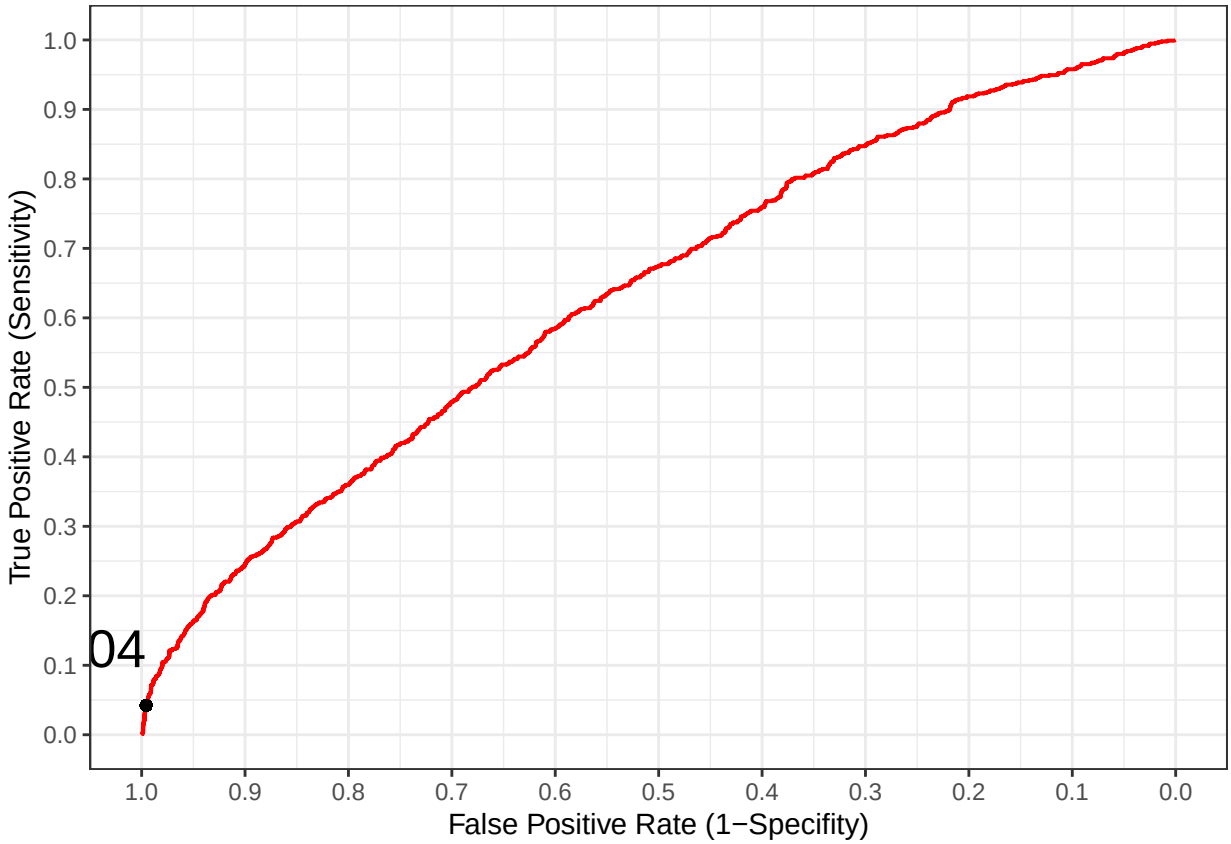


The summary outcome table of logit and Lasso models is very interesting. Because we set the bar for False Positive very high, less complex models provide an Infinite level of threshold. Meaning, the best prediction they can make is to avoid betting on any firm to grow to minimize losses. More complex models with more accurate predictions still provide a high threshold that allows for some companies to be classified as fast growing.

	Avg.of.optimal.thresholds	Threshold.for.Fold5	Avg.expected.loss	Expected.loss.for.Fold5
X1	Inf	0.5974843	1557.345	1554.387
X2	Inf	Inf	1557.016	1557.673
X3	Inf	0.6021872	1552.744	1539.599
X4	0.7066438	0.6552588	1538.611	1536.313
X5	0.7097497	0.6427426	1542.555	1536.313
LASSO	0.6887223	0.6427854	1548.832	1539.928

Below are some plots about optimal Model threshold. First, Model 4 logit has a high threshold essentially minimizing false positives.

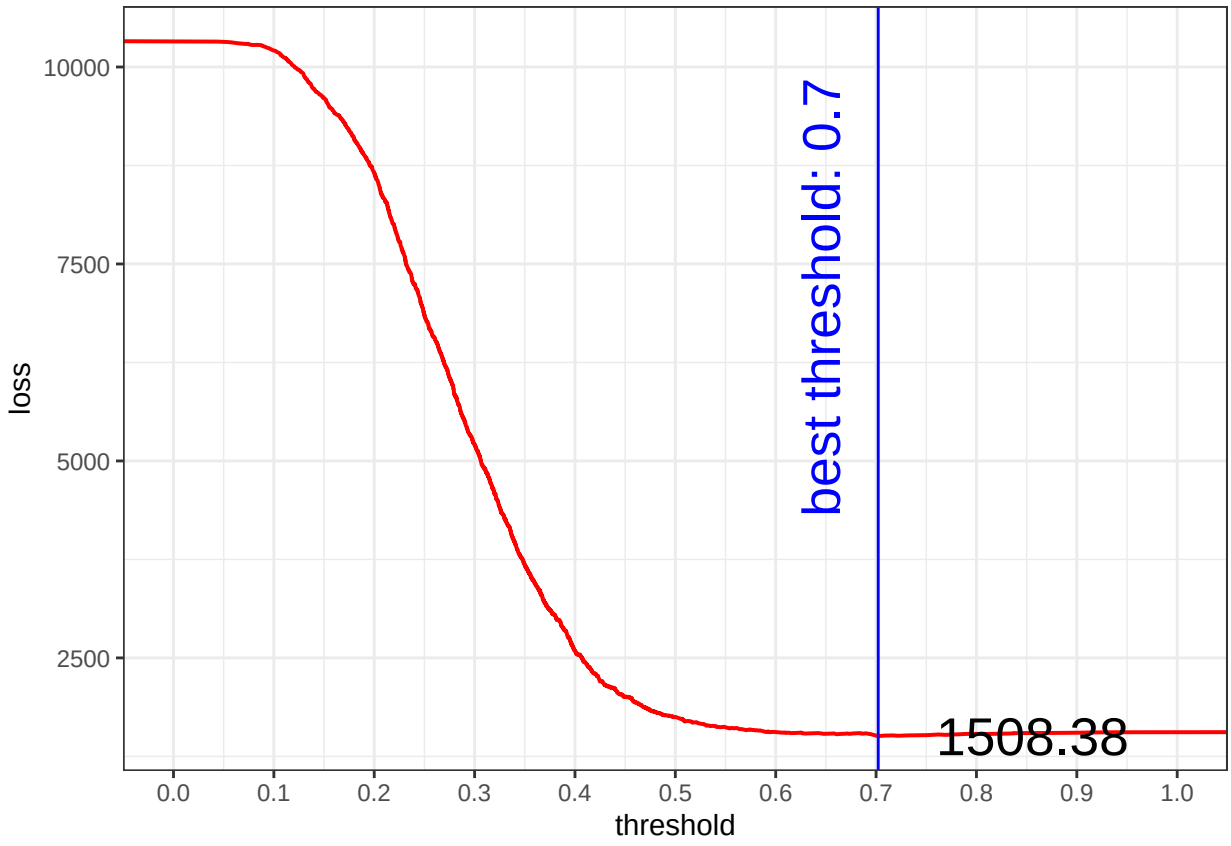


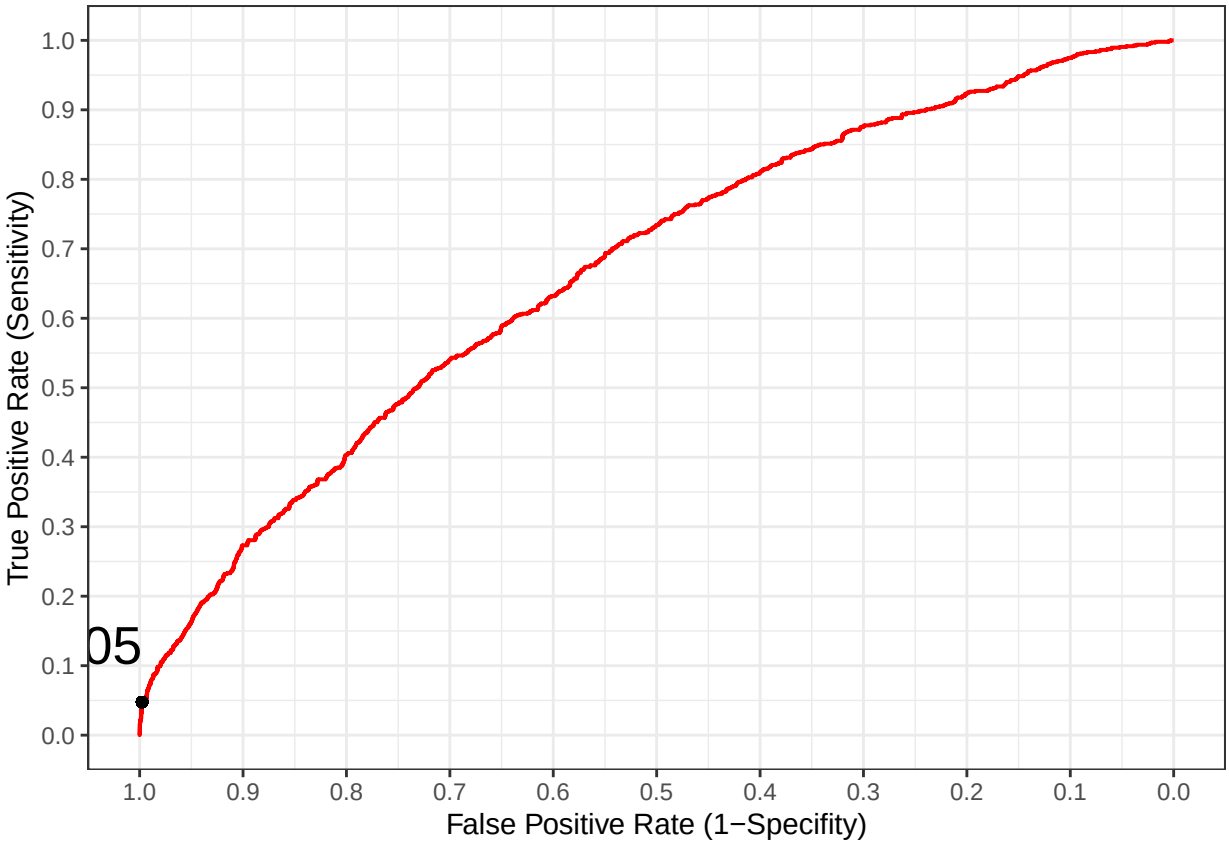


The confusion matrix in percentages shows the same result as in the main report but putting into perspective our general decision. We air on the side of not betting on companies 99% of the time if we listen to Model 4.

Finally the same plots about threshold for the Random Forest Model. Slightly higher threshold results in slightly lower expected loss.

	no_fast_growth	fast_growth
no_fast_growth	68.0	30.9
fast_growth	0.1	1.0




True Positive Rate (Sensitivity)
False Positive Rate (1−Specifity)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
05