Fast Grow Company Prediction

Code available: Github Repository

Introduction:

Nowadays, most companies aim to posses high annual growth rates and also want to know whether they maintain it. Therefore the aim of this analysis to predict wether an existing fim can sustain a 30% annual growth in sales or not in the long run. The prediction uses a dataset collected from a middle-size country in the European Union and focuses on the period between 2012 and 2014. In this report, there will be a brief introduction of the data handling such as cleaning and featuring, the process of choosing the type of method from simple multiply regression, random forest and logit model. After that, a cross validation will be implemented to pick the best model that be followed by a classification evaluation to determine an optimal threshold. In the end, that model will be decided on that has a the lowest prediction loss.

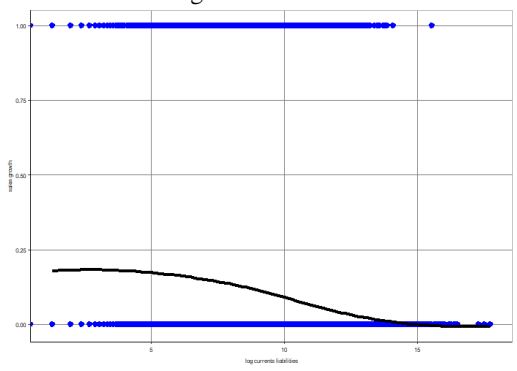
Data Preparation:

To begin with, those columns were droped where considerably high number of missing values were recorded. The prediction aims to determine fast growth, therefore all sales growth rates were calculated from 2012 to 2014. Due to the fact, that there is no data on growth in 2012, those variables were droped from that year after the gowth rate calculation. By this, a new varaible was created that has 1 as a high growth rate if it exceeds 30% and 0 if not. As a note, this fast growth rate is considered high that may influence the analysis significantly. Where the growth rate was missing due to lack of data from the previous year, they were filled with the mean value of the total growth rate after only having only 16 missing values. On the other hand, based on the same mechanism, a sales growth factor was created for high or not high growth rates.

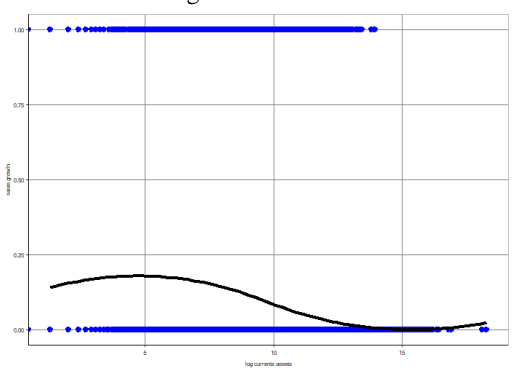
Descriptives:

In order to determine which variables may interact the best with the predicted variable, some data visualization was implemented. It turned out that the current liabilities and assets after a log transformation provide a visible regression pattern with high growth rates as it can be seen below:

Log Current Assets



Log Current Liabilities



In order to use them for the analysis, all of the infinite values were removed.

Models:

For the prediction, three different model methods were tested: multiply regression, random forest and logit model. From these were chosen that model that was used for the prediction. The section below discuss the details of each model and the argument for the final choice.

Simple Multiple Regression:

As it was determined before, there is a visible association among log current assets and liabilities therefore a simple multiple regression was created on growth rate binaries. It turned out that both coefficients are negative indicating that one percent higher in each descriptive variables would result in decrease of chance of having high growth rate (i.e. 1). However, considering the simplisity of the model and the low level of coeffients (i.e. -, it may not sufficient to carry out the prediction properly.

	factor	Estimate	Std. Error	AME
1	curr_assets_log	-0.01638392	0.0008915969	-0.01638392
2	curr_liab_log	-0.01126696	0.0008923671	-0.01126696

Random Forest:

While the simple multiple regression model might be too simple for the aim of the prediction, the random forest model may be sufficient as its purpose to find pattern in complex and messy data. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or mean/average prediction of the individual trees. During the analysis, two different random forest model was delivered that gave the following results:

MAE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
model_1	0.1657057	0.1658792	0.1668776	0.1681169	0.1704520	0.1716698	0
model_2	0.1663871	0.1665679	0.1672980	0.1686982	0.1709708	0.1722672	0

RMSE

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
model_1	0.2834528	0.2838637	0.2840958	0.2878082	0.2933042	0.2943246	0
model_2	0.2838402	0.2839900	0.2848133	0.2880793	0.2934456	0.2943073	0

Rsquared

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
model_1	0.04684652	0.04979767	0.05091069	0.05208113	0.05593120	0.05691959	0
model_2	0.04261909	0.04300856	0.04793186	0.04798082	0.05264531	0.05369930	0

The first model used 22 predictors while the second 50 predictors but only 2, 4 and 6 node sizes in both cases. The table shows that the avarge RMSEs are close to each other with 0,283 as much the minimums and maximums. As RMSE stands for the standard error of the residuals (i.e. difference between the predicted value and actual value) that measures the distance from the regression line. In other words, the lower the RMSE the better for the prediction if the model do not overfit the data. Mean Absolute Error (MAE) is another loss function used for regression models. MAE is the sum of absolute differences between the actual and predicted variables. So it captures the average enormousness of errors, without including their directions. So, in order to measure the loss function, MSE can provide a potential insight to measure it and as RMSE, the lower the better. Altough, MAEs are also close to each other in the models with approximately 0,168 as their mean value.

Logit Regression Model:

The logit regression model is a binomial regression model for finding association between binary or ordinal response probability and explanatory variables. The outcomes gives the probability of having either of the binary variables therefore it fits well for classification purposes. Because the purpose of the analysis to determine whether compnies can have a fast growth of 30% in their sales or not, this method was chosen for carrying out the prediction. In the following part, multiple logit models were tested with cross validation that are discussed and the best one will be chosen for the classification and prediction.

Cross Validation:

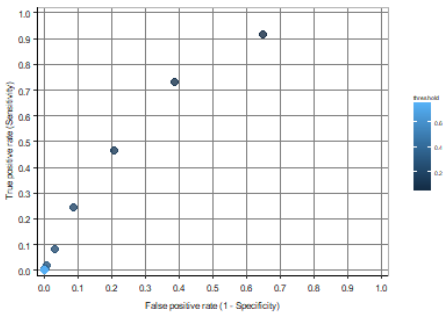
Cross-validation is a tool for evaluating models by training multiple models on subgropus of the dataset and evaluating them on the holdout subset from the remained dataset. The number of subsets can be given for the validation (i.e. also called folds) which is five in this analysis. The six models consideres all of the explanatory variables and also a LASSO for interactions. The validation gave the following outcomes of the models:

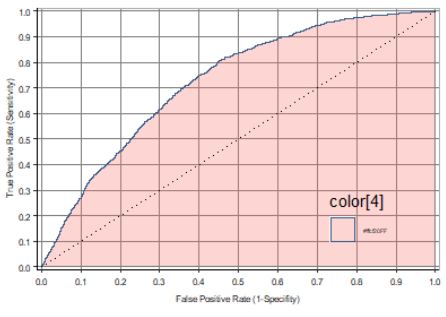
	Number.of.predictors	CV.RMSE	CV.AUC
X1	13	0.2879807	0.7086108
X2	20	0.2878334	0.7128212
X3	37	0.2872153	0.7208285
X4	80	0.2871434	0.7234684
X5	154	0.2874878	0.7198861
LASSO	1	0.2947971	0.7012614

As it was explain before, the RMSE stand for a measurement of the distance of the predicted values from the regression line. As at the random forest case, the RMSEs do not differ significantly from each other. As a second evaluation criterion, the AUC is for measure the area under the ROC curve that will be detailed in the later part of the report. Until now, in order to pick the best model is to know that the higher the AUC is the better. Based on these findings, the X4 logit model was chosen for testing. It turned out that while the training division had an RMSE 0,287 and an AUC 0,723, the testing subset had an RMSE of 0.2886916 that is considered close enough for carrying out the prediction.

Classification and Loss Function for Logit Model:

Classification refers to a predictive modeling problem where a class label is predicted for a given example of dataset. Loss functions for classification are representing the price paid for wrongness of predictions in classification problem. In orther to see the sensitivity of the model, a threshold can be selected. If the probability is greater than this threshold value, the event is predicted to happen otherwise it is predicted not to happen. In the following ROC charts you can see how the different thresholds from 0,05 to 0,75 (increased by 0,0,5) influences the trade off between True and False Positive. Also, the curve on the right show the mention AUC where the area above the 45° line.





During the choice of the best logit model, it was stated that the higher the AUC the better. It van be explained by that if the curve is steeper the tread off rate of having more True Prostives with less False Positive is lower. Therefore, the choice of optimal threhold is essential for the prediction.

As the Loss Function aims to determine the cost of wrongness, even if the best logit model was determine, all of them were checked that can be seen in the below table. For the X4 model, the average loss turned out to be the lowest from each that supports the model choice.

	Avg.of.optimal.thresholds	Threshold.for.Fold5	Avg.expected.loss	Expected.loss.for.Fold5
X1	0.1666253	0.1442747	0.8867096	0.9040471
X2	0.1660088	0.1914623	0.8772840	0.8991070
X3	0.1500907	0.1439600	0.8685439	0.8721262
X4	0.1569428	0.1440090	0.8663393	0.8770663
X5	0.1621814	0.1883223	0.8720398	0.8717462
LASSO	0.1521547	0.1382486	0.8986965	0.9013490

Best Threshold and Results:

In the end, before carrying out the full prediction, in order to have the less expected loss the optimal threshold need to be picked. The ratio of False Positive to False Negative was set to 5 to 2 which gave an optimal threshold of 0,156. This ratio was determined by a business problem whether to buy a new equipment worth of 10000 USD if they expect to grow by 30% in their sales but in the end not while stick to their previous equipment that values 2000 USD and having teh high growth rate.This threshold rate was used for the final confusion table for the X4 logit model in which if the probability of having a higher growth rate than 30% is lower than the optimal threshold than it does not have a high growth, otherwise yes.

	Reference	
Prediction	**no_high_growth**	**high_growth**
no_high_growth	8315	628
high_growth	1903	476

The results shows that the prediction was right for 8315 (not high sales growth) and 476 (have high sales growth) cases but was wrong in the other outcomes. As a false negative, high growth was predicted for 1903 firms which turned out to be false, while forecasted no high growth 628 times when it was wrong (false positive).

Conclusion:

Based on the above findings, from the three prediction method the logit regression was defined as the optimal model selection for carrying out the prediction. The analysis was aiming to forecast whether a comapny can have a 30% growth in their sales compared to their prevoius year perfomance. For the analysis, a dataset from Eurpoen Union was used which contained data on middle sized companies from the period between 2012 and 2014 applied. In the view of a consideribly high average expected loss, the prediction may have a better fit if the growth rate is lower than 30%. Although the cross-validation showed that the model generalization fits to other independent data, the ROC curve also showed a high trade off beween true and false positive that the model must be aware of.