Prediction and Classification of Models

Haaris Afzal Cheema & Mahrukh Khan

Introduction

The aim of this project was to predict fast growth of firms using probability prediction models. Detailed information on firm data from a European country will be used for this analysis. This data was collected, maintained and cleaned by Bisnode. The prediction of the binary target variable will be done using various financial metrics pertaining to balance sheets and income statements, along with other relevant management information. In this report, we will look at a cross-section of companies in 2013.

Label Engineering, Sample Design and Feature Engineering

In order to create our binary target variable for fast growth, firstly we created a variable for CAGR (Compound Annual Growth Rate) which was computed as the rate of sales growth between two years where the growth is assumed to compound exponentially. The formula for this computation can be seen below:

$$CAGR = (\frac{S}{S2})^{\frac{1}{t}} - 1$$

t being the time in years, set at 2 **S** is the Sales (in euro) currently **S2** is the Sales (in euro) in 2 years

Upon analyzing the distribution of CAGR, we decided that if the CAGR of any company exceeded 25% for the mentioned time interval, the company would be considered as fast growing. The reason for using two years in the calculation for CAGR was that one year of growth may not be able to capture sufficient and true trends. A 2-year interval was therefore used to in order to gauge companies as fast growing if they maintain their growth over a longer period of time. Greater time-horizons could also be used, but making predictions for a greater span of time will become increasingly difficult. The next task was pertaining to sample design. Observations pertaining to SMEs (Small and medium enterprises) were kept. This was done based on the annual sales. Companies with annual sales between 1000 and 10 million Euros. As a result of this, our final clean data contains 16,845 observations. We can see that out of these observations, 26.2 % (4424) of firms will experience fast growth. Thirdly, we move on to feature engineering, which was done based on exploratory analysis as well as domain knowledge. For instance, upon exploration, sales was observed as having a strongly skewed distribution for which a log transformation was done. Next, some financial variables such as intangible, fixed and current assets, which cannot take on negative values were flagged and the negative values were replaced with zero. Similar steps were followed in case of variables such as inventories and subscribed capital. Next, ratios were created where balance sheet items were divided by the size of the balance sheet whereas PNL items were scaled according to the total sales. This allowed for more ready interpretations and identification of skewness in variables. We also made use of winsorization to address extreme values. For instance, in case of variables like income before tax as a ratio of total sales, bounds were set between -1 and 1. Finally, lowess explorations were done, on the basis of which, non-linearities were captured and accordingly incorporated in our models.

Modelling and Evaluation

Prior to model-building, data partitions were created. A holdout set was created which contained randomly selected 20% observations from the original dataset. The training set (remaining 80%) was used for 5-fold cross validations, in order to obtain 5 different train and test samples which will be used for evaluating the different logit, LASSO and random forest models.

To assist with modelling, we divided variables into different categories. The categories can be seen below.

- **Main firm variables:** These are some of the key variables to predict fast growth, such as main profit and loss and balance sheet elements.

- **Quality variables:** Pertaining to balance sheet completeness and flags.
- **Engine variables 1:** Additional profit and loss and balance sheet elements such as income before tax, long term and short assets etc.
- **Engine variables 2:** Squared terms for some variables, such as profit and loss, income before tax and share of equity. These are the variables which are mostly to be between -1 and 1.
- **Engine variables 3:** Flags for engine 2 variables.
- **Growth Variables:** Variables pertaining to change in sales, along with flags.
- **Human Capital Variables:** Variables such as sex and age of CEO or whether their is foreign management.
- **Firm History Variables:** Age of company, region, industry etc.
- **Interactions 1 and 2:** interaction terms pertaining to industry as well as for sales.

Table 1: Models on which algorithms will be run

Model	Predictor Variables
X1	=Sales + Log Sales + Log of Change in Sales + Industry + Profit and Loss
X2	=X1 + Fixed assets + Equity + Current liabilities (with flags) + Age + Foreign management
X3	=Log Sales + Log Sales squared + Main firm variables + Engine variables 1 + Growth Variables
X4	=X3 + Engine variables 2 + Engine Variables 3 + Quality variables, Human Capital Variables
X5	=X4 + Interactions 1 and 2

Logit Models

RMSE

We initiated our analysis with 6 logit models including a LASSO logit model. The reason for using these instead of LPM models was to ensure that our predictions ranged between 0 and 1. As can be seen, models of increasing level of complexity (number of variables as well as functional forms) were used. The primary metric used to evaluate these models was the RMSE (root-mean-squared error). From the 6 models, the LASSO logit and model X4 produced the lowest RMSE.

Classification

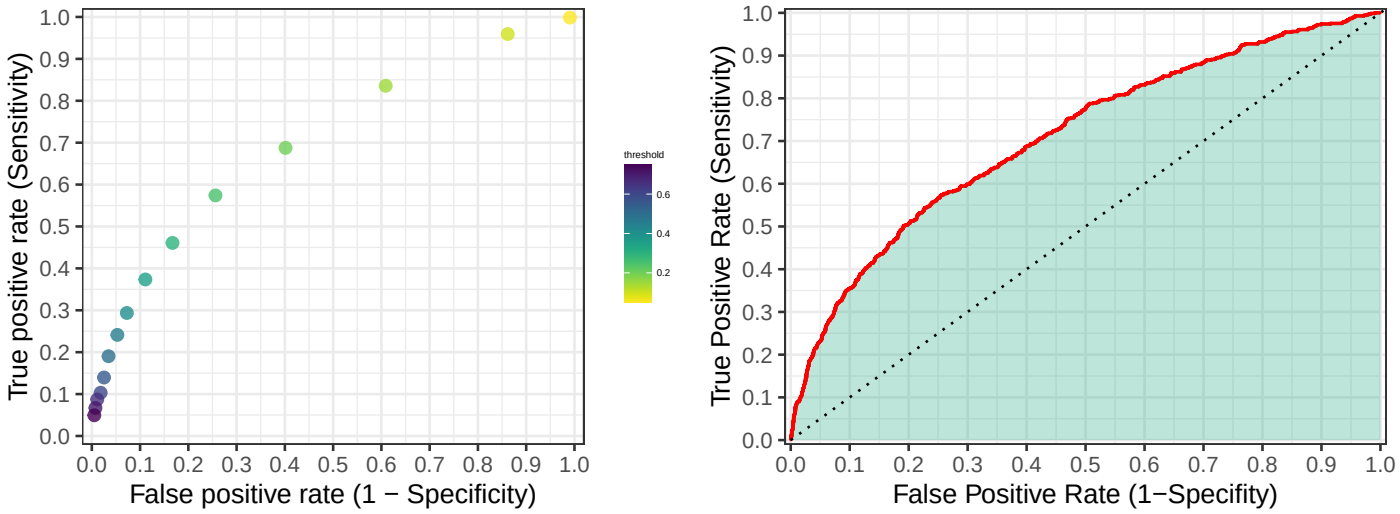

Figure 1: ROC Curves

Figure 1 (left) above shows the values for the ROC against the values of the selected threshold. Values for the threshold were varied between a sequence of 0.05 to 0.75 with gap of 0.05. The ROC is away from the 45 degree line indicating that it is giving reasonably good predictions. AUC for Model X4 is represented by the shaded area below the curve for the fifth fold. The X4 Model had a greater area under the curve in comparison to LASSO logit and is a simpler model, hence we will consider it as our benchmark model. The calibration curve for this model showed that for lower predicted probabilities (the case of firms that are not fast growing), the predictions are very well calibrated. For higher predicted probabilities, there is some deviation from the 45 degree line.

Best Optimal Threshold

The loss function will be defined from an investment perspective. In this scenario, not investing in a company experiencing rapid growth will have greater repercussions as opposed to investing in a company which doesn't end up growing rapidly. Due to this, in our loss function, the cost of a false negative would be three times more important as opposed to a cost associated with a false positive. Based on this information, we calculate the threshold which minimizes our expected loss. The optimal classification threshold based on the formula turned out to be 0.25 (1/1+3). We also used a more robust way to find this optimal threshold via a search algorithm. This was run on the work set with 5-fold CV. For all 6 models we observed that the average optimal threshold of algorithm was quite similar to the formula one.

We note that Model X4, which has the lowest RMSE, also has the lowest expected loss that is 0.509. We also used this optimal threshold to find expected loss on the holdout set which turns out to be 0.513 which is slightly larger than our training set. Below we have shown the confusion matrix for Model X4. The correctly predicted positive percentage is 40.6%.

Table 2: Confusion Matrix Model X4

	no_fast_growth	fast_growth
no_fast_growth	54.9	10.2
fast_growth	20.7	14.2

Random Forest

For our random forest model, we use the same variables as Model X4, without feature engineering. This model outperformed Model 4, with a cross validated RMSE of 0.397 and AUC of 0.703. We used predicted probabilities to find the optimal threshold (0.278) which was then used for classification. We re-estimated the model on the work set and then proceeded with the prediction on the holdout set. The RMSE for holdout set turned out to be 0.4017 and holdout AUC is 0.713. Both values are almost the same in both sets.

Table 3: Confusion Matrix Random Forest

	no_fast_growth	fast_growth
no_fast_growth	57.3	10.5
fast_growth	18.3	13.9

Finding the optimal threshold was done similarly through a search algorithm with the same loss function. In the confusion matrix, the correctly predicted positive percentage is 43.1%. (Higher than Model X4).The expected loss for the random forest model holdout set is 0.498 which is lesser than Model X4. If external validity is high, this means that we can expect to make a loss of 498 Euros per classification on the live data.

Summary

Table 4: Summary of Model Performance Measures

	Number.of.predictors	CV.RMSE	CV.AUC	CV.threshold	CV.expected.Loss
Logit X1	11	0.4083323	0.6299991	0.2866662	0.5582507
Logit X4	80	0.3990175	0.6971744	0.2442080	0.5026584
Logit LASSO	46	0.3983462	0.6784476	0.2468959	0.5330039
RF probability	36	0.3972292	0.7026842	0.2795713	0.5031884

Now that we have tried out multiple models, we can compare their performance in terms of prediction as well as classification. At an overall level, model X4 and random forest outperform the rest. When compared with each other, the RF model has a slightly lower RMSE, indicating a better prediction model. Model X4 has slightly lower expected loss, indicated a superior classification model. This could be due to random forest not optimizing for the best AUC. Nonetheless, we selected random forest to be our best performing model.

Additional Task

For this task, we decided to create sub-samples based on industry sectors: manufacturing and services. The sample design process for manufacturing was straightforward given that the industry was clearly defined. For services, a greater level of decision-making was involved. Food, accommodation and repairs had very few observations and so other services such as waste management and administrative duties were also clubbed together. Other services could also have been added but this would create a great mismatch in the sample sizes for comparing manufacturing and services.

The random forest model (best performing) was used. The RMSE for the prediction for manufacturing was 0.411 whereas in the case of the services, it was 0.396. Based on the AUC metric, manufacturing sector was 0.674 whereas services had a value of 0.709. Classification, too, was done separately for both sub-samples. The loss was function was kept the same, in which case a false negative is three times more important than a false positive. This will be kept the same for both sub-sample classifications. Using this loss function, optimal thresholds were evaluated for both industries. For manufacturing, the optimal threshold was 0.304 with an average expected loss of 0.543. In the case of services, the optimal threshold turned out to be 0.267 and the expected loss was equal to 0.494. The correctly predicted positive rate for manufacturing turned out to be 33.3% whereas for services it was 38.6%. The confusion matrices for both can be seen below.

Table 5: Confusion Matrix Service

	no_fast_growth	fast_growth
no_fast_growth	53.1	8.8
fast_growth	23.4	14.7

Table 6: Confusion Matrix Manufacturing

	no_fast_growth	fast_growth
no_fast_growth	59.6	12.7
fast_growth	18.5	9.2

With the loss function and the confusion matrices, we calculated the expected loss on the holdout sets. For manufacturing, it turned out to be 0.565 whereas for services it was 0.497.

Conclusion

The relevant metrics for both sub-samples for prediction show that services industry has a lower RMSE and a higher AUC value indicating superior prediction for this particular industry. Consequently, we can can expected to make a smaller error on average for predicting fast growth for firms in the services industry as opposed to the manufacturing sector. We also see that its best optimal threshold (0.267) is closer to our formula threshold value (0.25). Using classification, we can see from an investment perspective that the expected loss on both training and holdout sets is lower in the services industry and so, it would be the preferred sector for investment out of the two industries.