Technical Report: Assignment 3: Finding Fast growing firms

Data

The dataset employed in this assignment is derived from the "Bisnode firms" dataset featured in Gabor's Data Analysis book.The data set contains company data from a middle-sized country in the EU in 2005-2016 in three selected industries (auto manufacturing, equipment manufacturing, hotels and restaurants).It includes 287,829 observations (46 412 firms) and 48 variables for the period 2005-2016. The data was further filtered for the analysis and all the decisions and code documented in the sample_preparation.R script.

Sample Design

Firstly, I filter the data for for year 2012 - 2013 and created a new variable called "growth" which is the growth rate of sales between 2013 and 2012. Moreover, I focus on middle-sized firms with sales between 1000 and 10 000 000 euros. I defined the fast-growing firms as those with a growth rate of sales greater or equal to 50%. The final sample contains 16820 firms of which 17.7% are fast-growing firms. The sales variable is highly skewed to the right. Therefore, I took the log of sales to make it more normally distributed. The distribution of the log of sales is shown in the figure below.

Feature Engineering

I use similar features as was used in predicting firm exit prediction which consists of four groups: size, management, financials and other characteristics. Some of the financial features were winsorized to reduce the effect of outliers and keep them within a reasonable range.

The the illustrative code below:

```
# Winsorize the financial variables
zero <-  c("extra_exp_pl", "extra_inc_pl", "inventories_pl", "material_exp_pl",
```

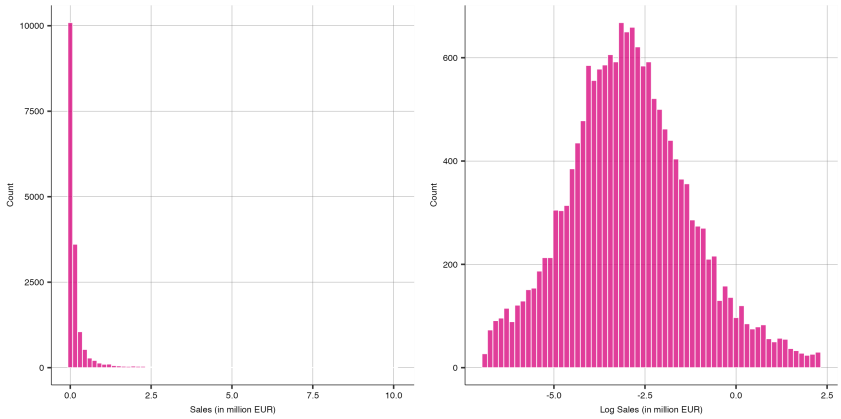

Figure 1: Distribution of Sales

```
            "personnel_exp_pl","curr_liab_bs", "fixed_assets_bs", "liq_assets_bs",
            "curr_assets_bs", "subscribed_cap_bs","intang_assets_bs")
data <- data %>%
  mutate_at(vars(zero), funs("flag_high"= as.numeric(.> 1))) %>%
  mutate_at(vars(zero), funs(ifelse(.> 1, 1, .))) %>%
  mutate_at(vars(zero), funs("flag_error"= as.numeric(.< 0))) %>%
  mutate_at(vars(zero), funs(ifelse(.< 0, 0, .)))
# for vars that could be any, but are mostly between -1 and 1
any <-  c("extra_profit_loss_pl", "inc_bef_tax_pl","profit_loss_year_pl", "share_eq_bs")
data <- data %>%
  mutate_at(vars(any), funs("flag_low"= as.numeric(.< -1))) %>%
  mutate_at(vars(any), funs(ifelse(.< -1, -1, .))) %>%
  mutate_at(vars(any), funs("flag_high"= as.numeric(.> 1))) %>%
  mutate_at(vars(any), funs(ifelse(.> 1, 1, .))) %>%
  mutate_at(vars(any), funs("flag_zero"= as.numeric(.== 0)))
```

As a result of feature engineering, the following groups of predictors were created:

The graphs in below shows some of the features that were used in the analysis and their relationship with the growth rate of sales. According to the boxplot, fast-growing firms tend to have lower sales, lower assets and higher current liabilities. Regarding the other characteristics, fast-growing firms tend to be younger and have young managers on average.

Probability prediction and Model selection

I performed probability prediction using the logit and select the best model by cross-validation. The cross-validation was performed using the 5-fold cross-validation method. Furthermore, the

Table 1: Fast growth predictor variables

Model No.	Used variables
Firm	Age of firm, squared age, a dummy if newly established, industry categories, location regions for its headquarters, and dummy if located in a big city.
Financial 1	Winsorized financial variables: sales, fixed, liquid, current, intangible assets, inventories, equity shares, subscribed capital, sales revenues, income before tax, extra income, material, personal and extra expenditure, extra profit.
Financial 2	Flags(extreme, low, high, zero – when applicable) and polynomials
HR	For the CEO: female dummy, winsorized age and flags, flag for missing information; foreign management dummy; labor cost, and flag for missing labor cost information.
Data Quality	Variables related to the data quality of the financial information, flag for a problem, and the length of the year that the balance sheet covers.
Interactions	Interactions with firm size, sales growth and industry.

chosen model was evaluated on holdout data. The working data was split into 80% training and 20% holdout data sets. Train data has 13456 observations (2341 fast growing firms) and holdout data has 3364 observations (479 fast growing firms). Moreover, I include Logit LASSO which start with our most complex model and then shrink the coefficients of the least important variables to zero. See the file model_selection.R for the detailed variables used in each model. The table below presents results of the cross-validated average RMSE and AUC(the average of five folds) for each model.

Table 2: RMSE and AUC for models

	Number of predictors	CV RMSE	CV AUC
X1	10	0.365	0.684
X2	17	0.357	0.722
X3	31	0.353	0.736
X4	74	0.350	0.743
X5	141	0.350	0.740
LASSO	75	0.349	0.718

The X4 and X5 models have the lowest RMSE and highest AUC. Following the rule of parsimony, I choose the X4 model as the best model and further evaluate it on holdout data and plot ROC curve. The holdout RMSE and AUC are 0.358 and 0.76 respectively

Using the X4 model, I construct confusion table for two thresholds: 0.5 and 0.18. The threshold

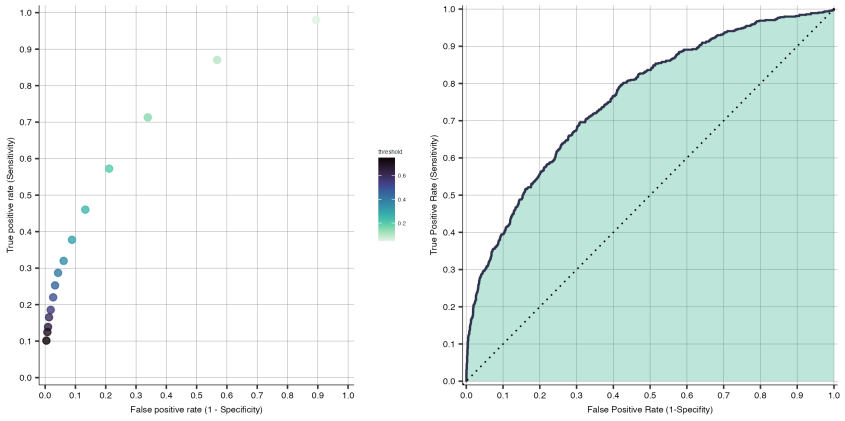

Figure 2: ROC Curve

of 0.5 is the default threshold for the logit model and 0.18 is the mean of predicted variables.

Table 3: Confusion Matrices with Different Thresholds

	no growth	growth	Total
no growth	79%	15%	94%
no growth	2%	4%	6%
Total	81%	19%	100%

Threshold: 0.5

	no growth	growth	Total
no growth	61%	7%	68%
no growth	20%	12%	32%
Total	81%	19%	100%

Threshold: 0.18

Having a threshold of 0.5 leads to predicted 6% of fast growing firms compared to 32% with the threshold of 0.18. Moreover, higher threshold has less false positive rate (2% vs 20%) and higher false negative rate (15% vs 7%). I also calculate measures such as sensitivity, specificity, precision, and accuracy. The results are shown in the table below.

Table 4: Accuracy Measures for two Thresholds

	Threshold: 0.5	Threshold: 0.18
Accuracy	83%	72%
Sensitivity	22%	75%
Specificity	97%	62%

Loss function

To define the loss function I assume that the cost of false negative is 2 times higher than false positive. Although we miss the opportunity to invest in a firm that will grow fast, there is a still possibility that we end up investing in a firm that will grow in sales but not as fast as the we defined. Once we have our loss function, we can calculate the optimum threshold that minimizes the loss. The table below shows our estimations:

Table 5: Summary of model performance measures

	CV RMSE	CV AUC	CV threshold	CV expected Loss
Logit X1	0.365	0.684	0.320	0.318
Logit X4	0.350	0.743	0.357	0.284
Logit LASSO	0.349	0.718	0.326	0.295
RF probability	0.347	0.749	0.381	0.281

As a result, the best model is the Random Forest model with 35 predictors. The optimum threshold is 0.381 and the expected loss is 0.281. The best model is the random forest model with a optimal probability threshold of 0.381. It has the lowest RMSE and AUC, however Logit X4 is marginally worse. Since the companies are obliged to report and interpret the predictors, I would recommend the Logit X4 model. The model has a good predictive power and is easy to interpret. Switching to logit X4 would result in just in 3 euro loss per firm compared to the random forest model. The graph below shows the ROC curve of the logit X4 model.

Appendix

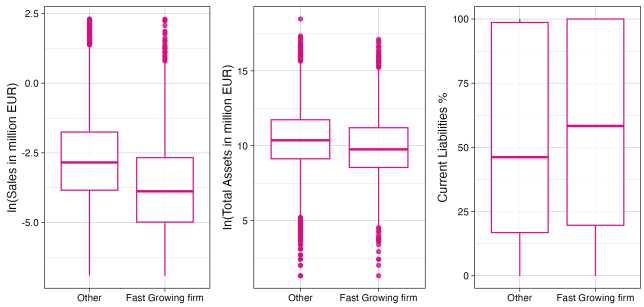

Figure 3: Boxplot of Financials

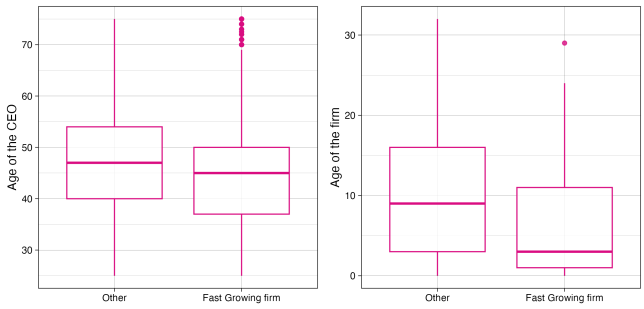

Figure 4: Boxplot of Age

Table 6: Average Marginal Effects (dy/dx) for Logit Model X2

Variable	Coefficient	dx/dy
age	-0.077	-0.010
curr_liab_bs	0.573	0.074
curr_liab_bs_flag_error	0.409	0.053
curr_liab_bs_flag_high	-0.270	-0.035
fixed_assets_bs	-0.109	-0.014
foreign_management	0.128	0.017
ind2_cat26	0.490	0.065
ind2_cat27	0.480	0.064
ind2_cat28	0.684	0.095
ind2_cat29	0.813	0.117
ind2_cat30	0.881	0.128
ind2_cat33	0.465	0.061
ind2_cat55	0.116	0.014
profit_loss_year_pl	-0.575	-0.074
sales_mil_log	-0.080	-0.010
sales_mil_log_sq	0.047	0.006
share_eq_bs	0.321	0.041

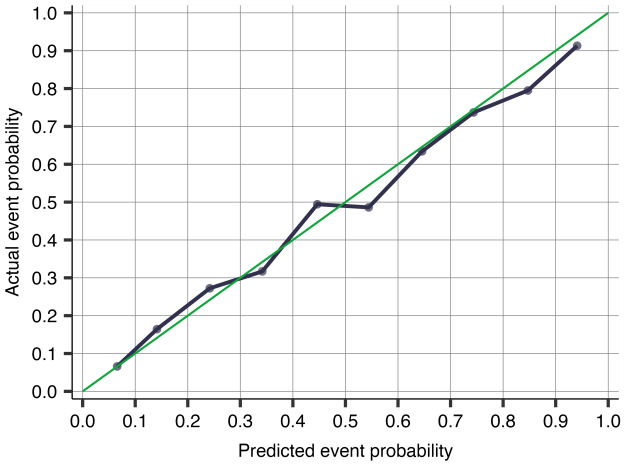

Figure 5: Calibration plot X4 Model