Assignment 3: Summary Report **Abay Jumabayev**

Detailed technical report and data can be found on my GitHub page.

Objective:

To find fast growing firms. This exercise is helpful for investors who want to generate maximum profits by investing in firms.

Data:

Bisnode-firms dataset – a panel data of firms located in some city of Europe over a period of 2010-2014. The data has 140,100 observations and 47 columns.

For the exercise I used a cross-section of 2012, which had 21,723 observations. The data for 2012 included only firms that existed in 2012, i.e. had financial reports for that year.

Variables:

The original data had financial, managerial, ownership, status (HQ) information about firms.

I created/transformed several variables including the dependent variable – *fast_growth*. *fast_growth* is 1 for firms who has at least 20% growth in sales for each year in a period 2012-2014. There are 3726 firms, or 17 % that are considered fast growing. The better alternatives would be employment growth, assets growth, or some financial ratios like ROI. However, the dataset either lacks values for those variables or has the values with lots of errors (for example, there are a lot of firms with negative assets).

I also created other variables which would help determine the fast growing company. Those include: industry category, age of a firm (plus age squared), gender of a CEO dummy, region category, log of sales (plus log squared), return on assets (ROA), return on equity (ROE), return on total assets (ROTA), current ration (CR), and historical growth of sales (*diff_ln_sales*). As the data was not clean enough, I flagged firms which had problems with financial variables. Also, I winsorized *diff_ln_sales* because historical growth rate would not predict fast growth well. According to the data, the lower the growth rate of sales, the higher the probability of fast growth. That is why I replaced *diff_ln_sales* values below -0.5 by 0.5.

Finally, I dropped all observations where a key variable is missing. I left with 19,036 observations.

Models:

I built 7 models to predict the fast growing firms: five logit models of different complexity, one lasso logit, and one random forest.

Model 1: Included log of sales, log of sales squared, difference in log of sales (winsorized), profit/loss, and industry dummies.

Model 2: Model 1 variables + fixed assets, share equity, current liabilities, age of a firm, foreign management dummy, and flags.

Model 3: Model 2 variables + all other financial variables including financial ratios, age squared, new firm dummy, region dummies, and urban dummies.

Model 4: Model 3 variables + quadratic terms of profits, income, and equity, firm characteristics, and all flags.

Model 5: Model 4 variables + interaction variables.

Model 6: Model 5, on which I applied Lasso.

Model 7: Random forest, which had all raw variables as dependent variabels.

Part I: Probability prediction

After I checked simple linear and logistic models, I used 80-20 train test split and trained six models (all except random forest). As it is seen from the table below, Model 4 (X4) performed best here. I would use it as a benchmark.

Table 1. Summary (without loss fn)

Model	Number of predictors	CV RMSE	CV AUC
X1	11	0.36774	0.591718
X2	18	0.36302	0.644893
X3	39	0.360019	0.664026
X4	83	0.358137	0.667756
X5	164	0.358701	0.666505
LASSO	145	0.358988	0.666603

Part II: Classification

Here I introduced a loss function. To do this, suppose I have 1000 EUR and I want to invest in firms. I know from the data that the median sales growth rate of fast growing firms is 78% while the median of non fast growing firms is -6%. Here I assume that the money invested will be returned proportionally to the sales growth. That is, if I invest in fast-growing firm, I will get 1781 EUR on average and if I invest in non-fast-growing firm, I will get 937 EUR on average. That is why I chose FP = 1781, FN = 937, which gives a cost of approximately 0.5.

As can be seen from the table below, the lowest cross-validated expected loss is achieved by the random forest model (rf_p).

Table 2. Summary (with loss fn)

Model	Number of predictors	CV RMSE	CV AUC	CV threshold	CV expected Loss
X1	11	0.36774	0.591718	0.933265	153.9356
X2	18	0.36302	0.644893	1.009283	153.8186
X3	39	0.360019	0.664026	1.214315	154.1206
X4	83	0.358137	0.667756	0.609045	152.1156
X5	164	0.358701	0.666505	0.819533	152.682
LASSO	145	0.358988	0.666603	1.049653	152.9408
rf_p	45	0.356338	0.67726	0.638283	151.242

To understand how random forest predicts the fast growing firms, let's look on the confusion matrix. Out of 3807 firms on holdout set, random forest model correctly predicted 3175 (that is 83%) firms.

Now return to our investor example. The model predicted that there are 23 fast-growing firms. Suppose I invest 1000 EUR to each company. I will spend 23,000 EUR and get in return 8*937 + 15*1781 = 34,211 EUR. That is 49% increase in profits.

Table 3. Confusion matrix for Random Forest model

	Predicted not fast-growing	Predicted fast-growing
Actual not fast-growing	3160	8
Actual fast-growing	624	15

Now suppose that I did not have a loss function. The benchmark model was Model 4, and the threshold = 0.3. Suppose I invest 1000 EUR to each company. I will spend 329,000 EUR and get in return 181*937 + 148*1781 = 433,185 EUR. That is 32% increase in profits, which is much less than in previous case.

Table 4. Confusion matrix for X4 (threshold = 0.3)

	Predicted not fast-growing	Predicted fast-growing
Actual not fast-growing	2987	181
Actual fast-growing	491	148