Summary Report

Motivation

The goal of the assignment is to employ standard binary classification techniques in machine learning to identify and predict which firms are likely to be considered as “fast-growing.” A “fast-growing” firm, in this context, refers to a company that exhibits substantial growth in the revenue. Revenue growth is closely associated with fast growth because it reflects the overall health and success of a company. It usually signals market demand, effective business strategies and competitive advantages.

Data and target definition

The dataset employed in this assignment is derived from the “Bisnode firms” dataset which contains company data from a middle-sized country in the EU in 2005-2016 in three selected industries. The target variable is a binary variable that indicates whether a firm is considered as “fast-growing” or not. I defined the fast-growing firms as those with a growth rate of sales greater or equal to 50%. The determination of whether a sales growth rate of 50% or higher is considered “fast” is a matter of debate and hinges on various factors, including industry norms, the company’s historical performance, and prevailing economic conditions. However, a growth rate of 50% is generally acknowledged as substantial and can serve as an indicative marker of rapid expansion. The final sample contains 16820 firms of which 17.7% are fast-growing firms.

Model estimation

Having evaluated the performance of several classification models, I have chosen the logit X4 and random forest models. To define the loss function, I am assuming that the cost of a false negative is two times higher than that of a false positive. While there is a risk of missing the opportunity to invest in a firm that will experience rapid growth, there is also the possibility of investing in a firm that will grow in sales but not at the pace we have defined. The table below presents results of the cross-validated average RMSE, AUC(the average of five folds) and Expected Loss for each model.

Table 1: 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
RF probability	0.347	0.749	0.381	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.

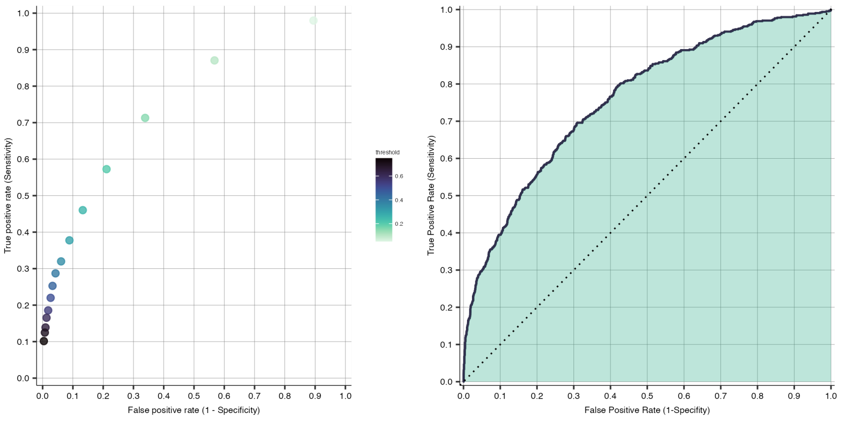

Figure 1: ROC Curve

Classification from predicted probabilities

The optimal probability threshold four our logit X4 is 0.357. The table below shows model measures for the holdout sample:

The sensitivity of the model is 31%. This means that the model correctly identifies 31% of the fast growing firms. The specificity of the model is 94%. This means that the model correctly identifies 94% of the non-fast growing firms. The accuracy of the model is 82%.

ht

Table 2: Accuracy Measures

	Threshold: 0.36
Accuracy	82%
Sensitivity	31%
Specificity	94%

Stability test

The stability test is performed by splitting the holdout into two subsamples (manufacturing and services industry) and estimating the model on each subsample. The table below shows the number of fast growing firms in each industry:

Table 3: Industry distribution of fast growing firms

	No Growth	Growth
Auto manufacturing	57	21
Equipment manufacturing	796	178
Hotels and restaurants	1870	442

Intuitively, the model should perform differently on each subsample. From the table above we can see that the number of fast growing firms is much higher in the hotels and restaurants industry. Hence the model should perform better on this subsample. The Expected Loss and AUC measures for the manufacturing industry is 0.32 and 0.70 respectively. The Expected Loss and AUC measures for the services industry is 0.30 and 0.78.

Conclusion

To sum up, using dataset of firms from a middle-sized country in the EU in 2005-2016 in three selected industries, I have employed standard binary classification techniques in machine learning to identify and predict which firms are likely to be considered as "fast-growing." I specified various models and evaluated their performance. Using the specified loss function, I determined the optimal probability threshold for each model. The best model is the random forest model with a optimal probability threshold of 0.381. It has the lowest RMSE and highest AUC. While the random forest model excels in predictive power, the logit X4, though marginally inferior, offers easier interpretability. Opting for the logit X4 incurs only a 3 euro loss per firm compared to the random forest model. The model demonstrates a 31% sensitivity, correctly identifying 31% of fast-growing firms, and a 94% specificity, accurately identifying 94% of non-fast-growing firms. The overall accuracy of the model stands at 82%. Notably, the model's performance varies across industry subsamples. Specifically, it performs better on the services industry, where the Expected Loss and AUC measures are 0.30 and 0.78, respectively. In contrast, for the manufacturing industry, these measures are 0.32 and 0.70.