CEU CENTRAL EUROPEAN UNIVERSITY

Summary Report: Predicting Fast-Growing Firms

1. Introduction

This assignment seeks to identify fast-growing firms based on the bisnode-firms dataset (2010–2015). The business goal is to prioritize investment prospects by predicting which firms experience at least 20% sales growth between 2012 and 2013. This threshold captures robust expansion (higher than typical 5–10% growth benchmarks) without being overly restrictive.

We considered other definitions, like comparing growth between 2012 and 2014 or using a 50% threshold, but the first would result in fewer observations and tha latter would eliminate many valid high potential firms.

Through the assignment we committed data preparation and feature engineering which included filtering to 2012-2013, removing excessive missingness and outliers, engineering ratios for financial statements. We built various models, two logit, a LASSO and a Random Forrest model, each of them treated with cross validation, in order to avoid overfitting. We compared the models by area under the curve, RMSE and expected loss. As a last step of the analysis we evaluated the best model seperately for manufacturing and services.

We reference code snippets from our Jupyter notebook for details on cleaning, variable creation, model training, and classification threshold selection.

2. Methodology

2.1 Data Preparation

In the original dataset we had around 288 thousand observation for 2010-2015. We decided to only examine sales growth between 2012 and 2013, which resulted in having around 15 thousand observations. We did lose many observatons, but our sample is still big enough to use it for comprehensive analysis. We decided not to use variables with many missing values, these were costs of sold goods, net domestic sales. There were some obvious mistakes in the dataset, so we setted negative asset entries to zero and flagged variables which seemes error as well, for example negative intangible assets. For missing CEO age we imputed the mean. We divided profit/loss variables by sales and balance sheet items by total assets, resulting in more robust variables, which are easier to compare.

For the target variable, we created a dummy that is one if the sales growth is at least 20% and 0 if not. That resulted in having 26% fast growing identified companies (3,810 out of 14,689). To avoid overfitting we splitted the dataset into a train set (80% of the observations) and a test set (20% of the observations). We performed cross validation with 5 folds on the train set.

2.2 Model Building

We created 2 logit models, one that includes the variables that seemed signficant, in the form that seemed adequate based on lowess and a second which included all variables in a linear way. For LASSO we also included potential interactions as well (e.g., curr_liab_bs * liq_assets_bs) and found that 20 coefficients remain non-zero at the optimal lambda (out of 23). For random forrest we optimized the parameters by grid search, trying different values maximum number of features (1 to 5) and minimum saple split (80 to 100 with 5 steps) and decided to have 500 trees. For max features it turned out that 5 is the best parameter and for minimum sample split it is 90.

2.3 Model comparison

We compared the models goodness by AUC and RMSE as well, and random forrest outperforms the other models by both of the metrics. The area under the curve is 0.66 which means that our model has a moderate explanatory power.

2.4 Classification & Cost-Sensitive Threshold

After probability prediction, we must set a classification threshold to balance false positives vs. false negatives. For the loss function we estimated a penalty for false positives and false negatives, which are 1 and 10. The code calculates an optimal threshold by scanning ROC curve points, aiming to minimize expected loss. For example, a threshold around 0.06 yields near-perfect recall, ensuring very few missed fast-growers at the expense of more false alarms.

2.4 Industry-Specific model evaluation

The manufacturing industries contained almost 5 thousand observations that was categorized as fast growing (29%) and services contained almost 10 thousand (24%). We re-ran the random forrest model with the same cost ratio, computed optimal thresholds, and recorded AUC, expected loss, and confusion metrics for each sector.

3. Results

3.1 Overall Model Comparison

MODEL	FEATURES	CV AUC	CV RMSE	AVG. CV EXPECTED LOSS
M1 (LOGIT WITH SPLINES)	26	0.637	0.436	~0.68
M2 (SIMPLE LOGIT)	20	0.629	0.437	~0.69
LASSO	20	0.632	0.436	~0.68
RF	18	0.660	0.431	~0.68

Random Forest leads in AUC (0.66) and RMSE (0.431) as well. Logits show decent performance (AUC ~0.63), but less flexibility for non-linear patterns. All produce similar expected loss estimates (0.68–0.69), but closer inspection shows RF has more consistent and robust classification across folds.

Holdout performance for Random Forrest model, out best performing model: Area under the curve is 0.65 and RMSE = 0.434, which means that our model is stable and is not overfitted. Confusion matrix at threshold 0.06 yields near-perfect recall, albeit with many false positives (precision ~0.33).

3.2 Sector-Specific Analysis

Using RF with cost ratio FN=10, FP=1:

INDUSTRY	EXPECTED LOSS	AUC	ACCURACY	PRECISION	RECALL	F1	BASELINE LOSS	IMPROVEMENT
MANUFACTURING	0.68	0.69	0.3252	0.301	0.998	0.46	2.90	2.22
SERVICES	0.68	0.68	0.3225	0.265	0.998	0.41	2.44	1.76

Manufacturing yields slightly higher AUC (0.691 vs. 0.683) and better F1. Both achieve recall ~0.998 at threshold 0.06, significantly reducing costly misses. Expected loss falls from ~2.90 (Manufacturing) and 2.44 (Services) to ~0.68, a major improvement given FN=10.

4. Conclusion

Splines capture non-linearities but can overfit if not well-chosen. LASSO helps reduce feature clutter, retaining 20 significant coefficients. However, Random Forest consistently delivered superior CV metrics and adaptability across industries.

The high FN cost (10× vs. FP) drives the optimal threshold down to ~0.06, ensuring near-perfect recall. Precision (~0.30) is relatively low; we accept more false positives to avoid missing truly fast-growers.

Manufacturing's slightly better AUC suggests more stable drivers of growth (e.g., labor or profit margins). Services are more heterogeneous, thus slightly lower precision and AUC.

AUC of 0.69 leaves scope for improvement: adding external data (market/industry trends) or refining cost ratios (e.g., FN=5) could reduce false alarms. Over-prediction can strain follow-up resources; firms flagged "fast-growing" need further due diligence.