Summary Report: Predictive Modeling of High-Growth Firms

Assignment 2 GitHub repo

Borbala Kovacs, Muheng Li

February 15, 2026

1 Data Processing

The objective was to predict sustained revenue growth while accounting for firm survival. We developed a predictive model to identify fast-growing firms using the *bisnode-firms* dataset. The goal is to support strategic decision-making by flagging high-potential companies for investment or resource allocation. To ensure the model is rigorous, replicable, and aligned with business objectives, we made several foundational design choices—outlined below—each justified by statistical best practices and domain knowledge.

Table 1: Summary of key design decisions

Decision	Choice	Rationale
Growth horizon	2 years (2012→2014)	Smooths volatility, captures sustained growth
Growth measure	Log sales growth	Standard in corporate finance, symmetric treatment of gains/losses
Fast-growth threshold	75th percentile of survivor growth	Data-driven, captures top quartile performers
Exiting firms	Coded as `fast_growth = 0`	Avoids selection bias; exits are not growth
Features	Same as exit-prediction pipeline	Same financial drivers, different direction of effect
Sample	Alive in 2012, sales 1K–10M	Consistent with seminar; excludes micro-firms and large corporates

2 Task 1: Predicting

2.1 Probability prediction

2.1.1 Model Specification

We define three logit models of increasing complexity, a LASSO, and a random forest:

Table 2: Model specifications for fast-growth firm prediction

Model	Description
M1	Baseline: log sales + sales growth + profitability + industry
M2	M1 + balance-sheet ratios + age + foreign management
M3	Full: all financial ratios + growth + HR + firm demographics + interactions
LASSO	Same variable set as M3, with L1 regularisation (automatic variable selection)
RF	Random forest on raw + engineered variables (no formula interactions needed)

2.1.2 Model Selection:

Table 3: Cross-validation performance metrics for fast-growth prediction models

Model	CV RMSE	CV AUC
M1	0.3878	0.6080
M2	0.3826	0.6500
M3	0.3800	0.6695
LASSO	0.3790	0.6724
RF	0.3759	0.6855

Based on the CV tables, Random Forest (RF) is the clear top performer overall. It achieves the lowest CV RMSE, meaning it produces the most accurate probability predictions in terms of calibration/error. RF also delivers the highest CV AUC values, indicating the strongest ability to discriminate between fast-growing and non–fast-growing firms.

Model choice for holdout evaluation: Therefore, RF is the best candidate to carry forward for holdout evaluation.

2.2 Classification

Table 4: Cross-validation performance and decision metrics for fast-growth prediction models

Model	CV RMSE	CV AUC	Avg optimal threshold	Avg expected loss
M1	0.3878	0.6080	0.3329	0.3717
M2	0.3826	0.6500	0.3505	0.3597
M3	0.3800	0.6695	0.3449	0.3469
LASSO	0.3790	0.6724	0.3569	0.3443
RF	0.3759	0.6855	0.3596	0.3410

The optimal decision thresholds are fairly similar across models, ranging from about 0.33 to 0.36, with RF using an average threshold of 0.3596. All thresholds are well below 0.5, which is expected in a cost-sensitive setting: because false positives are more expensive than false negatives (FP > FN), the classifier becomes more conservative and requires a higher predicted probability before labeling a firm as fast-growing.
Compared to the logistic models, RF not only improves RMSE and AUC but also delivers the lowest loss, suggesting that its ability to capture nonlinear patterns translates into better decisions when prediction errors carry asymmetric costs.

2.3 Defining the Loss Function

We frame the prediction problem as an investor/lender screening tool: a financial institution wants to identify firms with high growth potential, either to target them for equity investment, credit expansion, or partnership opportunities.
For the full sample analysis we set **FP = 1** and **FN = 2**, meaning that missing a fast-growing firm is twice as costly as wasting resources on a non-grower. This is a conservative asymmetry — in practice, growth-equity investors often face much higher FN/FP ratios (e.g. 5:1 or 10:1), but we keep it moderate for now because:

- Our ”fast growth” is top-quartile, not unicorn-level, so the upside of any single winner is bounded.
- FP costs are not negligible for a bank (misallocated credit lines carry risk).
- A 2:1 ratio is enough to meaningfully shift the optimal threshold below 0.5 and illustrate the classification trade-off.

2.4 Discussion of Results

2.4.1 Confusion Table on Holdout Set

Table 5: Confusion matrix for fast-growth prediction model (test set)

	Predicted: No fast growth	Predicted: Fast growth
Actual: No fast growth	2881	187
Actual: Fast growth	596	144

2.4.2 Model Performance

Cross-validation results were stable across folds, with Random Forest (RF) achieving the best performance (RMSE ≈ 0.376, AUC ≈ 0.686). Holdout set testing confirmed good generalization with no severe overfitting.

RF outperformed all other models (LASSO only marginally improved over logit model M3), as it captures nonlinear relationships and interactions missed by linear models.

The classifier is conservative: it has high specificity (94%) but low sensitivity (19%, detecting only 1 in 5 fast-growing firms), with precision of 44% (half of flagged firms are true fast growers).

Practically, the model works as an initial screening tool for shortlisting high-growth candidates, but is limited by low fast-growth event rates, missing key growth drivers (e.g., innovation, market conditions), and a bias toward avoiding false positives.

3 Sector-Specific Prediction

3.1 Results(Random Forest)

Table 6: Model performance by sector

Sector	AUC	Threshold	Test Avg Loss
Manufacturing	0.626	0.420	0.6741
Services	0.672	0.460	0.6715

Commentary: The Services sector exhibits higher predictability (AUC 0.672) compared to Manufacturing (AUC 0.626). This suggests that service-based growth is more linearly related to standard financial metrics (likely due to lower capital intensity). Despite the difference in raw predictive power, the optimized thresholds (0.420 vs 0.460) successfully equalized the business risk, resulting in nearly identical average loss scores (≈ 0.67) for both sectors.

3.2 Introducing a New Loss Function

We apply a new loss function for the second part of the analysis. We take an even more conservative stance and introduce a 5:1 loss function. This loss function mimics real life stakes better and it will be able to show differences in prediction of the manufacturing and services sectors more illustratively.

3.3 Manufacturing Analysis

3.3.1 Confusion Matrix (Test Set)

	Predicted: No fast growth	Predicted: Fast growth
Actual: No fast growth	240	361
Actual: Fast growth	30	127

Assessment: The asymmetric loss function ($Cost_{FN} = 5$) forced the model to cast a "wide net."

- **Recall Priority:** The model successfully identified 127 of 157 true high-growth firms (Recall $\approx 81\%$), missing only 30.
- **Screening Cost:** This high recall comes at the expense of 361 False Positives. In operational terms, the investment team must audit ≈ 3 false leads to find 1 true high-growth firm.

3.3.2 Top 5 Drivers

Driver	Importance
age	0.283741
d1_sales_mil_log	0.282239
sales_mil_log	0.250923
ind2_cat	0.077625
ceo_count	0.024400

Interpretation: Manufacturing growth is driven by a balanced mix of **Maturity** (Age), **Momentum** (Lagged Growth), and **Scale** (Log Sales). The high importance of '*d1_sales_mil_log*'

(0.28) indicates a "hot hand" effect—manufacturers already growing are the most likely to continue growing.

3.4 Services Analysis

3.4.1 Confusion Matrix (Test Set)

	Predicted: No fast growth	Predicted: Fast growth
Actual: No fast growth	1365	1113
Actual: Fast growth	187	385

Assessment: The Services model follows the same strategic pattern as Manufacturing but on a larger scale. The ratio of False Positives to True Positives remains approximately 2.9:1, confirming that the calibration logic holds consistent across different sample sizes.

3.4.2 Top 5 Drivers

Driver	Importance
age	0.370706
sales_mil_log	0.310829
d1_sales_mil_log	0.174474
ind2_cat	0.065455
gender_m_male	0.014296

Interpretation: A distinct structural difference appears here. In Services, **Age** (0.37) and **Size** (0.31) are overwhelmingly dominant, while Momentum (Lagged Growth) drops to third place (0.17). *Key Insight:* Unlike manufacturing, where recent growth predicts future growth, service firms appear to rely more on established market presence (Age) and critical mass (Size) to achieve the ”Fast Growth” designation.

3.5 Comparison

The Random Forest analysis reveals that the Services sector is inherently more predictable (AUC 0.67) than Manufacturing (AUC 0.63), likely due to a more direct linear relationship between financial inputs and revenue scaling in service industries. Despite this disparity in raw predictive power, our rigorous threshold optimization successfully equalized the economic outcome, resulting in a nearly identical average business loss (≈ 0.67) for both groups. The model adapts its strategy to the sector's nature: it utilizes momentum (past growth) to identify manufacturers but relies on structural maturity (age and size) to target service firms, ultimately satisfying the aggressive ”missed opportunity” penalty by voluntarily accepting a higher false-positive rate in the harder-to-predict manufacturing sector.

4 Executive Summary

We developed a predictive model to identify high-growth firms (top quartile revenue growth over 2012-2014) using Random Forest classification, which outperformed logistic regression and LASSO with an AUC of 0.686 and optimal expected loss of 0.341. The model demonstrates high specificity (94%) but low sensitivity (19%), functioning effectively as a conservative screening tool that minimizes false positives at the cost of missing many true fast-growers. Sector-specific analysis reveals that services firms are more predictable (AUC 0.672) than manufacturing firms (AUC 0.626), with services growth driven primarily by firm maturity and size, while manufacturing growth depends more heavily on recent momentum and past sales growth. When calibrated with an aggressive 5:1 loss function that heavily penalizes missed opportunities, the model achieves 81% sensitivity in identifying true fast-growers, though this requires screening approximately 3 false positives for every true positive. Despite differences in underlying predictability, optimized threshold selection successfully equalizes business risk across sectors, yielding nearly identical expected losses (0.67) for both manufacturing and services predictions.