Finding Fast-Growing Firms

DA3 Assignment 2 (Manager Summary)

Elene Zuroshvili and Luciana Razuri Solari

2026

1 Executive summary

We build a screening model that assigns each firm a *probability of fast growth*. The goal is not only to predict well, but to support a decision rule that reflects business priorities: missing a truly fast-growing firm is substantially more costly than flagging a slow-growing one.

Using Bisnode firm-level data (2010–2015) and an 80/20 train–holdout split, we compare five increasingly rich logistic regression specifications (M1–M5), a LASSO-regularized logistic model, and a tuned Random Forest. Models are compared using (i) **probability accuracy** (Brier RMSE), (ii) **ranking ability** (AUC), and (iii) **decision performance** via an asymmetric loss function (FN cost = 10, FP cost = 1).

Recommendation: The Random Forest is the preferred model for operational screening because it achieves the **lowest cross-validated expected loss** under the 10:1 FN:FP cost ratio, while also providing the best probability accuracy and ranking performance.

2 Problem setup and target definition

Fast growth is defined over a one-year horizon. A firm is labeled as fast-growing if its sales growth from year $t-1$ to t exceeds 50%. The model output is a probability score that can be used to rank firms for further screening, outreach, or investment analysis. This specific threshold is chosen based on the following corporate finance concepts:

- **Internal Growth Rate (IGR)**: A 50% growth rate likely exceeds the IGR of most firms, identifying companies that must rely on external financing (debt or equity) to sustain operations.
- **Information Asymmetry**: Identifying these "Gazelles" is critical for lenders. Rapid growth often creates a temporary "cash burn" despite long-term value, requiring sophisticated predictive models to distinguish high-potential firms from those at risk of insolvency.
- **Growth Options**: High-growth firms represent high real-option value. From a strategic management perspective, identifying these firms early allows for targeted investment before market valuations adjust.

We use a one-year sales growth definition because it aligns with short-horizon screening decisions where the objective is to identify firms that are currently scaling. From a corporate finance perspective, sales growth is closely related to market expansion and operational scaling, and it is often a leading signal that drives financing needs, investment opportunities, and strategic resource allocation.

Alternative definitions are possible. A two-year growth window would smooth short-term volatility and may better capture sustained growth, but it can delay identification of rapidly

expanding firms. Other outcomes such as asset growth or employment growth would capture investment and hiring dynamics, but they may lag behind revenue expansion and can vary systematically across industries due to differences in capital intensity. Given the goal of early identification for screening, a one-year sales-based definition represents a practical compromise.

3 Data and features (high level)

The dependent variable is the fast-growth indicator. The feature set combines:

- Firm size and recent dynamics (log sales, change in sales)
- Balance sheet structure and profitability components
- Firm age and entry indicators
- Management characteristics (e.g., CEO age, gender, foreign management)
- Accounting quality and missingness flags
- Controls for industry, region, and urbanization (via one-hot encoding)

The train–holdout split preserves the underlying class distribution: fast-growing firms represent about 21% of observations in the total sample (0.210), training set (0.210), and holdout set (0.213).

4 Model candidates

We evaluate three model families:

- **Logistic regression (M1–M5):** interpretable probability models with increasing feature richness (from a small core set to a large specification with interactions).
- **LASSO logistic regression:** a regularized variant that performs feature selection in high-dimensional settings.
- **Random Forest:** a flexible probability model that captures nonlinearities and interactions; tuned via cross-validation.

5 How we evaluate models

We separate **prediction quality** from **decision quality**:

- **Brier RMSE (probability RMSE):** average error of predicted probabilities (lower is better).
- **AUC:** how well the model ranks fast-growing firms above non-fast-growing firms (higher is better).
- **Expected loss:** the metric used for final model choice, based on business costs:

$$\text{Loss} = 1 \cdot 1(\text{false positive}) + 10 \cdot 1(\text{false negative})$$

We choose a classification threshold that minimizes expected loss (rather than defaulting to 0.5), and report cross-validated results and holdout performance.

6 Results and decision points

6.1 Decision 1: Does added complexity improve probability accuracy?

Table 1 shows Brier RMSE across the five CV folds for M1–M5. The results are stable across folds, and performance improves as we move from M1 to richer specifications, with M4–M5 achieving the lowest probability errors.

Table 1: Cross-validated Brier RMSE across folds (M1–M5)

Fold	M1	M2	M3	M4	M5
0	0.401409	0.396371	0.394639	0.392242	0.393152
1	0.407037	0.401621	0.399003	0.397369	0.396236
2	0.402871	0.397999	0.396171	0.393800	0.394784
3	0.400988	0.397664	0.394443	0.393851	0.393618
4	0.401564	0.395826	0.394008	0.392306	0.391965

6.2 Decision 2: Which model is best for the business objective (FN cost = 10)?

For operational screening, the key metric is **expected loss**. Table 2 compares all models on cross-validated Brier RMSE, AUC, and expected loss under the 10:1 FN:FP cost ratio. The Random Forest dominates the logistic models and LASSO on all three metrics and has the lowest expected loss, making it the recommended model.

Table 2: Model comparison (5-fold cross-validation)

Model	Brier RMSE (CV)	AUC (CV)	Optimal threshold (CV)	Expected loss (CV)
M1	0.403	0.596	0.123	0.784
M2	0.398	0.636	0.081	0.782
M3	0.396	0.652	0.084	0.782
M4	0.394	0.660	0.086	0.780
M5	0.394	0.664	0.077	0.771
LASSO	0.394	0.663	0.089	0.774
Random Forest	**0.391**	**0.673**	**0.100**	**0.759**

6.3 Interpreting the threshold

The optimal threshold for the Random Forest is approximately 0.10, substantially below the conventional 0.5 cutoff. This reflects the asymmetric cost structure: false negatives are ten times more expensive than false positives. Operationally, this means the screening rule intentionally flags more firms as potential fast growers to reduce the risk of missing truly fast-growing firms, accepting more false positives as the cheaper error.

6.4 Discrimination and decision threshold (plots)

Figure 1 shows the ROC curve for the Random Forest on the holdout set. The curve lies clearly above the diagonal, indicating meaningful discrimination between fast- and non-fast-growing firms. The corresponding AUC of approximately 0.67 confirms that the model provides substantially better ranking performance than random classification, even though the prediction task remains challenging.

Figure 2 shows the expected loss as a function of the classification threshold under the asymmetric cost structure (FN cost = 10, FP cost = 1). The loss-minimizing threshold is around

0.10, far below the conventional 0.5 cutoff. This reflects the high cost of false negatives: the decision rule intentionally flags more firms as potential fast growers to reduce the risk of missing truly fast-growing firms. Around the minimum, the loss curve is relatively flat, indicating that the chosen threshold is robust to small changes.

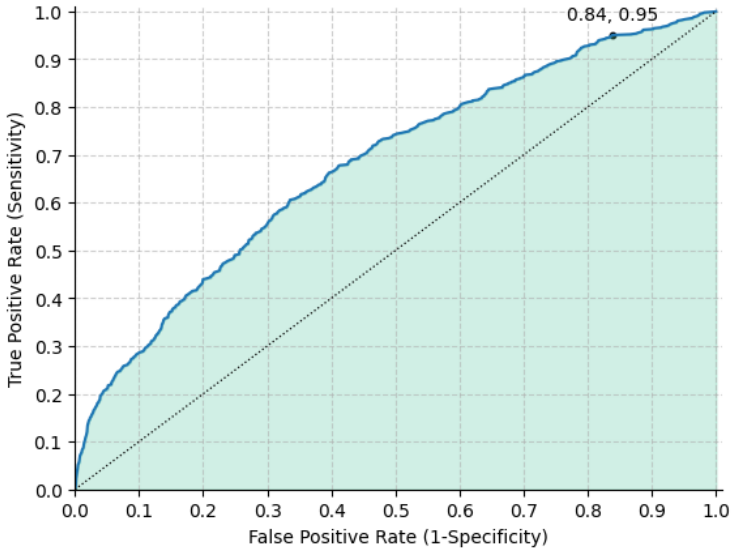

Figure 1: ROC curve for the Random Forest model on the holdout set.

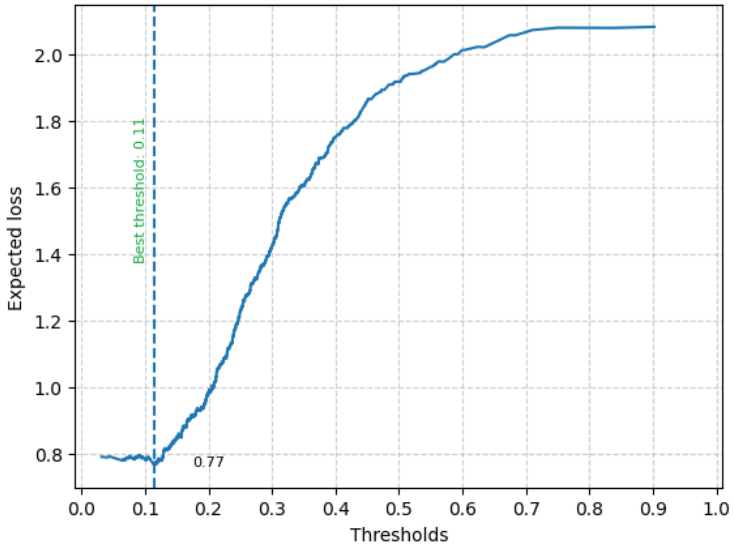

Figure 2: Expected loss as a function of the classification threshold (Random Forest). The vertical line indicates the loss-minimizing threshold under FN:FP = 10:1.

6.5 Holdout confusion matrix (discussion of classification)

Table 3 reports the confusion matrix for the Random Forest on the holdout set using the loss-minimizing threshold. The model correctly identifies most fast-growing firms (781 true positives) while missing only 29 of them (false negatives), which is important given that false negatives are assigned a much higher cost in the business objective. At the same time, the model produces a substantial number of false positives (2,651), meaning that many non-fast-growing firms are flagged as potential fast growers. This trade-off is intentional: under the 10:1

cost ratio, it is economically preferable to accept more false positives in order to avoid missing truly fast-growing firms. Overall, the confusion matrix confirms that the model's classification behavior is well aligned with the chosen loss function and the screening objective.

Table 3: Confusion matrix for the Random Forest model on the holdout set (using the optimal threshold).

	Predicted no fast growth	Predicted fast growth
Actual no fast growth	347	2,651
Actual fast growth	29	781

7 Industry-specific results (Task 2)

We repeat the same decision setup (FN:FP = 10:1) for Manufacturing (NACE 10–33) and Services (NACE 55–56, 95–96), using the Random Forest model. This allows us to assess whether fast growth is equally predictable across sectors and whether a single decision rule is appropriate in both cases.

Table 4: Industry-specific performance (Random Forest, holdout)

Industry	N	AUC	Optimal Threshold	Expected Loss
Manufacturing	5,562	0.554	0.021	0.754
Services	13,474	0.637	0.067	0.783

The results show substantial differences in predictability across sectors. The model achieves a much higher AUC for services firms (0.637) than for manufacturing firms (0.554), indicating that fast growth in services is more strongly related to observable firm characteristics such as size, recent dynamics, and management attributes. In contrast, growth in manufacturing appears harder to predict from these variables, likely reflecting the importance of lumpy investments, capacity constraints, and external demand shocks that are not fully captured by the available data.

The optimal classification thresholds also differ markedly across sectors. For manufacturing, the loss-minimizing threshold is extremely low (0.021), implying that the optimal strategy is to flag almost all firms as potential fast growers in order to avoid missing truly fast-growing firms. For services, the higher threshold (0.067) reflects the fact that the model provides more informative probability scores and allows for a more selective decision rule.

Finally, expected loss does not move one-to-one with AUC: although ranking performance is better in services, the expected loss is slightly higher than in manufacturing. This highlights that ranking quality and decision performance capture different aspects of model usefulness and reinforces the importance of aligning model evaluation with the underlying business objective rather than relying solely on standard accuracy measures. Overall, these findings suggest that sector-specific thresholds—or even separate sector-specific models—can improve decision quality in practical screening applications.

8 Conclusion

Predicting fast-growing firms is feasible, but model selection must be aligned with the underlying business objective rather than relying on default accuracy metrics or a 0.5 classification threshold. Under a 10:1 cost ratio for missing fast-growing firms versus false alarms, the Random Forest minimizes expected loss and is therefore recommended for operational screening.

In practical terms, the model should be used as a *screening and prioritization tool*: firms with higher predicted growth probabilities can be flagged for further review, outreach, or investment consideration. The low optimal threshold reflects a deliberate strategy to cast a wide net and avoid missing high-potential firms, accepting that this will generate more false positives that can be filtered out in subsequent, more detailed assessments.

Logistic regression models remain useful as transparent benchmarks and for economic interpretation, but they are consistently dominated by the Random Forest in terms of decision performance under the chosen cost structure. Finally, the strong differences between manufacturing and services firms suggest that a one-size-fits-all decision rule is suboptimal. In practice, sector-specific thresholds—or even separate sector-specific models—can further improve decision quality and better align the screening process with the economic realities of different industries.

Overall, the results show that data-driven screening tools can meaningfully support strategic and investment decisions, provided that model evaluation and deployment are explicitly tied to business costs and decision priorities.