Predicting Fast-Growing Firms with Bisnode Data

Technical Report

Elene Zuroshvili and Luciana Razuri Solari

2026

1 Introduction

This technical report documents the full modeling pipeline used to predict fast-growing firms using the Bisnode firm-level panel data. The purpose of this document is reproducibility and methodological transparency. In contrast to the managerial summary, the focus here is on the technical workflow: data construction, feature engineering, model specification, hyperparameter tuning, evaluation metrics, threshold optimization, and final model selection.

All analysis follows a strict train–holdout protocol. Model development and selection are performed exclusively on the training set using cross-validation. The holdout set is used only once for final evaluation.

2 Data and Sample Construction

The starting point is the Bisnode firm-level panel covering the period 2010–2015. The raw panel is processed using the provided data preparation pipeline, resulting in the cleaned dataset `bisnode_firms_clean_elene.csv`. The dataset contains balance sheet variables, profit and loss components, firm demographics, management characteristics, and several data quality flags.

The final sample is split into a training set (80%) and a holdout set (20%). A fixed random seed is used to ensure reproducibility. The class distribution is stable across splits: fast-growing firms represent approximately 21% of observations in the full sample (0.210), the training set (0.210), and the holdout set (0.213). This moderate class imbalance motivates the use of probability-based metrics and cost-sensitive evaluation rather than raw accuracy.

3 Construction of the Target Variable

The dependent variable, `fast_growth`, is constructed using a one-year growth definition based on sales. For each firm, sales growth is computed between year $t-1$ and year t. Firms whose growth exceeds 50% are labeled as fast-growing (value 1), while all others are labeled as 0.

The choice of a one-year horizon is driven by two considerations. First, many financial and strategic decisions (e.g., credit allocation, investment screening, acquisition targeting) are made on annual cycles. Second, a shorter horizon captures more immediate scaling dynamics and avoids smoothing effects that arise with longer growth windows. Alternative definitions, such as two-year growth or asset-based growth, were considered but would delay the identification of rapidly expanding firms.

4 Feature Engineering and Design Matrices

A large set of predictors is constructed from the cleaned dataset. These include:

- Firm size and dynamics (log sales, squared log sales, changes in sales),
- Balance sheet structure (assets, liabilities, equity, liquidity measures),
- Profit and loss components (inventories, material expenses, personnel expenses),
- Firm demographics (age, age squared, new firm indicator),
- Management characteristics (CEO age, gender, foreign management, management team size),
- Data quality and missingness flags.

Categorical controls for industry (NACE), region, and urbanization are encoded using dummy variables, with one category dropped in each group to avoid multicollinearity.

Several design matrices are constructed to allow controlled comparison across model complexity:

- **M1–M5:** Logistic regression models of increasing complexity.
- **LASSO set:** A large feature set including interactions, used with L1 regularization.
- **RF set:** A reduced set of numeric and categorical variables suitable for tree-based models.

This stepwise construction allows us to study how predictive performance evolves as model flexibility increases.

5 Models

Three classes of models are estimated:

1. **Logistic regression (M1–M5):** Used as transparent baseline probability models with increasing feature richness.
2. **LASSO-regularized logistic regression:** Used to handle high dimensionality and perform automatic feature selection. All features are standardized before fitting, which is required for regularization.
3. **Random Forest:** A non-linear, tree-based ensemble model used to capture interactions and nonlinearities not accessible to linear models.

All models are evaluated using 5-fold cross-validation on the training set.

6 Evaluation Metrics

Model performance is evaluated along three dimensions:

- **Brier RMSE:** Measures the accuracy of predicted probabilities.
- **AUC:** Measures ranking performance independently of any classification threshold.
- **Expected loss:** A decision-oriented metric based on the asymmetric loss function:

$$\text{Loss} = \frac{1 \cdot FP + 10 \cdot FN}{N}.$$

This reflects a setting where missing a truly fast-growing firm (false negative) is ten times more costly than falsely flagging a slow-growing firm (false positive).

7 Cross-Validation Results

We first compare all candidate models using 5-fold cross-validation on the training set. Table 1 reports average performance across folds.

Table 1: Cross-validated model performance

Model	# Coefficients	CV RMSE	CV AUC	CV Expected Loss
M1	12	0.4028	0.5958	0.7838
M2	19	0.3979	0.6365	0.7823
M3	36	0.3957	0.6523	0.7820
M4	80	0.3939	0.6604	0.7798
M5	154	0.3940	0.6637	0.7708
LASSO	112	0.3938	0.6626	0.7737
RF	n.a.	**0.3911**	**0.6727**	**0.7587**

As model complexity increases, both RMSE and AUC improve, but with diminishing returns. The Random Forest achieves the best performance on all three metrics and is therefore retained as the leading candidate for decision-making.

8 LASSO Regularization

The LASSO model is trained on the largest feature set using a grid of penalty parameters. Features are standardized prior to estimation. The selected model retains 112 non-zero coefficients. Its performance is similar to M5 but slightly worse than the Random Forest, indicating that non-linearities and interactions captured by the forest add predictive value beyond linear feature selection.

9 Random Forest Hyperparameter Tuning

The Random Forest is tuned using grid search over:

- `max_features` $\in \{5, 6, 7\}$,
- `min_samples_split` $\in \{11, 16\}$.

Table 2: Random Forest grid search (cross-validation)

max_features	min_samples_split	CV AUC	CV RMSE
5	11	0.6721	0.3912
5	16	**0.6727**	**0.3911**
6	11	0.6713	0.3914
6	16	0.6725	0.3912
7	11	0.6705	0.3915
7	16	0.6724	0.3913

The selected configuration is `criterion = gini`, `max_features = 5`, and `min_samples_split = 16`.

10 Threshold Optimization and Expected Loss

For each model, the classification threshold is chosen to minimize expected loss under the asymmetric cost structure. Table 3 reports the average optimal thresholds and losses across folds.

Table 3: Threshold optimization and expected loss

Model	Avg. optimal threshold	Avg. expected loss	Fold 5 loss
M1	0.123	0.784	0.786
M2	0.081	0.782	0.772
M3	0.084	0.782	0.782
M4	0.086	0.780	0.784
M5	0.077	0.771	0.779
LASSO	0.089	0.774	0.779
RF	0.100	**0.759**	0.766

11 Discrimination and Threshold Choice

Figure 1 shows the ROC curve for the Random Forest on the holdout set, illustrating its ranking performance. Figure 2 shows expected loss as a function of the classification threshold, with a clear minimum around 0.10, far below the conventional 0.5 cutoff.

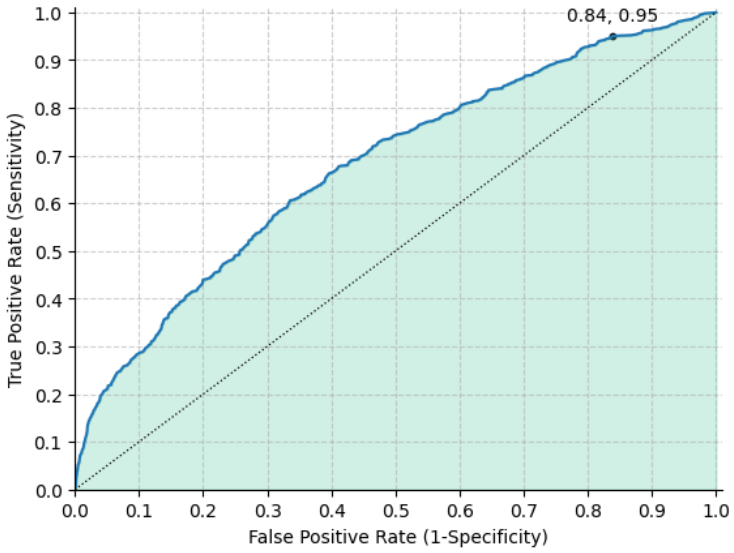

Figure 1: ROC curve for the Random Forest on the holdout set.

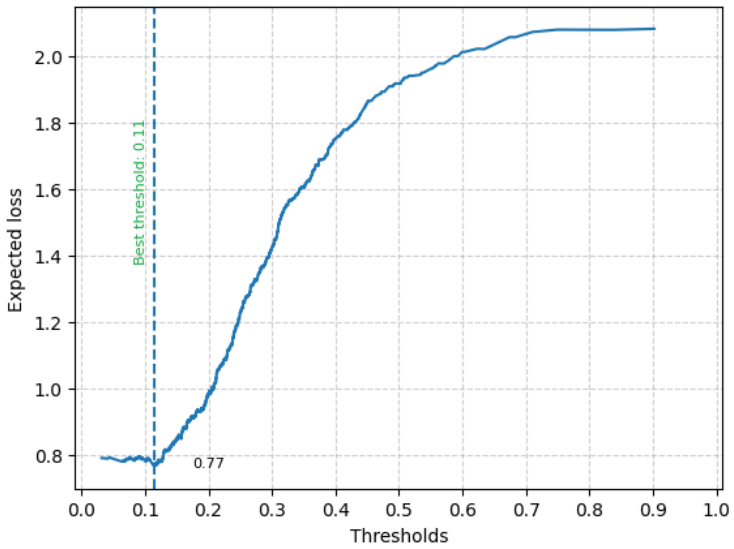

Figure 2: Expected loss as a function of the threshold (Random Forest).

12 Holdout Performance and Confusion Matrix

On the holdout set, the Random Forest achieves:

- RMSE = 0.395,
- AUC = 0.664,
- Expected loss = 0.772.

Using the loss-minimizing threshold, the confusion matrix is shown in Table 4.

Table 4: Confusion matrix (Random Forest, holdout)

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

13 Industry-Specific Analysis (Task 2)

The same pipeline is applied separately to Manufacturing (NACE 10–33) and Services (NACE 55–56, 95–96) using the Random Forest and the same loss function.

Table 5: Industry-specific results (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

14 Discussion and Limitations

The Random Forest dominates linear benchmarks in decision-oriented performance but is less interpretable. Results depend on the chosen loss function, and different business contexts would

imply different thresholds. Predictive performance in manufacturing remains limited, suggesting that important structural drivers of growth may not be fully captured by the available variables.

15 Conclusion

This report documents a full, reproducible pipeline for predicting fast-growing firms using firm-level data. The analysis highlights the importance of cross-validation, hyperparameter tuning, cost-sensitive evaluation, and threshold optimization. Under the chosen loss function, the Random Forest achieves the lowest expected loss and is therefore selected as the final model.