Predicting Fast-Growing Firms

Evidence from Bisnode Firm-Level Panel Data (2010–2015)

Business motivation and objective

Identifying fast-growing firms is a central problem in corporate finance and applied data science. Fast growth firms disproportionately contribute to job creation, innovation, and economic dynamism, yet they are rare and difficult to predict ex ante. From a managerial perspective, early identification of such firms supports better capital allocation, targeted policy interventions, and portfolio decisions.

The objective of this project is to develop and evaluate predictive models for identifying fast-growing firms using firm-level accounting and demographic data from the Bisnode panel. The focus is on **probability prediction and cost-sensitive classification**, reflecting realistic business decision-making rather than pure statistical accuracy.

Data and sample design

Data source

We use the Bisnode firm-level panel dataset covering the years **2010–2015.** The raw panel is transformed using a data preparation pipeline.

Data cleaning and transformation

Financial variables are cleaned and transformed following standard practices in empirical corporate finance:

- Monetary variables are scaled using balance sheet or income statement totals
- Log-transformations are applied to size-related variables to reduce skewness
- Extreme values are handled using winsorization
- Indicator flags are created to capture problematic observations (e.g. zero, extreme, or inconsistent values)

These steps help stabilize model estimation while retaining economically meaningful variation.

Sample construction

- Balanced firm-year panel constructed for 2010–2015
- Financial variables cleaned and standardized
- Accounting ratios expressed relative to balance sheet or profit-and-loss totals
- Outliers handled via winsorization and indicator flags
- Industry categories collapsed into broad groups (manufacturing vs services)

The final cleaned dataset (bisnode_firms_clean_growth.csv) contains approximately **X firms** and **Y firm-year observations**, with 118 engineered features.

Target definition: fast growth

Conceptual considerations

Defining "fast growth" is one of the most important intellectual steps in this project. Growth can be measured in many ways, and each choice has implications for interpretation and prediction.

From a corporate finance perspective, growth should:

- Be comparable across firms of different sizes
- Reflect sustained expansion rather than temporary shocks
- Be observable and reliably measured

Baseline definition

A firm is classified as **fast-growing** if it belongs to the **top 20% of the two-year sales growth distribution**:

$$\text{Growth}_i = \log\left(\text{Sales}_{i,2014}\right) - \log\left(\text{Sales}_{i,2012}\right)$$
$$\text{FastGrowth}_i = \mathbf{1}(\text{Growth}_i \geq P_{80})$$

This definition balances:

- robustness to short-term noise,
- comparability across firm sizes,
- and alignment with empirical growth literature.

Alternative definitions considered

Several alternative growth definitions were considered:

1. **One-year growth (2013 vs 2012)**
   This measure is more timely but highly sensitive to transitory shocks.
2. **Absolute sales growth**
   While intuitive, this measure disproportionately selects large firms.
3. **Employment-based growth**
   Economically meaningful but affected by missing data and reporting inconsistencies.

Overall, the chosen two-year log sales growth measure provides the best balance between stability, interpretability, and data availability.

Feature engineering and selection

Predictors are grouped into economically meaningful blocks:

- **Firm size & age**: log sales, firm age, squared terms
- **Profitability**: profit-to-sales, income before tax ratios
- **Financial structure**: liquidity, leverage, equity shares
- **Growth history**: lagged sales growth
- **Human capital**: CEO age, gender composition, workforce size
- **Industry & location**: industry dummies, urban indicators

LOWESS diagnostics were used to assess non-linear relationships between key predictors (e.g., size, age, profitability) and the probability of fast growth. These diagnostics motivated the inclusion of quadratic terms and transformations for selected variables.

Models and evaluation strategy

Models estimated

Three models were trained and evaluated using 5-fold cross-validation:

1. **Logit model** - Provides a transparent and interpretable baseline.
2. **Regularized logit (L1/L2) -** Controls overfitting and stabilizes coefficient estimates.
3. **Random forest classifier -** Captures non-linearities and complex interactions**.**

Probability prediction performance

Model performance is evaluated using cross-validated metrics such as AUC and log-loss. Results show that:

- Logit models perform reasonably well but struggle with non-linear effects
- Regularization improves stability but only marginally improves performance
- The random forest consistently delivers the best predictive accuracy

Based on these results, the random forest is selected as the preferred probability model

Cost-sensitive classification

Business loss function

To reflect asymmetric decision costs:

- **False Positive (FP)**: costly misallocation of resources
- **False Negative (FN)**: missed high-growth opportunity

Loss function and threshold selection

A single loss function is defined, assigning different costs to false positives and false negatives. For each model:

- Predicted probabilities are converted into classifications
- Classification thresholds are optimized to minimize expected loss
- Expected loss is averaged across cross-validation folds

For each model:

- Probabilities are converted into classifications using the threshold minimizing expected loss.
- Expected loss is averaged across cross-validation folds.

Model selection

The **random forest model** achieves the lowest average expected loss and is selected as the preferred model, despite slightly lower interpretability compared to logit.

Classification results and interpretation

A confusion matrix is examined for a representative validation fold. The results indicate:

- A meaningful trade-off between sensitivity and specificity
- Reasonable detection of fast-growing firms given class imbalance
- Performance consistent with the chosen loss function

Overall, the model is best interpreted as a **screening tool** rather than a deterministic decision rule.

Industry-specific performance

Growth dynamics differ across industries. Applying a single model without industry validation may lead to misleading conclusions.

Manufacturing vs services

The classification exercise is repeated separately for manufacturing and service firms, using the same loss function and modeling approach.

Results show:

- Stronger predictive performance in manufacturing
- Weaker and noisier performance in services

This suggests that growth in services is more heterogeneous and harder to predict using standard financial variables.

Discussion and managerial implications

The results suggest that fast-growing firms can be predicted with economically meaningful accuracy using standard firm-level data. Size, profitability, and prior growth history are the most consistent predictors, while demographic variables play a secondary role.

From a managerial perspective:

- Probability-based targeting dominates simple rule-based screening.
- Cost-sensitive thresholds materially affect which firms are selected.
- Industry context should guide deployment and interpretation.

Overall, the model provides a practical tool for prioritizing firms under limited attention and resource constraints.

Conclusion

This project demonstrates how modern predictive methods, combined with careful target design and business-oriented evaluation, can support strategic decision-making in corporate finance contexts. Future work could incorporate longer horizons, alternative growth metrics, or dynamic updating as new data arrive.

Appendix (optional, if space allows)

- Confusion matrix (selected fold)
- Feature importance (random forest)
- Additional LOWESS plots