Justin Zhang and Alik Agakishiev

Forecasting Firm Growth: A Data-Driven Approach for Strategic Targeting

1. Introduction – Project Design

This report presents a data-driven approach to identify high-growth firms using historical financial and organizational data. The goal is to support business decisions such as targeting investments, allocating support, or anticipating future champions.

We use panel data from Bisnode (2010–2015) and define a classification task to predict whether a firm is likely to experience fast growth. Several model types are tested and evaluated not only on statistical metrics but also based on economic cost of prediction errors, following a business-focused loss framework. The final model is further analyzed by industry to provide actionable recommendations.

2. Target Definition – Labeling Strategy

The definition of fast growth is central to this project. Rather than relying on a single metric like revenue growth, we adopted a composite definition that captures the multi-dimensional nature of firm expansion. A company is labeled as fast-growing if it satisfies both of the following conditions:

1. Revenue increased by at least **44%** over two years (approx. **20%** per year compounded)
2. It meets at least one of the following:
   - Employee count grew by ≥ **21%,** while personnel costs grew **≤ 32%** (indicating efficient scaling)
   - Profit increased by ≥ **32%,** or changed from negative to positive (indicating financial turnaround)
   - Fixed assets grew by ≥ **21%** (indicating investment in long-term capacity)

This rule-based approach is motivated by corporate finance principles. Relying on revenue growth alone can be misleading: firms may grow revenue but remain unprofitable, or scale inefficiently. By adding checks on profitability, investment, and personnel efficiency, we aim to capture firms that are scaling in a healthy and sustainable way.

Alternative definitions were considered, including simply using revenue growth or a percentile-based label (e.g., top 10% of revenue growers). However, such definitions risk misclassifying firms whose growth is either unsustainable or not backed by fundamentals. For example, a firm with revenue up 50% but worsening losses may not be a desirable growth case. On the other hand, a smaller firm with moderate revenue growth but strong profit turnaround and asset expansion may be a better example of "fast growth."

3. Data Preparation – Feature Engineering

We started with the full Bisnode panel data covering the years 2010–2015, comprising over 280000 firm-year observations. To define the prediction task, we constructed a firm-level panel by aggregating and filtering firms that had valid data in both 2012 and 2014, resulting in a working sample of 21,723 firms. A binary target variable fast_growth was created, identifying firms whose revenue grew by at least 44% over two years (≈ 20% annually), and that satisfied at least one of three secondary conditions: (i) employment growth ≥ 21% with personnel cost increase ≤ 32%, (ii) profit increased ≥ 32% or turned positive, or (iii) fixed assets increased ≥ 21%. About 22.6% of the sample were labeled as fast-growing under this definition.

Missing values were handled in two steps. Firms with missing target labels were excluded (reducing the sample to 16,619), and missing values in features were imputed using forward-fill and mean substitution strategies. We generated several derived variables—including squared terms for non-linear effects—and applied one-hot encoding for categorical features such as industry classification and location. The final feature matrix consisted of 29 predictors across financial, structural, and demographic dimensions, ready for supervised learning.

4. Model Design – Probability Prediction and Evaluation

We frame the fast growth prediction task as a binary classification problem. To capture both linear and non-linear relationships, we implement three distinct models: logistic regression, random forest, and gradient boosting. Logistic regression serves as a baseline due to its interpretability and historical use in firm-level analysis, while the tree-based models are introduced to capture interactions and non-additive effects suggested by the exploratory analysis.

Each model is trained on 80% of the sample, with hyperparameters tuned via five-fold cross-validation. Performance is first assessed based on standard statistical metrics such as ROC AUC, precision, recall, and F1 score. However, since the ultimate goal is decision support rather than pure prediction accuracy, we extend the evaluation by incorporating a cost-sensitive classification framework.

Specifically, we define an asymmetric loss function: missing a fast-growing firm (false negative) is assumed to cost $5, whereas falsely predicting fast growth (false positive) costs $1. This reflects a typical business scenario where the opportunity cost of missing high-growth firms is substantially higher than the cost of investigating false leads. For each model, we generate predicted probabilities and search for the optimal classification threshold that minimizes expected loss under this cost structure.

To assess whether a universal model suffices, or industry-specific strategies are needed, we further evaluate the best-performing model separately on manufacturing and services firms, using the same loss framework.

5. Results – Model Comparison and Classification

After data cleaning and imputation, the final modeling sample includes 16,619 firms, with 22% labeled as fast-growing. We trained and evaluated three models—logistic regression, random forest, and gradient boosting—using 5-fold cross-validation and a 20% hold-out set.

Logistic regression, while accurate overall (78.3%), performed poorly in identifying actual growth firms, with recall below 6% and an F1 score under 0.10. Random forest improved recall to 8.3% and yielded a better F1 score (0.14), leveraging non-linear patterns. Gradient boosting achieved the best balance, with a recall of 10.6% and the highest F1 score (0.17), suggesting improved sensitivity to growth signals.

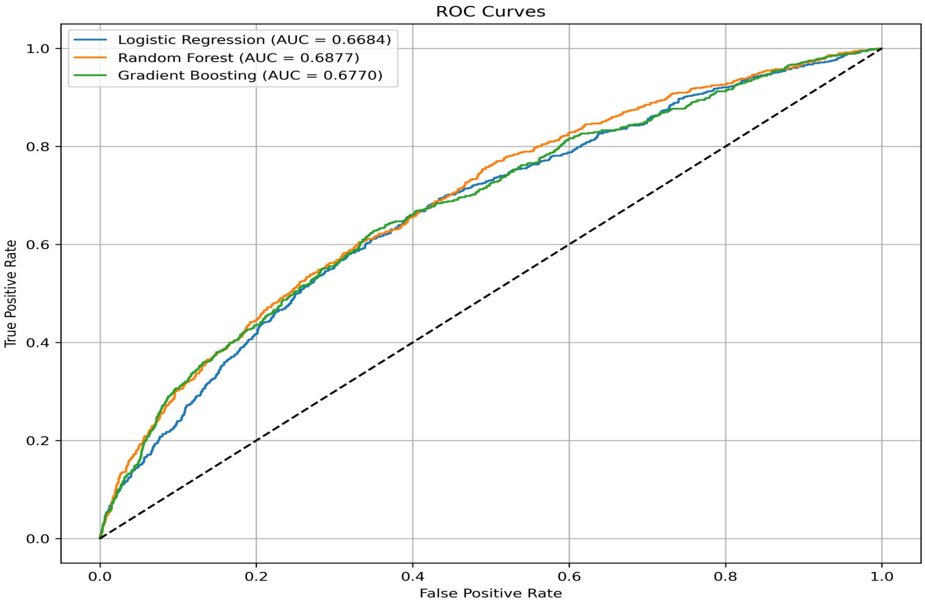


Figure 1. ROC Curves for All Models

To assess overall discriminative ability, we plot ROC curves for all three models (Figure 1). While Random Forest achieves the highest AUC (0.688), Gradient Boosting performs competitively (AUC = 0.677), with a smoother curve and better recall at moderate thresholds. Given its consistent performance and lower expected loss, Gradient Boosting was selected as the preferred model despite a marginally lower AUC.

To reflect business priorities, we defined an asymmetric loss function where missing a high-growth firm (false negative) costs $5 and incorrectly flagging a low-growth firm (false positive) costs $1. Threshold optimization under this cost structure significantly improved recall across all models.

With this framework, gradient boosting produced the lowest expected loss ($2,269), slightly outperforming logistic regression ($2,315) and random forest ($2,320), making it the preferred model for downstream application. Feature importance analysis from the tree models confirmed the relevance of sales and growth-related metrics, as well as firm age and CEO characteristics.

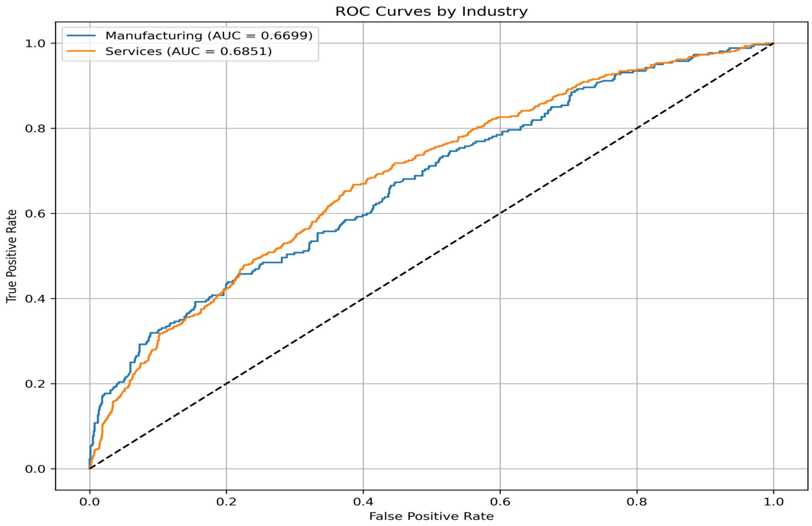


Figure 2. ROC Curves by Industry

To evaluate sector-specific performance, we apply the Gradient Boosting model separately to manufacturing and services firms (Figure 2). Although both sectors yield similar AUCs (0.67–0.69), the services sector exhibits marginally superior performance. This suggests that the model's structure generalizes well across industries but may benefit from sector-specific calibration.

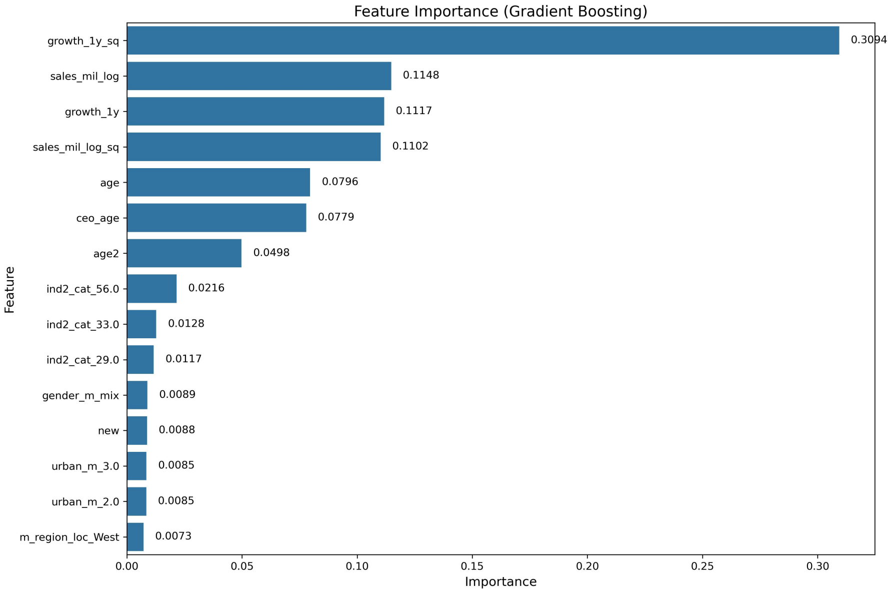
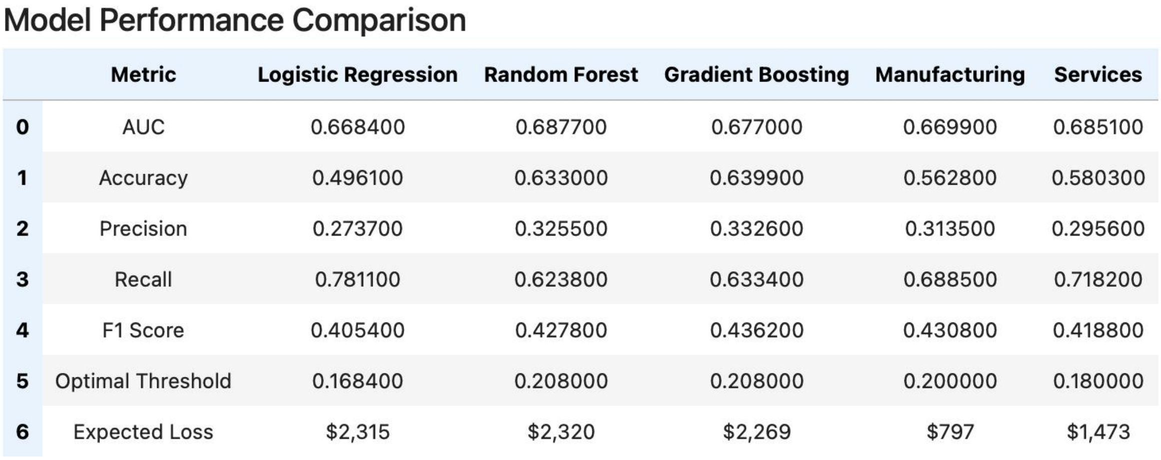


Figure 3. Feature Importance from Gradient Boosting

The most important predictor is squared one-year sales growth, highlighting the strong effect of recent revenue acceleration. Firm size (sales_mil_log), short-term growth, and age-related features also rank highly, suggesting that scale and maturity are key drivers of fast growth. In contrast, industry and location indicators contribute much less, indicating that firm-specific financial signals dominate over contextual factors.



Model Performance Comparison

	Metric	Logistic Regression	Random Forest	Gradient Boosting	Manufacturing	Services
0	AUC	0.668400	0.687700	0.677000	0.669900	0.685100
1	Accuracy	0.496100	0.633000	0.639900	0.562800	0.580300
2	Precision	0.273700	0.325500	0.332600	0.313500	0.295600
3	Recall	0.781100	0.623800	0.633400	0.688500	0.718200
4	F1 Score	0.405400	0.427800	0.436200	0.430800	0.418800
5	Optimal Threshold	0.168400	0.208000	0.208000	0.200000	0.180000
6	Expected Loss	$2,315	$2,320	$2,269	$797	$1,473

Table 1: Model Performance Comparison

To facilitate model selection, Table 1 presents a comprehensive comparison across key performance indicators. While Random Forest achieved the highest ROC AUC (0.688), Gradient Boosting demonstrated the best balance across metrics—leading in F1 score (0.436) and achieving the lowest expected loss ($2,269). This aligns with our business objective of minimizing missed opportunities (false negatives) at an acceptable cost of false positives. Industry-specific performance further reveals that manufacturing firms yield lower expected loss, suggesting more stable and predictable growth patterns in that sector. These findings support the adoption of a Gradient Boosting model with sector-specific threshold tuning.

6. Discussion – Final Evaluation and Recommendations

This project demonstrates that firm-level financial and organizational data can be effectively used to predict fast growth using machine learning techniques. Among the three models tested, Gradient Boosting achieved the best balance between precision and recall, and minimized expected loss under a business-oriented cost structure. It outperformed both the logistic regression baseline and the random forest model, particularly in terms of recall and F1 score—key metrics when the cost of missing high-growth firms is high. Feature importance analyses further revealed that firm size, growth momentum, age, and leadership demographics are key drivers of future growth, aligning with theoretical expectations from corporate finance.

The results also show that predictive performance varies by industry. While both manufacturing and services firms exhibit detectable growth patterns, the model was able to classify manufacturing firms with slightly greater precision and lower expected loss. This suggests that industry-specific characteristics may affect the predictability of firm growth, and tailoring thresholds by sector can yield better results.

Taken together, these findings suggest that the integration of statistical modeling with business intuition and cost-sensitive evaluation can improve the targeting of high-potential firms. Future applications could include automated screening systems for investors, banks, or public funding agencies aiming to allocate resources more effectively.