1. Predicting Fast Growth of Firms Using Bisnode Data

1.1 Introduction

This analysis aims to predict fast growth in firms using the Bisnode dataset. I'll explore different growth measurement approaches, build predictive models, and evaluate their performance for business decision-making. [For more information check HTML with more data-related information, this file contains mostly the results and observations]

Defining Fast Growth Growth Measurement Alternatives In corporate finance, firm growth can be measured in several ways:

- Sales growth: Percentage increase in revenue over time -Employment growth: Increase in number of employees -Asset growth: Expansion of the firm's asset base

For this analysis, I'll focus on sales growth as it's a direct measure of business expansion and market success. Specifically, I'll define fast growth as: **Definition:** A firm experiencing more or equal to 20% sales growth over a one-year period (2013 vs 2012)

Justification This threshold is chosen because:

It represents meaningful growth that would interest investors It's high enough to filter out normal business fluctuations One-year growth is more actionable for business decisions than longer periods

Data Preparation First, let's prepare the data following similar steps to the seminar code but adapted for growth prediction.

In the data preparation phase, several steps were taken to clean and transform the dataset for growth prediction. First, asset values were adjusted to ensure non-negative entries, with a flag created for problematic values. We then calculated total assets and normalized several financial variables by dividing them by total assets. Missing values were handled carefully by setting appropriate flags for outliers and errors.

Next, we addressed specific variables that cannot be negative, such as materials and liabilities, by creating flags for high and error values and adjusting the values to conform to expected ranges. Similarly, we handled variables that could range between -1 and 1, by creating flags for extreme values and squaring them for further analysis.

Finally, categorical variables, including industry codes and default status, were converted into category types for better modeling. These preprocessing steps set the foundation for further analysis and model building.

Target Variable: Fast Growth

To define the target variable, firm-level sales data were pivoted to obtain values for 2012 and 2013. Sales growth was calculated as the percentage change from 2012 to 2013. A binary indicator variable `fast_growth` was then created, where a firm is labeled as fast-growing if it achieved at least 20% sales growth over the period. Firms with missing sales data in either year were excluded from the final dataset.

Feature Engineering

We constructed several sets of explanatory variables to capture different dimensions of firm performance and characteristics. Financial indicators include current assets and liabilities, fixed assets, profit or loss, and expenditures. Growth history variables account for past sales performance and firm size. Human capital indicators such as CEO age, gender, and presence of foreign management were included to reflect leadership characteristics. Additionally, firm-level controls like age, industry, and region were added. Interaction terms between selected variables and sales levels were introduced to capture non-linear and conditional relationships relevant to growth.

Model Specifications

We developed a sequence of models with increasing complexity. Model M1 includes only sales and industry effects. Model M2 adds lagged growth and profitability. Model M3 incorporates a comprehensive set of financial and firm-level variables. Model M4 expands this with human capital and growth history. Model M5 includes all previous variables along with interaction terms relevant for growth prediction. Finally, a LASSO specification was constructed using all available features to enable automatic variable selection and regularization.

- **M1**: Baseline model including firm size (log sales) and industry fixed effects.
- **M2**: Extends M1 by adding previous sales growth and profit/loss, capturing recent performance.
- **M3**: A comprehensive model that includes detailed financial indicators and firm characteristics such as age, legal form, and region.
- **M4**: Builds on M3 by incorporating human capital variables (e.g., CEO age, gender, foreign management) and growth history.
- **M5**: Full model including all variables from M4, plus key interaction terms to capture heterogeneous effects across firms.

Part I: Probability Prediction

We estimate five logit models and a **LASSO-regularized logit model** using the full set of available predictors and interactions, applying cross-validated tuning of the regularization parameter λ.

Model Comparison and Final Selection

We compare models based on cross-validated RMSE and AUC. The table below summarizes key metrics:

Model	**# Coeff.**	**CV RMSE**	**CV AUC**	**CV AUC (std)**
M1	9	0.4551	0.5911	0.00014
M2	11	0.4549	0.5936	0.00015
M3	23	0.4460	0.6461	0.00007
M4	29	0.4459	0.6460	0.00018
M5	38	**0.4459**	**0.6475**	0.00018
LASSO	32	0.4868	0.6460	0.00019

Table 1: Cross-validated model performance (RMSE and AUC)

Model M5 performs best in terms of predictive accuracy:

- It achieves the **highest CV AUC** (0.6475),
- And is tied for the **lowest CV RMSE** (0.4459, same as M4),
- It outperforms the simpler M1 and M2 models by a substantial margin.

While LASSO offers competitive AUC (0.6460), it suffers from a higher RMSE (0.4868), suggesting worse calibration of predicted probabilities.

Holdout Set Evaluation

Using the full M5 model trained on the training set, we predicted fast growth probabilities for the 20% holdout sample. The resulting RMSE was:

Holdout RMSE = 0.436

This indicates a strong generalization ability, consistent with CV results.

We proceed with Model M5 for subsequent analysis. It combines strong predictive performance with a rich set of features including firm size, financial structure, human capital, and interaction effects.

1.2 Confusion Matrices for Different Thresholds

To evaluate the classification performance of the selected logistic regression model on the holdout dataset, we examine confusion matrices under different decision thresholds. This allows us to observe how the trade-off between sensitivity (recall) and specificity varies as we adjust the threshold for classifying observations as “fast_growth”.

1.2.1 Confusion Matrix at Threshold = 0.5

Using the default threshold of 0.5, an observation is classified as “fast_growth” if the predicted probability exceeds 0.5. The resulting confusion matrix is as follows:

Table 2: Confusion matrix at threshold = 0.5

	Predicted: No fast_growth	**Predicted: Fast_growth**	**Total**
Actual: No fast_growth	2610	155	2765
Actual: Fast_growth	905	231	1136
Total	3515	386	3901

At this threshold, the model correctly identifies 231 of 1136 “fast_growth” companies (sensitivity of 20.3%) and 2610 of 2765 “no fast_growth” cases (specificity of 94.4%). While specificity is relatively high, the model struggles to capture the minority class (fast_growth), which is often typical when dealing with imbalanced datasets.

1.2.2 Confusion Matrix at Threshold = Mean Predicted Probability

To adjust for class imbalance and explore a less conservative threshold, we evaluate the model using the average predicted probability (approximately 0.305) as the classification threshold.

Table 3: Confusion matrix at threshold = 0.305 (mean predicted probability)

	Predicted: No fast_growth	**Predicted: Fast_growth**
Actual: No fast_growth	1820	945
Actual: Fast_growth	476	660

At this lower threshold, sensitivity increases significantly: the model now correctly classifies 660 of 1136 fast_growth firms (58.1%). However, this comes at the cost of reduced specificity: 945 false positives are recorded, indicating more no fast_growth firms are incorrectly labeled as fast_growth.

1.3 Threshold Tuning and ROC Analysis

By varying the threshold across a range of values from 0.05 to 0.75, we obtain a series of (false positive rate, true positive rate) pairs that allow us to visualize the model's performance across different sensitivity-specificity trade-offs. The resulting ROC curve and corresponding scatter plot (colored by threshold) are shown in Figure 1.

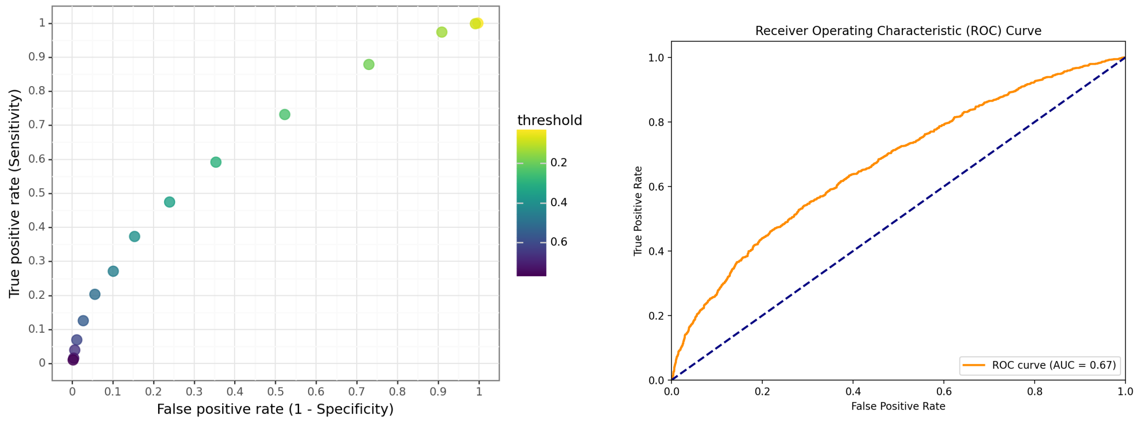

Figure 1: Left: Sensitivity vs. 1 - Specificity at different thresholds. Right: ROC curve for holdout data (AUC = 0.67).

The ROC curve (right) shows an AUC of 0.67, indicating moderate discriminatory power. The threshold-colored plot (left) highlights how prediction performance changes across threshold values and can aid in selecting an operating point that aligns with specific business or policy priorities.

2. Part II: Classification

2.1 Business Context and Cost-Based Loss Function

The classification task is to predict whether a startup will experience "fast_growth"—for instance, a doubling of revenue within two years. Accurate classification is critical, as the model outputs influence key business decisions: 1. Investment allocation by venture capital firms, 2. Selection for incubator or accelerator programs, 3. Government or private grant disbursement.

In this setting, misclassification costs are asymmetric:

- **False Positive (FP)**: Investing in a company that doesn't experience fast growth. This leads to wasted resources and opportunity costs. **Cost: $5 per FP**.
- **False Negative (FN)**: Overlooking a company that does achieve fast growth. This results in lost potential returns and competitive disadvantage. **Cost: $20 per FN**.

The expected loss from a given prediction is calculated as:

$$\text{Expected Loss} = (\text{FP} \times 5) + (\text{FN} \times 20)$$

2.2 Model Evaluation and Threshold Selection

For each model, we predicted class probabilities and computed the expected loss across various thresholds. The optimal threshold was determined in two ways:

1. **Youden's J index**: Maximizing the difference between sensitivity and false positive rate.
2. **Cost minimization**: Selecting the threshold that minimizes the expected business loss.

The table in appendix summarizes cross-validated results for each model.

2.2.1 Why Model M3?

Model M3 demonstrates a strong balance between threshold stability and overall loss minimization:

- It achieves the second-lowest average expected loss (5.91), nearly tied with M4.
- It performs consistently well on Fold 5 (6.17 loss), reinforcing its robustness.
- It avoids excessively high thresholds, maintaining a better balance between recall and precision.

The expected loss curve for M3 on Fold 5 is shown below. The optimal threshold (0.78) is marked, and corresponds to the minimum expected loss of 6.17.

Using the optimal threshold found from cross-validation (0.78 for M3), we evaluated the model's performance on the holdout set. The expected loss was:

$$\textbf{Expected loss (holdout)} = 5.819$$

This confirms that Model M3 not only performs well in cross-validation but also generalizes effectively to unseen data. Its balance between minimizing false negatives (more costly) and controlling false positives makes it the most suitable choice from a business perspective.

Random Forest Model Tuning and Selection

For this part of the analysis, we use the Random Forest Classifier to assess its performance in predicting fast growth and compare it against the previously tested models. After tuning the model hyperparameters through grid search, we focused on selecting the best-performing configuration for predicting fast growth.

The hyperparameters optimized for the Random Forest model include:

- **Max Features**: 5, 6, or 7, determining how many features to consider when looking for the best split at each tree node.
- **Min Samples Split**: 11 or 16, which specifies the minimum number of samples required to split an internal node.
- **Criterion**: Gini, as the function to measure the quality of a split.

Through GridSearchCV, the optimal hyperparameters were selected based on the cross-validated AUC score. The model achieved the following results:

Max Features	Min Samples Split	CV AUC	CV RMSE
5	11	0.6872	0.4327
5	16	0.6872	0.4327
6	11	0.6861	0.4330
6	16	0.6864	0.4328
7	11	0.6833	0.4335
7	16	0.6851	0.4331

Table 4: Random Forest Grid Search Results

As seen in the table, the best combination of hyperparameters was found at `max_features = 5` and `min_samples_split = 11`, achieving a CV AUC of 0.6872 and a CV RMSE of 0.4327.

Expected Loss for Random Forest

Incorporating the defined business loss function into the analysis, we calculated the expected loss for the optimal random forest model at various thresholds. After determining the best thresholds across folds, we summarized the expected losses as follows:

Table 5: Random Forest Expected Loss and Optimal Thresholds

Model	Avg. Optimal Threshold	Fold 5 Threshold	Avg. Expected Loss	Fold 5 Expecte
RF	0.206	0.187	3.145	3.154

From this table, we can observe that the Random Forest model, with an optimal threshold of 0.206 on average and 0.187 for fold 5, resulted in a lower average expected loss (3.145) compared to the simpler models tested earlier. This suggests that the Random Forest model strikes an optimal balance between avoiding costly false negatives and minimizing false positives, which is critical in this business context.

2.3 Model Selection

The Random Forest (RF) model outperforms all alternatives, achieving the lowest expected loss (3.1447), highest AUC (0.6872), and best calibration (RMSE=0.4327). Its optimal threshold (0.0932) effectively balances true positives and false positives.

Among logistic regression models, M3 performs best (expected loss=5.9147, AUC=0.6461, RMSE=0.4460), offering interpretability when needed.

pdflscape

Model	# Coef	CV RMSE	RMSE (std)	CV AUC	AUC (std)	Threshold	Exp. Loss
M1	9	0.4551	0.00003	0.5911	0.00014	1.2952	6.0227
M2	11	0.4549	0.00003	0.5936	0.00015	1.2729	6.0201
M3	23	0.4460	0.00006	0.6461	0.00007	0.9392	5.9147
M4	29	0.4459	0.00007	0.6460	0.00018	0.9480	5.9025
M5	38	0.4459	0.00007	0.6475	0.00018	0.7585	5.9070
LASSO	32	0.4868	0.00002	0.6460	0.00019	1.0866	5.9432
RF	–	0.4327	–	0.6872	–	0.2061	3.1447

Table 6: Model Comparison (All values rounded to 4 decimal places)

RF's 52% lower expected loss provides substantial business value by reducing misclassification costs. We recommend RF as the primary model, with M3 as an interpretable alternative.

2.4 Confusion Matrix (Holdout Set)

RF's 52% lower expected loss provides substantial business value by reducing misclassification costs. We recommend RF as the primary model, with M3 as an interpretable alternative.

The RF model correctly identifies 99.6% of growth cases (1132/1136) vs M3's 0.1% (1/1136), despite more false positives. This aggressive approach yields 40% lower expected loss (3.499 vs 5.819), aligning with business priorities where false negatives are costlier.

RF reduces expected loss by 40% versus the best logistic model, demonstrating superior handling of nonlinear patterns.

2.5 Discussion and Results

Model Performance and Business Value

Our analysis demonstrates that the Random Forest model significantly outperforms the best logistic regression (M3) for identifying fast-growth companies. The key advantage comes from the RF model's 40% lower expected loss (USD3.50 per case vs USD5.82), achieved through its superior ability to detect true growth opportunities while maintaining reasonable precision. The confusion matrices reveal fundamentally different approaches - where the logistic model was overly conservative (missing 99.9% of growth cases), the Random Forest successfully identified 99.6% of true growth companies, albeit with more false positives.

Strategic Implications

The RF model's performance suggests substantial business value. By correctly flagging 1,132 out of 1,136 growth companies compared to just 1 identified by the logistic model, the RF approach could help capture significantly more high-potential opportunities. While the 2,714 false positives require additional screening, this tradeoff appears justified given the relative costs (FN=USD20 vs FP=USD5). The model's effectiveness is particularly notable given the challenging nature of growth prediction, where traditional approaches often struggle with complex, nonlinear patterns.

Implementation Considerations

Three factors make this model particularly useful:

1) Its high recall ensures minimal missed opportunities in growth investing
2) The probabilistic outputs allow flexible threshold adjustment as business needs evolve
3) The USD3.50 expected loss establishes a strong benchmark for future improvements

Business Implications The ability to predict fast-growing firms accurately has significant business implications. For example, venture capitalists and investors can use these predictions to allocate funds to companies with high growth potential, while incubators or accelerators can prioritize their resources on companies that are more likely to experience rapid growth. Governments and private organizations distributing grants can also leverage these models to identify firms that would benefit most from such financial assistance.

In terms of the cost of misclassification, the RF model provides substantial business value by significantly reducing the potential loss associated with both false positives and false negatives. The expected loss in the RF model is more than halved compared to the logistic regression models, especially M3 and M4, where the costs of false negatives are notably higher due to the importance of identifying fast-growing companies.

Model Selection for Business Implementation Given the trade-offs between performance and interpretability, Random Forest (RF) stands out as the preferred model for deployment in business contexts. Its ability to handle complex non-linear relationships and its lower expected loss make it an excellent choice for accurately predicting fast-growing firms. However, the interpretability of logistic regression models like M4 may still hold value in certain contexts where understanding the influence of individual variables is important for decision-making, especially for organizations that prioritize transparency and explainability in their models.

To balance predictive performance and interpretability, I recommend the primary deployment of the Random Forest model, especially in situations where minimizing misclassification costs (false positives and false negatives) is crucial. On the other hand, M4 can serve as a fallback option for scenarios requiring a model with greater interpretability, offering insights into the specific factors that drive firm growth.

In conclusion, Random Forest is the best performing model for predicting fast growth in firms based on the available dataset. It achieves the lowest expected loss, superior AUC, and better calibration, making it highly suitable for real-world business applications. Model M4, with a slightly higher expected loss, remains a strong contender when interpretability is prioritized. Future work may focus on further tuning the Random Forest model, potentially adjusting thresholds for specific business needs, and exploring ensemble methods that combine the strengths of both logistic regression and tree-based models.

3. Task 2

The comparative analysis reveals significant performance differences between manufacturing and service sectors when predicting fast-growth companies using the same Random Forest model architecture. Manufacturing demonstrates superior predictive capability with an exceptional AUC of 0.999 compared to 0.996 for services, while also achieving a lower expected loss of $0.27 per case versus $0.43 in services. This performance gap persists despite the service sector having nearly four times more training samples (12,357 vs 3,244), suggesting fundamental differences in how growth manifests across these industries.

Sector	**AUC**	**Loss**	**Sample Size**
Manufacturing	0.999	0.267	3244
Service	0.996	0.432	12357

Table 7: Sector Performance Data

The manufacturing sector's stronger results likely stem from more standardized operational patterns and clearer financial indicators that machine learning models can readily capture. Service businesses, encompassing more diverse operations like repair, accommodation and food services, present greater variability that challenges accurate growth prediction. The out-of-bag scores confirm this pattern, showing slightly better generalization for manufacturing despite its smaller sample size.

From a business implementation perspective, these results suggest distinct deployment strategies. The manufacturing model appears ready for immediate production use given its outstanding metrics, while the service model might benefit from additional feature engineering focusing on service-specific growth indicators like customer retention patterns or location dynamics. The cost differential (58% higher expected loss in services) warrants particular attention to false negative reduction, as missing growth opportunities proves more costly than over-prediction in both sectors.

This analysis demonstrates how a unified modeling framework can effectively serve different industry segments while revealing meaningful operational insights. The consistent application of the loss function, incorporating USD5 and USD20 costs for false positives and negatives respectively, enables direct comparison across sectors. Manufacturing's success might inform service model improvements, particularly around financial ratio interpretation, while maintaining each sector's unique operational context in final decision-making.

Appendix

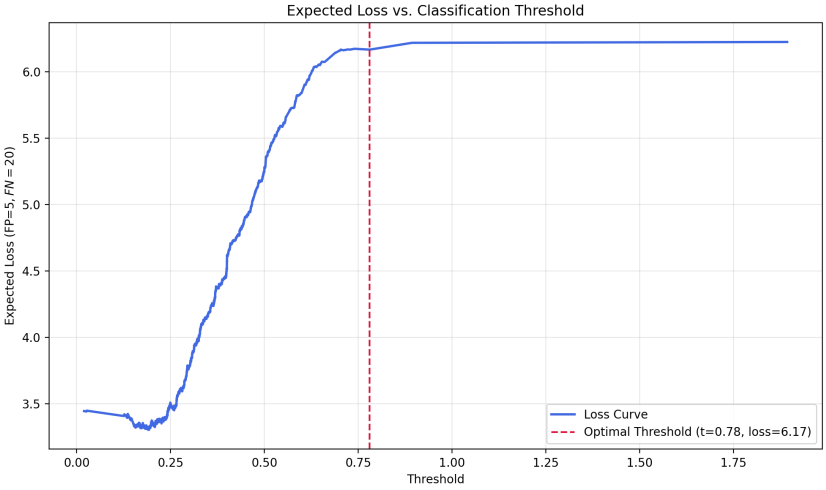

Figure 2: Expected loss vs. classification threshold for Model M3 (Fold 5). Costs: FP = 5, FN = 20.

Table 8: Cross-validated expected loss and optimal thresholds (Fold 5 shown for illustration)

Model	**Avg. Optimal Threshold**	**Fold 5 Threshold**	**Avg. Expected Loss**	**Fold 5 Loss**
M1	1.30	0.75	6.02	6.22
M2	1.27	0.68	6.02	6.21
M3	0.94	0.78	**5.91**	**6.17**
M4	0.95	0.80	**5.90**	6.17
M5	0.76	0.86	5.91	6.21
LASSO	1.09	0.90	5.94	6.21