Predicting Fast Growth in Firms

This report presents a detailed study on predicting fast-growing firms using the Bisnode-firms panel data (2011–2014). Our goal was to design a "fast growth" target, build multiple predictive models, and then select the best model according to both probability prediction performance and classification loss. We also demonstrate how the predictive exercise can be split by industry (manufacturing vs. services) to tailor decision-making.

I. Target Definition and Rationale

The target variable—fast growth—was defined using a two-year growth rate calculated from sales figures. Specifically, growth was computed as the relative change in sales between a baseline year (2012) and a later year (2014). The threshold is set at the 75th percentile of the growth distribution, which in our data corresponds to at least 50% growth in sales, so that firms with growth rates above this cutoff were labeled as fast-growing. This threshold was chosen because it gives a balance: it is high enough to capture firms that demonstrate significant, sustainable growth yet not so restrictive that the number of fast-growing firms becomes too small for reliable modeling. An alternative approach could have been to set the threshold even higher, targeting only the most explosive growth cases. However, doing so would result in a much smaller group of "slim fish" firms exhibiting extremely high growth, making it difficult for the predictive models to learn from and generalize to new data.

• **Comparability Across Firms:** By using a relative growth metric, we account for the size differences between firms, ensuring that the model captures meaningful changes rather than absolute sales volume.

• **Robustness:** The use of the 75th percentile as a cutoff ensures that the target class is neither too rare nor too common, giving a balance that benefits model training.

• **Alternative Definitions Considered:** Other options included defining growth over a single year or using absolute increases in sales. However, a two-year measure was preferred because it smooths out short-term fluctuations and reflects a more sustained performance trend, which is in line with corporate finance's emphasis on long-term value creation.

II. Data Preparation and Feature Engineering

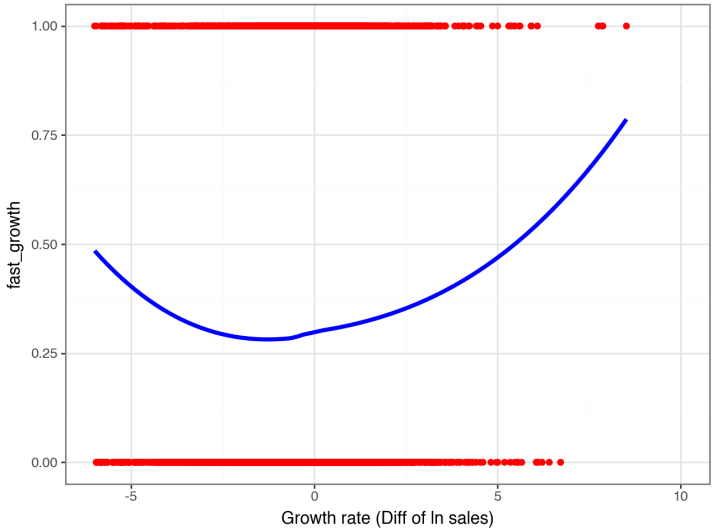
Data Cleaning and Sample Design

Starting with the panel data, we filtered the dataset to include only the years 2011 through 2015 and removed unnecessary variables to streamline the analysis. The sample was further restricted to firms with positive sales in 2012—the baseline year. The data was then reshaped into a panel format with one observation per firm per year, after we compute a two-year sales growth metric and derive a binary "fast growth" variable based on this measure. The sample was further restricted to firms with positive sales in 2012—the baseline year.

Creation of the Target and Predictor Variables

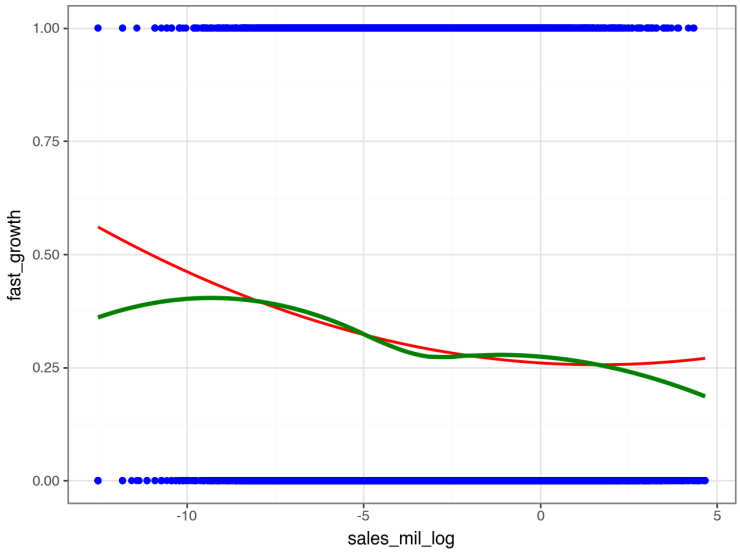
After computing the two-year growth rate (2012 to 2014), each firm was labeled as "fast growing" (1) if its growth exceeded the 75th percentile (about 50% growth), and "not fast growing" (0) otherwise. We transformed sales figures using natural logarithms (e.g., ln_sales, sales_mil_log) to reduce skewness, and then took the difference in log sales (d1_sales_mil_log) as a measure of growth. To manage extreme values, this difference was winsorized at specified cutoffs. We also

created financial ratios by scaling balance sheet items (e.g., current assets, liabilities) by total assets and introduced flag variables to identify potential data issues such as negative or implausibly large asset values. Lastly, categorical features like industry classification and region were converted into dummy variables, ensuring they could be properly used by our modeling algorithms.



Graph 1: Relationship Between Growth Rate and Fast Growth Probability

In Graph 1 the red points represent individual firms, while the blue curve is a smoothed line LOESS illustrating how the probability of fast growth changes across different log-sales-growth values. The curve shows a lower likelihood of fast growth when sales drop compared to the last year, aligning with the intuitive expectation that stronger sales expansions compared to the last year translate into a higher probability of being classified as a fast-growing firm. Interestingly, the curve suggests that at very low (negative) growth rates, there is an unexpected rise in the probability of being classified as fast growing. This phenomenon may be influenced by outliers. To address these extremes, we created a modified version of the growth variable (d1_sales_mil_log_mod), which "winsorizes" or caps any differences in log sales below -1.5. This helps ensure that extreme negative values do not unduly skew the model's perception of growth or inflate the predicted fast-growth probabilities in those rare cases.



Graph 2: Relationship Between Log of Sales and Fast-Growth Classification

In graph 2 the blue points correspond to individual firm observations. Two smooth lines are overlaid: the red curve is a quadratic polynomial fit, and the green curve is a nonparametric LOESS smooth. Both lines generally show that as sales increases (i.e., as firm size grows), the probability of being fast growing tends to decline. This pattern aligns with the intuition that smaller firms (with lower sales) often have an easier time achieving high percentage growth than larger firms. The polynomial (red) and LOESS (green) curves capture slightly different shapes, but both emphasize that the "fast growth" label is somewhat more common among firms with relatively lower sales levels.

III. Model Building and Comparison

We tested a range of models for predicting fast growth, starting with a baseline Ordinary Least Squares (OLS) regression on a simple predictor set (X1). We then built logistic regression models (from X1 to X5) to handle the binary target. To manage higher-dimensional data, we used LASSO for variable selection. We also used a Random Forest to capture non-linearities and interactions, and tested boosting methods Gradient Boosting Machines (GBM), XGBoost, and CatBoost—for their ability to optimize predictive accuracy through iterative learning

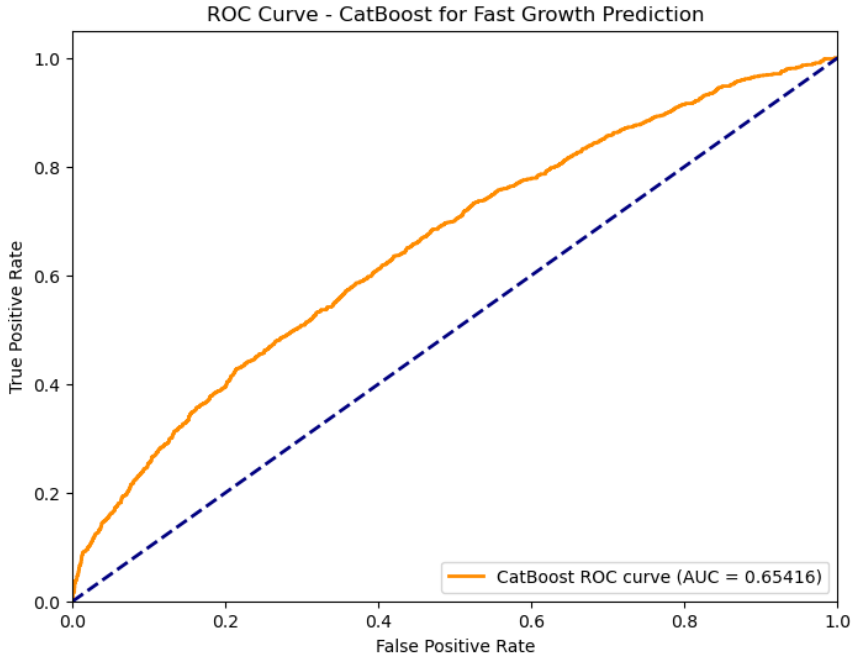
Each model was assessed using five-fold cross-validation. The primary metrics included cross-validated RMSE (to gauge the accuracy of probability estimates) and cross-validated AUC (to measure discriminative power). In addition, we introduced a business-specific loss function—assigning different costs to false positives and false negatives—to derive an optimal classification threshold. This allowed us to compute each model's expected loss and compare them in terms of overall cost-effectiveness. The final performance summary table combines these metrics to highlight the most promising model for predicting fast growth.

IV. Results

Table 1. summary results

Model	Number of Predictors	CV RMSE	CV AUC	CV Threshold	CV Expected Loss
X1	11	0.4524	0.5526	inf	0.2906
X2	18	0.4482	0.5983	0.5130	0.2893
X3	35	0.4471	0.6056	0.5296	0.2887
X4	79	0.4452	0.6171	0.5242	0.2861
X5	163	0.4458	0.6184	0.5871	0.2876
LASSO	107	0.4454	0.6207	0.5681	0.2878
Random Forest	44	0.4443	0.6239	0.6285	0.2850
GBM	44	0.4415	0.6406	0.5831	0.2849
XGBoost	44	0.4415	0.6402	0.7880	0.2873
CatBoost	44	0.4405	0.6467	0.5436	0.2831

Among the models (Table 1), the boosting methods, particularly CatBoost, gives the best performance with the lowest cross validation RMSE and expected loss, and the highest AUC. The selection of CatBoost was guided not only by its overall prediction accuracy but also by its stable performance across cross-validation folds and its ability to handle categorical variables efficiently. For those who require a model with high interpretability, LASSO logit is a strong choice because it naturally selects a few key predictors, resulting in a sparse model that clearly outlines the impact of each variable.



Graph 3. ROC Catboot

The ROC curve (Graph 3) shows CatBoost's classification performance, with an AUC of around 0.65, well above the diagonal baseline, highlighting its strong ability to distinguish fast-growing firms. As we shift the classification threshold toward the bottom-right region of the ROC curve, the model becomes more lenient and can capture a substantial fraction of fast-growing firms early on. This confirms CatBoost as the most promising model for predicting fast growth.

Classification Strategy and Business Considerations

In our classification strategy, misclassifications are directly tied to business costs. We set the loss function with a cost of 1.3 for false positives (FP) and 1.0 for false negatives (FN). Although a higher FP cost could have been used, that would have resulted in too few fast-growth signals, limiting the model's usefulness. Using five-fold cross-validation, we optimized the classification threshold to minimize expected loss.

Table 2. Lasso holdout result

	Predicted no fast growth	**Predicted fast growth**
Actual no fast growth	2781	14
Actual fast growth	1156	37

Table 2, summarizing 3988 cases from holdout set, reveals that while the model correctly identifies the vast majority of non-fast-growth firms (2781 correct out of 2795, yielding a very low false positive rate of about 0.5%), it misses a substantial number of fast-growth firms, with only 37 correctly predicted out of 1193, resulting in a very low true positive rate (TPR) of approximately 3.1%. This shows that the model, while highly specific, is overly concerned and may require threshold adjustments if capturing more fast-growth firms is needed, depending on the relative costs of false negatives versus false positives.

Additionally, recognizing that industry dynamics vary, we applied the same loss function to both manufacturing and service sectors. This segmentation analysis aids in determining whether distinct thresholds or separate models might enhance predictive accuracy, thereby enabling tailored strategies for each industry.

V. Final Recommendations

Our analysis employed a rigorous and systematic approach, beginning with extensive data cleaning and feature engineering to capture the nuanced performance trends of firms over time. By evaluating a range of predictive models, we identified that CatBoost gives the best balance between predictive accuracy and business impact. CatBoost not only delivered the lowest expected loss (0.2831) and the highest AUC (0.6467) among all tested models but also demonstrated robust performance across different segments and efficient handling of categorical variables. While the loss parameters (FP=1.3 and FN=1.0) can be fine-tuned to target even higher growth thresholds, we believe that a broader investment strategy—allocating resources across a larger portfolio of companies identified by our model—yields better results than concentrating on only a few top picks. In finance, diversification often leads to more robust returns, and our approach aligns with that principle.

For practical deployment, we strongly recommend adopting the CatBoost model for predicting fast-growing firms. Its consistent, high-level performance, coupled with the ability to adjust classification thresholds based on business loss considerations makes it a valuable tool for identifying high-growth opportunities. Moreover, a tailored approach that fine-tunes the model separately for manufacturing and service sectors may yield additional benefits, ensuring that our predictive strategy remains both accurate and aligned with strategic business objectives.