Summary Report: Predicting Fast-Growing Firms

Link to the code: github

1. Target Variable Design

In this project, we aimed to identify firms with fast growth potential, using the Bisnode panel dataset for the years 2010–2015. The target variable was constructed based on year-over-year sales growth, specifically comparing 2013 sales to 2012. We defined a firm as "fast-growing" if its annual sales increased by more than 20% in that year, a threshold aligned with the OECD's definition of high-growth enterprises. This indicator captures short-term, actionable growth signals while maintaining a connection to long-term value creation principles in corporate finance.

Alternative target definitions were considered, including two-year cumulative growth and absolute growth thresholds. However, the one-year relative growth measure was preferred for its simplicity and relevance to forecasting immediate business performance. From a financial perspective, rapid sales growth may indicate competitive advantages, operational scalability, or market expansion, all key drivers of firm valuation and attractiveness to investors.

2. Modeling

To predict fast-growing firms, we developed and evaluated three models: logistic regression (logit), random forest, and XGBoost. The dataset was preprocessed and cleaned to remove missing and extreme values. Feature selection was based on a mix of business intuition and exploratory data analysis, including LOWESS plots and cross-tabulations. Variables included historical performance (sales, profit), firm age, sector classification, and ownership characteristics.

The logistic regression model offered a baseline interpretability. However, its predictive power was limited, particularly for capturing non-linear patterns. In contrast, the random forest model provided improved accuracy and robustness to noise, leveraging its ensemble structure. According to the table 1, the third model, XGBboost, achieved the highest cross-validated AUC score, and lowest expected loss suggesting superior generalization.

	Number of Coefficients	CV RMSE	CV AUC	CV treshold	CV expected Loss
M1	16.0	0.452987	0.618541	0.200771	0.672017
M2	23.0	0.447289	0.651345	0.213057	0.663537
M3	40.0	0.442029	0.672425	0.194024	0.645736
M4	83.0	0.439566	0.676471	0.212004	0.642862
M5	197.0	0.440077	0.676408	0.193520	0.640690
LASSO	101.0	0.439519	0.677835	0.210698	0.640759
RF	n.a.	0.438771	0.676892	0.199925	0.642090
XGB	n.a.	0.437491	0.681730	0.203971	0.639988

Table 1: Summary statistics of models' performance on CV

3. Evaluation

To align the model evaluation with business goals, we introduced a custom loss function that assigns different costs to classification errors. In our context, a false positive (predicting a firm will grow fast when it won't) incurs a unit cost of 1, while a false negative (failing to identify a fast-growing firm) has a cost of 4. This cost ratio (FN/FP = 4) reflects the reality that missing a truly high-growth firm—potentially overlooking investment or partnership opportunities—is significantly more costly than mistakenly targeting a slow-growing one.

Each model was evaluated using 5-fold cross-validation. Probabilities were calibrated and optimal thresholds for classification were determined by minimizing the expected loss. The XGBoost model consistently outperformed the others in terms of minimizing the average expected loss across folds.

The final classification results showed a balanced performance with high precision and recall under the optimal threshold. A confusion matrix for a representative fold highlighted the trade-off between over-prediction and under-prediction of growth potential (table 2).

```
0  (non high-growth):  755
1  (high-growth):  2813
```

	Predicted non-HG	Predicted HG
Actual non-HG	636	1822
Actual HG	119	991

Table 2: Confusion matrix of XGBoost

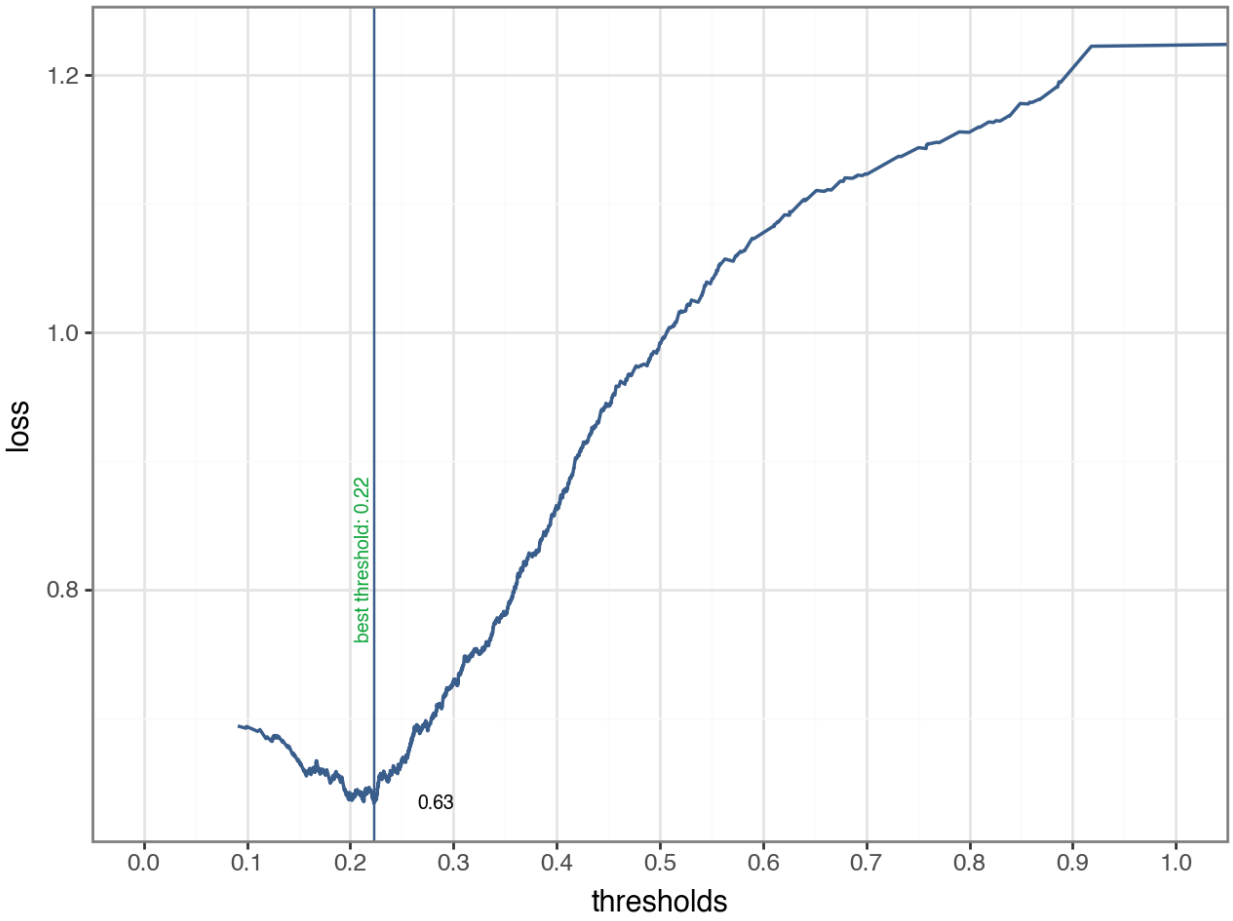

Figure 1: Loss plot of XGBoost with the best threshold

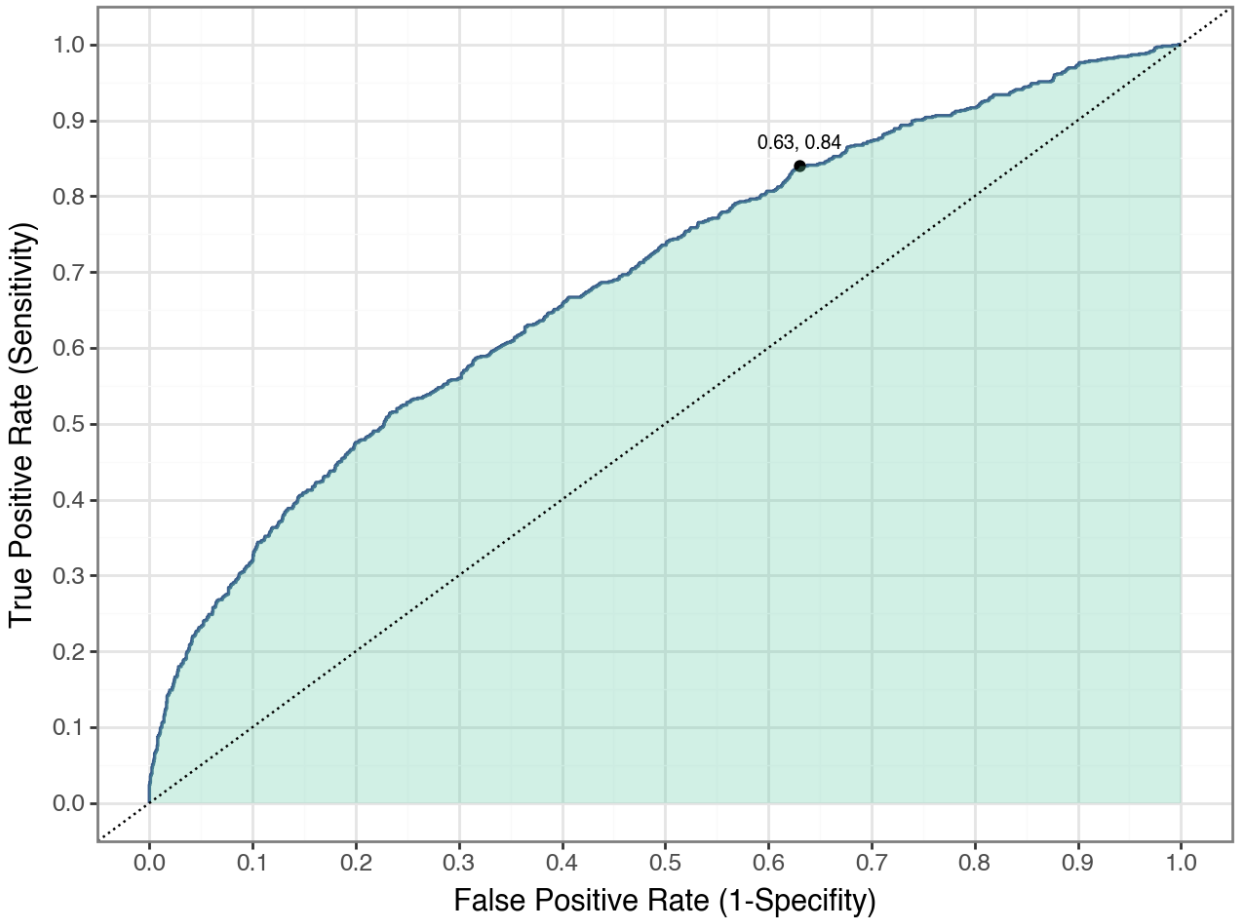

Figure 2: ROC plot of XGBoost with optimal point

4. Industry Subgroup Analysis

To better understand performance differences, we repeated the classification task separately for manufacturing firms and service firms. While using the same loss function, the model performed better for manufacturing firms, likely due to more stable financial indicators and clearer growth patterns. Service firms exhibited higher variability in sales, which may have introduced noise and reduced model precision.

	Metric	2012_hg_workfile_M	2012_hg_workfile_S
0	CV RMSE	0.439	0.432
1	CV AUC	0.677	0.701
2	Avg Threshold	0.194	0.218
3	Avg Expected Loss	0.640	0.615

Table 3: Comparison of expected loss by industry segment

5. Conclusion

This analysis demonstrates that it is feasible to predict fast-growing firms using a combination of historical financial data and machine learning. XGBoost emerged as the most effective model for probability prediction and classification, with a clear advantage in minimizing business-relevant losses. The analysis also revealed sector-specific differences, suggesting that tailored models or thresholds may be appropriate in certain cases.

From a managerial standpoint, this model can support strategic decision-making in areas such as firm acquisition, credit scoring, and growth-focused investment. Future extensions could include incorporating macroeconomic indicators, sentiment analysis from news or reviews, and network-based firm metrics to further enrich the prediction.