

# DA3 Assignment 2: Finding fast growing firms

## Prediction and Classification Using Bisnode Firm-Level Data

### Summary Report

Authors: Elsa Andrea Rodriguez, Petra Ilic | February 2026

#### 1. Executive Summary

This report presents a classification framework for identifying fast-growing firms using the Bisnode firm-level database. Firms in the bottom 20th percentile of log sales are labeled as low-growth (0); the rest as growth firms (1) with the variable `growth_binary`. We evaluate five logistic regression specifications of increasing complexity (M1–M5) plus a LASSO-penalized logit. M4 is selected as the best model based on cross-validated AUC and expected loss. In a second stage, M4 is applied separately to manufacturing and services sub-samples to assess whether predictive performance varies by industry.

#### 2. Data and Target Variable

We defined growth as a binary outcome using the 20th percentile of log sales as the cutoff. This threshold was chosen to isolate the bottom quintile of firms, those most clearly underperforming, rather than using an arbitrary value. The 80/20 class split creates an imbalanced dataset, which we account for in threshold optimization later. The train/holdout split (80/20) was applied before any model fitting to ensure unbiased evaluation.

The dataset covers 19,036 firms: manufacturing (`ind2_cat` 26–30;  $N = 3,790$ ) and services, repair, accommodation, food (`ind2_cat` 33, 55, 56;  $N = 15,246$ ).

	Whole dataset (task 1)	Manufacturing (task 2)	Services (task 2)
N (total)	19,036	3,790	15,246
Growth (1)	80.0%	86.6%	78.3%
No growth (0)	20.0%	13.4%	21.7%
Work / Holdout Ratio	15,229 / 3,807	3,032 / 758	12,197 / 3,050

Table 1. Sample overview

#### 3. Feature Engineering and Model Specifications

Five logit models of increasing complexity were built, plus a LASSO variant. Features cover: (i) 15 balance-sheet/P&L ratios + 4 quadratic terms, (ii) zero/error/outlier flags, (iii) CEO demographics and labor metrics, (iv) firm age, sub-industry, region, urban/rural, and (v) industry interaction terms.

We adopted an incremental modeling strategy, starting from a minimal specification (M1: just profit/loss and industry) and progressively adding feature groups. This allows us to measure the marginal contribution of each feature category. The jump from M2 to M3 (adding region, urban, and all financial ratios) raised AUC from 0.681 to 0.816. The jump from M3 to M4 (adding quadratic terms, flags, and HR variables) raised it further to 0.917. Adding interaction terms (M5) or applying LASSO regularization provided no further improvement, confirming that M4 captures the relevant signal without overfitting.

Model	General Description
M1	Profit/loss + industry code
M2	Core financials + age + foreign management
M3	Firm characteristics + region + urban + all engineered financial vars
M4	M3 + quadratic terms + quality/error flags + HR variables
M5	M4 + industry $\times$ demographic/firm interaction terms
LASSO	M5 variable set with L1 regularization (*72 non-zero coefficients selected)

**Table 2.** Model specification

## 4. Task 1: Model Comparison (Full Sample)

### 4.1 Cross-Validation Performance

All models fitted with 5-fold CV using LogisticRegressionCV. LASSO tuned over 10 lambda values. The table below shows averaged CV metrics.

Model	Coeffs	CV RMSE	CV AUC	Avg Threshold	Avg Exp. Loss
<b>M1</b>	3	0.377	0.633	$\infty$	0.803
<b>M2</b>	10	0.372	0.681	$\infty$	0.803
<b>M3</b>	24	0.343	0.816	0.875	0.608
<b>M4</b>	68	0.295	0.917	0.883	0.35

<b>M5</b>	76	0.301	0.911	0.878	0.368
<b>LASSO</b>	72	0.295	0.918	0.895	0.35

**Table 3.** CV performance summary.

M1 and M2 are too simple to produce a finite optimal threshold under the asymmetric loss function, their discriminatory power is insufficient for the 10:1 cost structure. M3 marks the first viable classification model. M4 and LASSO achieve nearly identical performance across all metrics (RMSE = 0.295, AUC  $\approx$  0.917–0.918, loss  $\approx$  0.350), but M4 requires no regularization tuning and is fully interpretable. M5 actually worsens performance despite adding 8 more variables, a clear sign of overfitting to noise in the interaction terms.

M4 has been selected as the best model. It ties LASSO on CV RMSE (0.295) and expected loss (0.350), and nearly ties on AUC (0.917 vs 0.918). M4 is preferred because it is simpler (no regularization) and more interpretable. M5 adds 8 interaction terms but worsens both AUC and loss.

#### 4.2 Holdout Evaluation (M4)

M4 holdout RMSE = 0.298 (vs CV 0.295), confirming no overfitting. Mean predicted probability = 0.794.

	<b>Prediction No Growth</b>	<b>Prediction Growth</b>	<b>Prediction No Growth</b>	<b>Prediction Growth</b>
	thr = 0.5		thr = 0.794	
<b>Actual: No Growth</b>	457	355	697	115
<b>Actual: Growth</b>	121	2,875	551	2,445

**Table 4.** Confusion matrices on holdout (N = 3,808)

The two confusion matrices illustrate the threshold trade-off: at 0.5, the model flags 3,230 firms as growth but makes 355 false positive errors. At the mean predicted probability (0.794), false positives drop to 115 but at the cost of 551 false negatives. Neither threshold is optimal, the loss function in Section 5 determines the right balance.

### 5. Classification with Asymmetric Loss

We set FP cost = 10 and FN cost = 1 to reflect a business scenario where the cost of investing resources in a firm falsely identified as growing is ten times higher than the cost of overlooking a genuinely growing firm. The resulting optimal threshold of 0.883 is far above the conventional 0.5, the model must assign an 88% probability of growth before flagging a firm. This conservative approach dramatically reduces false positives at the expense of some missed growth firms, which is the correct trade-off given the cost structure.

Loss function: FP cost = 10, FN cost = 1. Rationale: falsely investing in a non-growth firm is far costlier than missing a growth opportunity.  $E[\text{Loss}] = (\text{FP} \times 10 + \text{FN} \times 1) / N$ . The optimal threshold is found by searching [0.05, 0.95] in each CV fold.

<b>Metric</b>	<b>Value</b>
Avg optimal threshold (CV)	0.883
Fold 5 threshold	0.887
Avg expected loss (CV)	0.350
Fold 5 expected loss	0.340

**Table 5.** M4 threshold and loss

The high threshold (~0.88) is driven by the 10:1 cost ratio. The model must be very confident before classifying a firm as growing, minimizing expensive false positives. This is far above the default 0.5.

## 6. Industry Comparison

M4 was applied separately to manufacturing (ind2\_cat 26–30) and services (ind2\_cat 33, 55, 56) using the same loss function and CV procedure. The 20th percentile was computed on the full dataset before splitting.

<b>Metric</b>	<b>Manufacturing</b>	<b>Services</b>
<b>N observations</b>	3,790	15,246
<b>Growth = 1 prevalence</b>	86.6%	78.3%
<b>CV RMSE</b>	0.267	0.307
<b>CV AUC</b>	0.926	0.911
<b>Holdout RMSE</b>	0.248	0.298
<b>Optimal threshold</b>	0.897	0.883
<b>Expected loss (CV)</b>	0.261	0.370

**Table 6.** M4 performance: Manufacturing vs. Services

	<b>Manufacturing</b>		<b>Services</b>	
	<b>Prediction No Growth</b>	<b>Prediction Growth</b>	<b>Prediction No Growth</b>	<b>Prediction Growth</b>
<b>Actual: No Growth</b>	54	38	416	267
<b>Actual: Growth</b>	26	640	122	2,245

**Table 7.** Confusion matrices at threshold = 0.5 (Task 2 holdout)

Manufacturing outperforms services across all metrics. AUC is 1.5pp higher (0.926 vs 0.911) and expected loss is 42% lower (0.261 vs 0.370). This reflects greater homogeneity in manufacturing financial reporting versus the heterogeneous services sector (repair + accommodation + food).

## 7. Key Decisions, Conclusions, and Recommendations

The analytical framework defines the performance target at the 20th percentile of log sales, isolating a minority class of approximately 20% representing truly low-performing firms. Following an extensive evaluation of candidate architectures, the M4 logit model (utilizing 68 coefficients) emerged as the optimal selection. With an AUC of 0.917 and an expected loss of 0.350 on the full sample, the M4 model achieved performance parity with LASSO in terms of RMSE and loss. Given that regularization offered no marginal improvement, the unregularized M4 was selected for its superior simplicity and interpretability over more complex iterations, such as the M5 model, which failed to yield gains despite the addition of interaction terms.

The loss function was calibrated to reflect the significant financial risk of misidentifying growth potential, setting a 10:1 cost ratio for False Positives (FP) versus False Negatives (FN). This high cost of investing in falsely identified growth firms necessitated a conservative classification threshold, which was cross-validated and optimized at 0.88–0.90 rather than the default 0.5. This ensures that only high-confidence predictions are flagged for action.

The model architecture was applied through independent pipelines for the manufacturing and services sectors to account for distinct industrial dynamics. The results indicate that manufacturing is significantly more predictable, yielding a 42% lower expected loss (0.261) compared to services (0.370). This disparity suggests that standard financial features possess stronger predictive power within the manufacturing sector.

Conversely, the services sector exhibits higher loss levels, likely due to the inherent heterogeneity across sub-industries such as repair, accommodation, and food services. While the manufacturing model can be deployed with high confidence, the services model requires further refinement, potentially through sector-specific features or sub-industry modeling.

### Technical Notes

*Models: scikit-learn LogisticRegressionCV, 5-fold CV (random\_state=42). Standard logit: C=10<sup>20</sup> (unregularized), newton-cg solver. LASSO: L1 penalty, liblinear solver, 10 lambda values (10<sup>-1</sup> to 10<sup>-4</sup>). Design matrices via patsy. Threshold search: 0.05–0.95 per fold, cost-weighted. Data: Bisnode firm-level database. Code:*

[https://github.com/petrailic/assignment\\_2](https://github.com/petrailic/assignment_2)