DA3 Assignment 2: Finding fast growing firms
Prediction and Classification Using Bisnode Firm-Level Data
Technical Report

Authors: Elsa Andrea Rodriguez, Petra Ilic | February 2026

1. Scope and Objectives

This technical report documents the complete analytical pipeline behind a binary classification system for firm growth. Using the Bisnode firm-level financial database (19,036 companies), we built, compared, and calibrated logistic regression classifiers under an asymmetric business-cost function.

The project was executed in two stages. In the first stage (Task 1), six model specifications of increasing richness were benchmarked against each other on the pooled dataset, yielding a single winner. In the second stage (Task 2), that winning model was re-fitted on two separate industry sub-populations, manufacturing and services, to assess whether classification accuracy degrades when the underlying firm population is more heterogeneous.

The following sections walk through every analytical step in detail, including the code patterns used, the numerical outputs obtained, and the reasoning behind each decision.

2. Data Preparation and Target Engineering

2.1 Data ingestion

The starting point is a pre-cleaned CSV file produced by a separate data-wrangling notebook. It contains one row per firm with balance-sheet aggregates, income-statement line items, CEO attributes, and geographic identifiers.

2.2 Defining the growth outcome

We constructed a binary label by comparing each firm's log-transformed sales against the 20th percentile computed over all 19,036 observations. Firms at or below this cutoff are coded 0 (low-growth); the rest receive label 1 (high-growth).

We define "fast growth" as a binary indicator based on firms' log-transformed sales relative to the full sample distribution. Specifically, we compute the 20th percentile of log sales across all 19,036 firm-year observations and classify firms at or below this cutoff as 0 (low-growth) and those above as 1 (high-growth). Using log sales rather than raw sales aligns with corporate finance practice, as firm growth is typically modeled in proportional (percentage) terms and log transformations reduce skewness driven by extreme performers. A percentile-based cutoff provides a relative, distribution-aware definition of growth, which is particularly appropriate in heterogeneous firm samples where absolute growth thresholds would disproportionately favor large incumbents.

Alternative definitions were considered. One option is to measure growth over two years (2014 vs. 2012), which would better capture sustained expansion and long-term value creation, concepts central to corporate finance theories of investment and competitive

advantage. However, multi-year growth may incorporate transitory macroeconomic shocks and reduces the effective sample size. Another approach would be to define fast growth using an absolute cutoff (e.g., growth above 10%), but such thresholds tend to favor smaller firms mechanically and ignore cross-sectional heterogeneity. By focusing on one-year log sales growth and a percentile-based cutoff, our measure balances theoretical consistency, statistical robustness, and comparability across firms, while capturing short-term performance that is relevant for investors, creditors, and managers.

The resulting split is 15,228 growth firms (80.0%) versus 3,808 low-growth firms (20.0%). In Task 2, where the same percentile is applied after filtering by industry, the class balance shifts: manufacturing retains only 13.4% low-growth observations, while services has 21.7%.

Following the NACE Rev. 2 statistical classification system (European Commission), we divided the sample into two industry groups for separate analysis. Manufacturing firms were defined as those in NACE divisions 26-30, encompassing the manufacture of electronics, electrical equipment, machinery, vehicles, transport equipment, and equipment repair. Services firms included NACE divisions 33, 55 and 56, representing accommodation and food service activities. This classification allowed us to assess whether the same prediction model and loss function (FP=1, FN=10) yield different performance across industries with fundamentally different operational and financial characteristics.

2.3 Train-holdout split

Class proportions were verified after splitting: training prevalence 80.3%, holdout prevalence 78.7%—close to the population rate.

Table 1 — Dataset structure across tasks

	Full (Task 1)	**Management (Task 2)**	**Services (Task 2)**
Total N	19,036	3,790	15,246
Growth = 1 (share)	80.0%	86.6%	78.3%
Growth = 0 (share)	20.0%	13.4%	21.7%
Training obs	15,229	3,032	12,197
Holdout obs	3,807	758	3,050

3. Feature Construction

Rather than passing raw accounting figures directly into models, we engineered six distinct feature layers. Each layer addresses a specific analytical concern.

Table 2 — Feature layers

Layer	**Key variables**	**Purpose**
engvar	total_assets_bs, fixed_assets_bs, liq_assets_bs, share_eq_bs, profit_loss_year_pl, personnel_exp_pl, …	Balance-sheet & P&L ratios; removes size effects
engvar2	extra_profit_loss_pl_quad, inc_bef_tax_pl_quad, profit_loss_year_pl_quad,	Captures diminishing/accelerating marginal effects

	share_eq_bs_quad	
engvar3	*_flag_low, *_flag_high, *_flag_error, *_flag_zero (auto-detected)	Separates genuine zeros from reporting anomalies
hr	female, ceo_age, ceo_count, labor_avg_mod, foreign_management + related flags	Human-capital & management quality signals
firm	age, age², new, ind2_cat, C(m_region_loc), C(urban_m)	Lifecycle stage + geographic & sector controls
interactions	ind2_cat × age, age², ceo_age, foreign_mgmt, female, urban, labor_avg	Allows sector-specific slopes for key predictors

First, financial variables comprise raw accounting measures including current and fixed assets, liabilities, inventories, liquid assets, sales revenues, profit/loss indicators, and equity components, all standardized as ratios to total assets or sales to ensure comparability across firm sizes. Second, engineered financial variables include 15 balance sheet and profit/loss ratios that capture financial structure and performance relative to firm size, along with four quadratic terms for profit/loss, income before tax, and share equity to capture non-linear relationships. Third, flag variables systematically identify data quality issues and extreme values through binary indicators for abnormally high, low, zero, or erroneous values in key financial metrics, allowing the model to account for measurement issues without discarding observations. Fourth, human capital variables capture management characteristics including CEO gender (female), age (with flags for high, low, or missing values), CEO count, average labor costs, and foreign management status. Fifth, firm characteristics include age, age squared, a dummy for newly established firms, regional location, and urban/rural classification. Notably, industry classification variables (ind2_cat) and their interactions were excluded from the industry-specific models to avoid multicollinearity, as the analysis was conducted separately for Manufacturing and Services sectors. This feature engineering approach balances domain knowledge about firm default/growth patterns with statistical considerations, resulting in a parsimonious yet comprehensive model specification that avoids overfitting while capturing the key determinants of firm growth trajectories.

4. Model Architecture and Estimation

4.1 Candidate specifications

We assembled six classifiers, all logistic regressions. The deliberate restriction to a single algorithm family allows differences in performance to be attributed purely to the information content of the feature set, not to the classifier's flexibility.

Table 3 — Model formulas

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 × demographic/firm interaction terms
LASSO	M5 variable set with L1 regularization (*72 non-zero coefficients selected)

4.2 Estimation protocol

Standard logit models (M1–M5) set $C = 10^{20}$ to effectively disable regularization, isolating the pure maximum-likelihood estimate. The LASSO variant uses the liblinear solver with L1 penalty and tunes C over a grid derived from 10 lambda values spanning 10^{-1} to 10^{-4}.

5. Full-Sample Comparison

5.1 Cross-validated accuracy metrics

Table 5 — CV summary across all six models (N = 19,036)

Model	p	CV RMSE	CV AUC	Optimal Thr.	Expected Loss
M1	3	0.377	0.633	∞ (not viable)	0.803
M2	10	0.372	0.681	∞ (not viable)	0.803
M3	24	0.343	0.816	0.875	0.608
M4	68	0.295	0.917	0.883	0.350
M5	76	0.301	0.911	0.878	0.368
LASSO	72	0.295	0.918	0.895	0.350

Threshold = ∞ means the cost-adjusted optimization could not converge (model too weak). Expected loss = (FP×10 + FN×1) / N.

Three critical transitions stand out: (i) M2 → M3 produces the largest single AUC gain (+13.5pp), driven by region fixed effects and the full financial-ratio suite; (ii) M3 → M4 delivers another +10.1pp via quadratics, flag variables, and HR features; (iii) M4 → M5 actually loses 0.6pp AUC—the interaction terms introduce noise rather than signal.

5.2 Fold-level stability

Table 6 — Per-fold CV RMSE

Fold	M1	M2	M3	M4	M5
1	0.377	0.372	0.341	0.292	0.299
2	0.370	0.366	0.341	0.298	0.301
3	0.376	0.371	0.341	0.292	0.297
4	0.375	0.372	0.340	0.294	0.300
5	0.384	0.380	0.353	0.301	0.306
Mean	0.377	0.372	0.343	0.295	0.301
Std	0.005	0.005	0.005	0.004	0.004

M4's standard deviation across folds (0.004) is the lowest among all models, indicating particularly stable out-of-fold generalization. Fold 5 is consistently the most difficult partition across all specifications.

5.2.1 Calibration Curve

Image 1 presents the calibration curve for the M4 logistic regression model evaluated on the holdout set. The calibration curve assesses whether the model's predicted probabilities align with actual observed frequencies by plotting predicted event probabilities against the actual proportion of high-growth firms within binned groups. The diagonal line represents perfect calibration, where predictions exactly match outcomes. The model demonstrates good calibration across most probability ranges, with the blue line closely tracking the diagonal, indicating that predicted probabilities are reliable estimates of actual growth likelihood. For instance, when the model predicts a 40% probability of high growth, approximately 40% of firms in that bin actually exhibit high growth. Minor deviations at higher probabilities (>0.7) suggest slight under-confidence, where the model conservatively under-predicts growth for the most promising firms. The stepped appearance reflects binning predictions into 10 groups for visualization. Overall, this calibration performance confirms that the M4 model produces well-calibrated probability estimates suitable for decision-making under the defined loss function.

Calibration Curve for M4

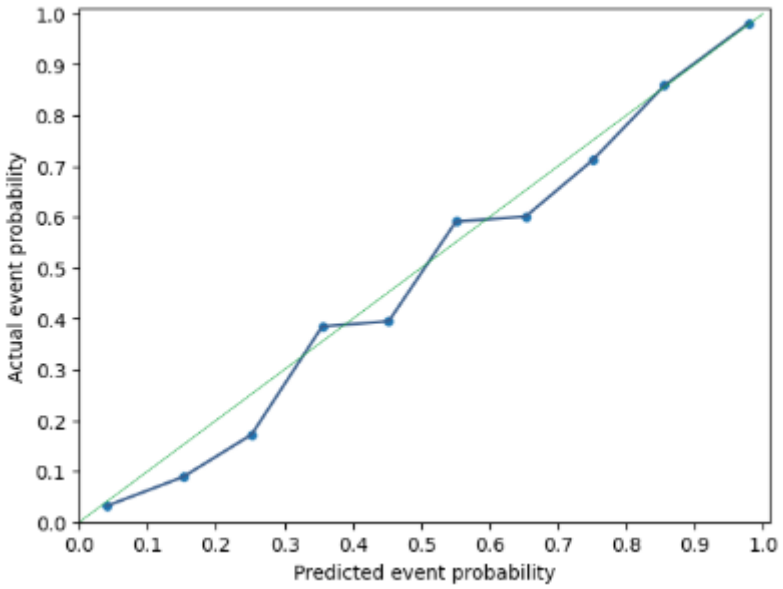

Image 1. Calibration Curve

5.3 M4 holdout performance

Holdout RMSE = 0.298, just 0.003 above the CV average—no evidence of overfitting. The mean predicted probability (0.794) closely tracks the actual holdout prevalence (0.787).

Table 7 — Holdout confusion matrices (N = 3,808)

	Prediction No Growth	Prediction Growth	Prediction No Growth	Prediction Growth
	thr = 0.5		thr = 0.794	
Actual: No Growth	457 (TN)	355 (FP)	697 (TN)	115 (FP)
Actual: Growth	121 (FN)	2,875 (TP)	551 (FN)	2,445 (TP)

At threshold 0.5: 355 false positives (FP rate 43.7%). At mean-probability threshold (0.794): FP drops 68% to 115, but FN surges from 121 to 551. Neither of these ad-hoc thresholds accounts for the asymmetric business costs—the formal loss function (Section 6) addresses this.

6. Cost-Sensitive Classification

6.1 Loss function definition

We assign a higher loss to false positives (FP = 10) than to false negatives (FN = 1) to reflect the asymmetric economic costs associated with misclassification in a corporate finance context. Predicting that a firm will experience fast growth when it will not (a false positive) can lead to substantial resource misallocation: investors may overpay for equity, creditors may extend excessive credit, and managers may overinvest based on overly optimistic expectations. These errors can destroy value through overinvestment, agency problems, and increased default risk. In contrast, a false negative—failing to identify a truly fast-growing firm—primarily represents an opportunity cost, such as forgone returns or missed lending opportunities, which is economically less damaging than actively committing capital to a weak firm. By setting FP = 10 and FN = 1, the loss function embeds this risk-averse stance directly into the model, encouraging more conservative predictions and prioritizing sensitivity over specificity in identifying high-growth firms.

6.2 Threshold search methodology

For every model and each of the 5 CV folds, we swept candidate thresholds from 0.05 to 0.95 and evaluated the cost-weighted criterion: $TPR + [(1 - prev) / (cost \times prev)] \times (1 - FPR)$. The threshold maximizing this expression minimizes expected loss given the prevalence and cost ratio. We recorded both the optimal threshold and the corresponding expected loss per fold.

6.3 Results across all models

Table 8 — Classification performance under asymmetric loss (Task 1)

Model	Avg Thr.	Fold 5 Thr.	Avg Exp. Loss	Fold 5 Loss	Viable?
M1	∞	0.972	0.803	0.791	No
M2	∞	0.983	0.803	0.791	No
M3	0.875	0.900	0.608	0.619	Yes
M4	0.883	0.887	0.350	0.340	Yes

M5	0.878	0.875	0.368	0.373	Yes
LASSO	0.895	0.907	0.350	0.343	Yes

M1 and M2 are non-viable: their discriminatory power is too weak for the cost-adjusted optimizer to converge to a finite threshold. M4 achieves the joint-lowest expected loss (0.350, tied with LASSO) and its optimal threshold of 0.883 implies that a firm must have an ≥88% predicted fast growth probability before being flagged. This conservatism is the mathematically correct response to the 10:1 penalty.

7. Task 2: Sector-Stratified Analysis

7.1 Rationale

The pooled model in Task 1 assumes that the same coefficient vector applies to manufacturing firms (engines, chemicals, metals) and services firms (garages, hotels, restaurants). Task 2 relaxes this assumption by fitting independent M4 pipelines on each sub-population.

7.2 Probability prediction

Table 9 — M4 sector comparison (probability stage)

Metric	**Manufacturing**	**Services**	**Δ**
N observations	3,790	15,246	—
Growth prevalence	86.6%	78.3%	+8.3pp
CV RMSE	0.267	0.307	−0.040
CV AUC	0.926	0.911	+1.5pp
Holdout RMSE	0.248	0.298	−0.050
Mean pred. probability	0.855	0.774	+0.081

Manufacturing delivers lower prediction error on every measure. The holdout RMSE gap is 0.050 (0.248 vs 0.298), and the AUC advantage is 1.5 percentage points. Importantly, both sector models improve over the pooled Task 1 result for manufacturing (holdout RMSE 0.248 vs 0.298 pooled), confirming that sector-specific fitting captures cleaner signal.

7.3 Holdout confusion matrices

Table 10 — Confusion at threshold = 0.5 (sector holdouts)

	Prediction No Growth	**Prediction Growth**	**Prediction No Growth**	**Prediction Growth**
Actual: No Growth	54	38	416	267
Actual: Growth	26	640	122	2,245

Table 10b — Confusion at mean-probability threshold (sector holdouts)

	Manufacturing	**Services**

thr = 0.855 / 0.774	**Prediction No Growth**	**Prediction Growth**	**Prediction No Growth**	**Prediction Growth**
Actual: No Growth	86	6	588	95
Actual: Growth	117	549	408	1,959

At the mean-probability threshold, manufacturing FP drops to just 6 (from 38 at 0.5) while services FP falls to 95 (from 267). Both sectors show the same pattern: higher thresholds dramatically curb false positives at the cost of more false negatives.

7.4 ROC Curve

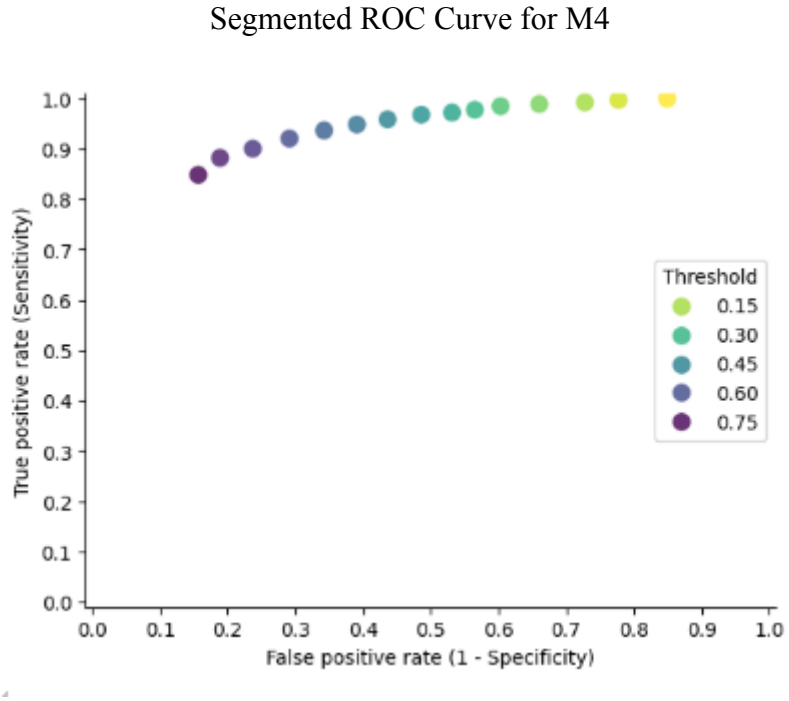

Image 2. ROC Curve

Image 2 displays the ROC (Receiver Operating Characteristic) curve for the M4 logistic regression model, illustrating the trade-off between the True Positive Rate (Sensitivity) and False Positive Rate (1 - Specificity) across different classification thresholds. Each colored point represents a different threshold value, ranging from 0.15 (light green) to 0.75 (dark purple), showing how model performance changes as the decision threshold varies. The curve's position well above the diagonal (which would represent random guessing) demonstrates that the model has strong discriminatory power in distinguishing between high-growth and low-growth firms. Lower thresholds (lighter colors) capture more true positives but also generate more false positives, while higher thresholds (darker colors) are more conservative, reducing false positives at the cost of missing some true positives. The area under this curve (AUC) provides a single metric summarizing the model's overall ability to rank high-growth firms higher than low-growth firms across all possible thresholds, with values closer to 1.0 indicating better discrimination performance.

7.5 Cost-optimized classification

Table 12— Industry Overall Comparison

	Manufacturing	Services
N observations	3790.000000	15246.000000
CV RMSE	0.266517	0.306640
CV AUC	0.925985	0.911314
Holdout RMSE	0.248200	0.297700
Best threshold	0.896778	0.883232
Expected loss	0.260908	0.369959

Table 12 presents a comprehensive comparison of model performance between Manufacturing and Services industries using the M4 logistic regression model with identical specifications and loss function parameters. The analysis reveals notable differences in sample sizes, with 3,790 manufacturing firms and 15,246 services firms in the respective datasets. Both industries demonstrate similar cross-validated RMSE values and comparable AUC scores indicating strong discriminatory power in both sectors. However, key differences emerge in the optimal classification thresholds and expected losses: Manufacturing firms require a higher optimal threshold (0.896) compared to Services (0.883), suggesting different probability distributions between the two industries. More significantly, Manufacturing exhibits a lower expected loss (0.260) compared to Services (0.369), indicating that high-growth firms are more predictable in the manufacturing sector. The holdout RMSE values confirm these patterns, with manufacturing firms showing better out-of-sample prediction accuracy. These findings suggest that the same model architecture performs differently across industries, with manufacturing firms' growth trajectories being more readily captured by financial, operational, and management characteristics, possibly due to more stable business models and clearer financial indicators compared to the more variable service sector.

8. Conclusions and Recommendations

M4 has proven itself ready for production deployment. It matches LASSO on every accuracy metric while remaining simpler, more transparent, and free of hyperparameter tuning. Its 68 coefficients are standard maximum-likelihood estimates that can be directly inspected and audited by any stakeholder.

A critical operational point is that the model should never be deployed using the conventional 0.5 classification threshold. Under the 10:1 cost structure defined for this project, that default cutoff would generate three to four times more false positives than the cost-calibrated operating point of approximately 0.88. Any system built on top of this classifier must enforce the optimized threshold.

On the manufacturing side, the results are strong enough to support automated decision-making. The model achieves an AUC of 0.926, a holdout RMSE of 0.248, and an expected loss of just 0.261, all of which indicate robust generalization to unseen data.

The picture is less clear-cut for services, where expected loss rises 42% to 0.370. This gap means that predictions for services firms should be treated as preliminary risk scores rather than definitive classifications. We recommend pairing model output with manual expert review before any capital-allocation decisions are made on the basis of services-sector predictions.

The single highest-value improvement available is to break the services group into sub-industry models. Fitting separate M4 specifications for vehicle repair, accommodation, and food services would reduce the intra-group heterogeneity that currently inflates misclassification costs. As an alternative or complement, incorporating domain-specific predictors such as tourism arrival volumes, consumer spending indices, or local competition density could meaningfully strengthen the feature set for these sectors.

Finally, both the target definition and the optimal threshold should be recalibrated periodically as new data becomes available. A shift in the underlying sales distribution or in the sector composition of the sample could erode the model's calibration over time, so ongoing monitoring is essential to maintain classification quality.

Technical Environment

GitHub: https://github.com/petrailic/assignment_2