**Summary Report: Predicting Fast-Growing Firms 2025**

**1. Introduction**

This assignment seeks to **identify fast-growing firms** based on the **bisnode-firms** dataset (2010–2015). The **business goal** is to prioritize investment prospects by predicting which firms experience **≥20% sales growth** between 2012 and 2013. This threshold captures robust expansion (higher than typical 5–10% growth benchmarks) without being overly restrictive.

We considered other definitions—like comparing 2014 vs. 2012 or using a 50% threshold—but:

1. **2014 data limitation**: Fewer firm observations.
2. **50% growth**: Would eliminate many valid “high potential” firms.

**Key Tasks**

1. **Data Preparation & Feature Engineering**: Filter to 2012–2013, remove excessive missingness and outliers, engineer ratios for financial statements, handle potential leakage (e.g., dropping sales12).
2. **Task 1**: Build multiple models (Logit with splines, simpler Logit, LASSO Logit, Random Forest), compare via cross-validation on:
   * AUC (discrimination),
   * RMSE (probability quality),
   * Expected Loss under FP=1, FN=10.
3. **Task 2**: Evaluate the best model (Random Forest) separately for **Manufacturing** (NACE 1000–3400) and **Services**(NACE ≥4500), using the same cost ratio and comparing performance and expected loss.

We reference code snippets from our Jupyter notebook for details on cleaning, variable creation, model training, and classification threshold selection.

**2. Methodology**

**2.1 Data Preparation**

1. **Source & Initial Filtering**
   * Started with cs\_bisnode\_panel.csv (~287,829 observations for 2010–2015).
   * Reduced to 2012–2013 data (56,943 rows), focusing on 2013 for the outcome prediction (~14,689 rows after cleaning).
2. **Cleaning & Imputation**
   * Dropped columns with high NA (e.g., COGS, net\_dom\_sales), set negative asset entries to zero, and capped certain variables (e.g., intangible assets, which should not be negative).
   * Imputed missing CEO age with the mean (~50.5).
   * Removed any rows with missing critical variables (e.g., liq\_assets\_bs).
3. **Feature Engineering**
   * **Financial Ratios**: Divided P&L items by sales (e.g., personnel\_exp\_pl) and balance sheet items by total assets (e.g., liq\_assets\_bs).
   * **Quadratic Terms**: For certain items like profit\_loss\_year\_pl to capture non-linearities.
   * **Spline Transformations**: Custom function lspline for piecewise linear approximations (e.g., amort with knot at 125,000).
   * **Flag Variables**: Marked out-of-range items (e.g., extremely high or negative shares).
   * **Target**: f\_growth=1 if sales growth ≥20% from 2012 to 2013, resulting in **26% positives** (3,810 out of 14,689).
4. **Train/Holdout Split**
   * 80% training (~11,751 rows), 20% holdout (~2,938 rows).
   * Verified class proportions remain similar (around 74% vs. 26% in each split).

**2.2 Model Building**

We implemented **four core models** in sklearn and statsmodels:

1. **M1 – Logit (with Splines):**
   * 15 variables (e.g., lspline(amort, [125000]), lspline(material\_exp\_pl, [0.9]), foreign\_management), capturing non-linear patterns in accounting items.
   * Trained via LogisticRegressionCV with cross-validation (5 folds).
2. **M2 – Simpler Logit:**
   * 19 variables in linear form (e.g., amort, profit\_loss\_year\_pl), intentionally less complex.
3. **LASSO Logit:**
   * 23 variables, including interactions (e.g., curr\_liab\_bs \* liq\_assets\_bs), penalized by L1-regularization.
   * Used LogisticRegressionCV with penalty='l1' and found that **20 coefficients** remain non-zero at the optimal lambda.
4. **Random Forest (RF):**
   * 18 variables, focusing on major financial indicators plus CEO attributes (ceo\_age, foreign\_management).
   * Tuned via GridSearchCV over (max\_features ∈ [1..5], min\_samples\_split ∈ [80..100]); used 500 trees (n\_estimators=500).

**Cross-validation** used the same 5-fold splits, comparing:

* **AUC** (0.5 = random, 1.0 = perfect),
* **RMSE** of predicted probabilities vs. actual labels,
* **Expected Loss** with a cost function:Loss=(FP×1+FN×10)/Sample Size.Loss=(FP×1+FN×10)/Sample Size.

**2.3 Classification & Cost-Sensitive Threshold**

After probability prediction, we must set a **classification threshold** to balance false positives (FP=1) vs. false negatives (FN=10). The code calculates an **optimal threshold** by scanning ROC curve points, aiming to minimize expected loss. For example, a threshold around **0.06** yields near-perfect recall (~1.0), ensuring very few missed fast-growers at the expense of more false alarms.

**2.4 Industry-Specific (Task 2)**

* **Manufacturing**: NACE 1000–3400, 4,861 firms, 29% positives.
* **Services**: NACE ≥4500, 9,828 firms, 24% positives.
* Re-ran RF with the same cost ratio, computed optimal thresholds, and recorded AUC, expected loss, and confusion metrics for each sector.

**3. Results**

**3.1 Overall Model Comparison (Task 1)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Features | CV AUC | CV RMSE | Avg. CV Expected Loss |
| M1 (Logit w/ Splines) | 26 | 0.637 | 0.436 | ~0.68 |
| M2 (Simple Logit) | 20 | 0.629 | 0.437 | ~0.69 |
| LASSO | 20 | 0.632 | 0.436 | ~0.68 |
| RF | 18 | 0.660 | 0.431 | ~0.68 |

* **Random Forest** leads in **AUC (0.66)** and **RMSE (0.431)**.
* Logits show decent performance (AUC ~0.63), but less flexibility for non-linear patterns.
* All produce similar expected loss estimates (0.68–0.69), but closer inspection shows RF has more consistent and robust classification across folds.

**Holdout** performance for RF:

* **AUC** ≈ 0.65, **RMSE** = 0.434,
* Confusion matrix at threshold 0.06 yields near-perfect recall, albeit with many false positives (precision ~0.33).

**3.2 Sector-Specific Analysis (Task 2)**

Using RF with cost ratio FN=10, FP=1:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Industry | Expected Loss | AUC | Accuracy | Precision | Recall | F1 | Baseline Loss | Improvement |
| Manufacturing | 0.68 | 0.69 | 0.3252 | 0.301 | 0.998 | 0.46 | 2.90 | 2.22 |
| Services | 0.68 | 0.68 | 0.3225 | 0.265 | 0.998 | 0.41 | 2.44 | 1.76 |

* **Manufacturing** yields slightly higher AUC (0.691 vs. 0.683) and better F1.
* Both achieve recall ~0.998 at threshold 0.06, significantly reducing costly misses.
* Expected loss falls from ~2.90 (Manufacturing) and 2.44 (Services) to ~0.68, a major improvement given FN=10.

**4. Discussion of Findings**

1. **Modeling Choices**
   * Splines capture non-linearities but can overfit if not well-chosen. LASSO helps reduce feature clutter, retaining 20 significant coefficients.
   * However, **Random Forest** consistently delivered superior CV metrics and adaptability across industries.
2. **Cost-Sensitive Threshold**
   * The high FN cost (10× vs. FP) drives the optimal threshold down to ~0.06, ensuring near-perfect recall.
   * Precision (~0.30) is relatively low; we accept more false positives to avoid missing truly fast-growers.
3. **Sector Differences**
   * Manufacturing’s slightly better AUC suggests more stable drivers of growth (e.g., labor or profit margins).
   * Services are more heterogeneous, thus slightly lower precision and AUC.
4. **Limitations & Next Steps**
   * AUC ~0.69 leaves scope for improvement: adding external data (market/industry trends) or refining cost ratios (e.g., FN=5) could reduce false alarms.
   * Over-prediction can strain follow-up resources; firms flagged “fast-growing” need further due diligence.

**5. Conclusion & Recommendations**

* **Random Forest @ Threshold ~0.06** is our recommended approach:
  + Highest overall CV AUC (~0.66) and best holdout RMSE (0.434).
  + Very high recall (~0.998), crucial for investment screening where missed opportunities can be costly.
* **Key Implementation Steps**:
  + **Ingest & Clean Data**: Continue dropping or imputing missing assets, ensure correct ratio transformations.
  + **Refine Features**: Incorporate macroeconomic or sector-specific indicators, especially for Services.
  + **Revisit FN:FP Ratio**: If operational overhead from false positives is too high, try lowering FN cost to 5 and tune threshold accordingly.
  + **Deployment**: Integrate into a firm-scoring dashboard that flags predicted fast-growers above 0.06 for deeper analysis.

In summary, this pipeline robustly identifies potentially fast-growing firms. By combining advanced data cleaning, feature engineering, and cost-sensitive machine learning, we pinpoint high-potential opportunities in both Manufacturing and Services,significantly reducing missed “blockbuster” firms relative to baseline.