

Summary Report: Strategic Targeting of High-Growth Firms

Date: February 2026

To: Data Science Team Leaders & Senior Management

From: Yllke Berisha & Bo Wang

Subject: Predictive Modeling for 2013 Venture Capital Allocation

1. Executive Summary

Objective

To identify high-potential firms ("unicorns") using 2012 financial data, enabling an optimized investment strategy for 2013 that maximizes returns while managing due diligence costs.

Key Findings

- **Best Model:** The **Random Forest (RF)** model outperformed Logistic Regression and LASSO, achieving the highest predictive power (**AUC \approx 0.67**) and lowest expected financial loss.
- **Strategic Threshold:** We recommend an **aggressive investment strategy**. Instead of a standard 50% probability cutoff, we should investigate any firm with a probability of high growth greater than **18%**.
- **Sector Strategy:** The model is significantly more effective in the **Services** sector. We recommend a bifurcated strategy: aggressive entry in Services and a more conservative approach in Manufacturing.

2. Business Problem & Risk Appetite

The "Why": The Cost of Missing Out

In Venture Capital, the asymmetry of risk is extreme. Missing a "unicorn" (False Negative) is significantly more damaging to our portfolio than investigating a non-starter (False Positive).

To reflect this reality, we quantified our risk appetite:

- **Missed Opportunity Cost (FN):** 10 units (Lost ROI).
- **Wasted Investigation Cost (FP):** 2 units (Admin fees).
- **Ratio 5:1:** This dictates that our model must be **risk-seeking**—we are willing to tolerate more false alarms to ensure we capture the winners.

Decision Matrix (Cost Impact)

	Decision: Don't Invest	Decision: Invest
Actual: Low Growth	Cost = 0 (Correct Decision)	Cost = 2 (Wasted Due Diligence)
Actual: High Growth	Cost = 10 (Missed "Unicorn")	Cost = 0 (Successful Investment)

3. Methodology & Model Performance

Data Scope

We analyzed active firms with sales between €1k and €10m (excluding micro-enterprises and giants). "High Growth" is defined as sales growth exceeding **20%** in the following year.

Model Selection

We tested three approaches: Logistic Regression (Baseline), LASSO (Feature Selection), and Random Forest (Complex Interactions).

Performance Summary

Random Forest (RF) was the clear winner. It minimizes our business loss function better than any linear model.

Model	Description	CV AUC (Predictive Power)	CV Expected Loss (Lower is Better)
Logit M4	Simple Baseline	0.665	1.37
LASSO	Automated Selection	0.670	1.35
Random Forest	Champion Model	0.672	1.34

4. Key Drivers of Growth

Our analysis of the Random Forest model reveals that firm growth is driven by non-linear factors that simple regressions miss. The two strongest predictors are:

1. **Past Growth Momentum (`d1_sales_mil_log`):** The strongest predictor. Firms that have already started an upward trajectory are statistically more likely to continue it.
2. **Scale (`sales_mil`):** Current size significantly impacts future growth potential, capturing the reality that it is easier for smaller firms to double in size than for larger ones.

(Note: Technical plots such as Variable Importance and ROC Curves are available in the full technical report.)

5. The Decision: When to "Pull the Trigger"

Optimal Threshold Analysis

Standard classification models use a 50% threshold (i.e., "invest if it's more likely than not"). However, given our high cost of missing winners (Cost=10), a 50% threshold is **too conservative** and would result in missed opportunities.

We optimized the decision threshold to minimize total expected loss. The data dictates an **optimal threshold of ~18%**.

- **Interpretation:** If the model says a firm has an **18% chance**