Report: Predicting High-Growth Firms

Data and code available at: https://github.com/icecodesred/Data-Analysis-3/tree/main/Assignment2

Introduction and Target Definition

Objective:
Identify "high-growth" firms using 2012 firm-level data by predicting revenue growth from 2012 to 2014. A firm is deemed "fast growing" if it falls within the top 20% of revenue growth over this period.

2-Year Growth as an Indicator:

- **Smoother Trajectory:** Measuring growth over two years smooths out one-off events (e.g., temporary contracts or market expansions) and better reflects a firm's sustained performance.
- **Investment Alignment:** Investors typically assess a firm's medium-term outlook (2–3 years) to gauge its financial stability and operational efficiency, making a 2014 endpoint more representative of true growth potential.

Alternative Approaches:

- **One-Year Growth:** Using only 2013 vs. 2012 might capture transient spikes rather than lasting performance improvements.
- **Non-Revenue Metrics:** While asset, employee, or market share growth can be informative, revenue growth is a more direct indicator of operating performance and profitability.

By focusing on a two-year horizon and the top 20% of revenue growth, we capture meaningful and sustained expansion that is most relevant for investors.

Data and Features

We used a cleaned dataset of firms from 2012, with **43 predictors** that capture:

- **Financial Statement Variables:**
  For instance, sales, profit_loss_year, curr_assets, material_exp, personnel_exp, etc. These measure the firm's liquidity, operational efficiency, and profitability.
- **Balance Sheet Quality Indicators:**
  Such as balsheet_flag or balsheet_length to gauge reliability and completeness of reported financials.
- **Management/HR Features:**
  Including ceo_age, female, foreign_management. These can correlate with managerial style, innovation capacity, and organizational strategy.

- **Engineered Variables:**
  We also included aggregated or transformed features like total_assets_bs or inc_bef_tax_pl to standardize and compare across different firm sizes.

Our **target** variable, fast_growth, is **1** if the firm's 2014 revenue growth places it in the top 20% of all firms, and **0** otherwise.

Modeling Approach

1. **Train/Holdout Split (80/20):**
   - We split the data into 80% training and 20% holdout, ensuring we keep the holdout set completely unseen for final evaluation.
2. **Cross-Validation (on Training Set):**
   - We compared three models: **Lasso Logistic Regression**, **Random Forest**, and **XGBoost**.
   - We performed **5-fold stratified cross-validation** to maintain class balance across folds.
   - For each model and fold, we calculated:
     - **RMSE** (on predicted probabilities)
     - **ROC AUC**
     - **Expected Loss** using a custom cost function to reflect our **business objective** (detailed below).
3. **Cost Function (Loss):**
   - A missed high-growth firm (false negative) is **2.5 times** as costly as wrongly investing in a non-high-growth firm (false positive).
   - Mathematically, we set: \text{Loss} = \text{cost_fp} \times \text{FP} \;+\; \text{cost_fn} \times \text{FN}, with \text{cost_fn} = 2.5 \times \text{cost_fp}.
   - We searched thresholds from 0 to 1 to find the threshold that **minimized** this expected loss for each model and fold.
4. **Final Evaluation (Holdout Set):**
   - After identifying the best model (Random Forest) based on cross-validation loss, we retrained Random Forest on the **full 80% training data** and evaluated on the 20% holdout, examining confusion matrices, classification reports, and ROC curves.

Results

Cross-Validation Performance:

- **RMSE & AUC:**
  Logistic Regression, Random Forest, and XGBoost showed AUCs in the **0.68–0.72** range. RMSE values were similar, indicating comparable probability calibration.

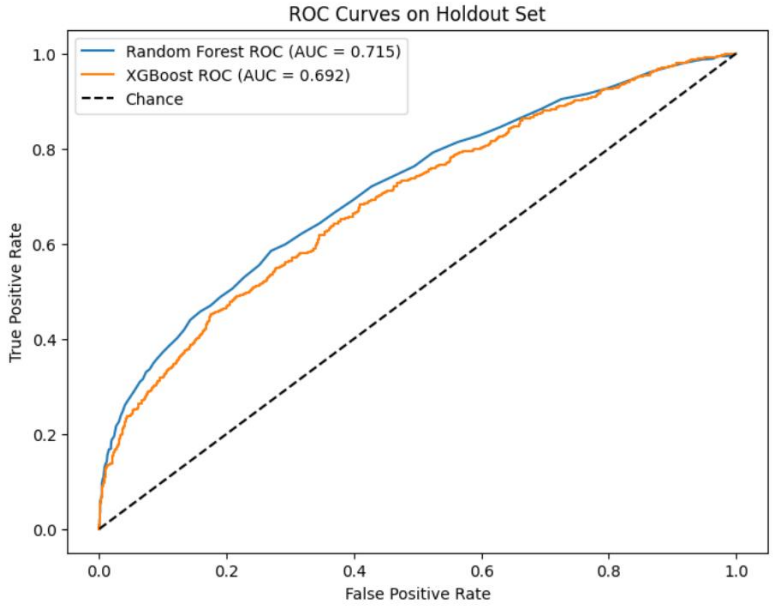
- **Expected Loss:**
    - Random Forest achieved the **lowest average expected loss (~2166)**, indicating it balanced false positives and false negatives best under our cost function.
    - The best thresholds for Random Forest were in the **0.20–0.30** range, reflecting the emphasis on avoiding missed high-growth firms.

Holdout Set Evaluation:

- **Confusion Matrix & Classification Report:**
    - Random Forest correctly identifies many high-growth firms but at the cost of labeling a substantial number of non-high-growth firms as "high growth." Given the high penalty for missing a truly top-growth firm, this trade-off is acceptable.
    - Overall accuracy is moderate (~48–50%), but more importantly, the model reduces false negatives relative to a higher threshold approach

ROC Curve

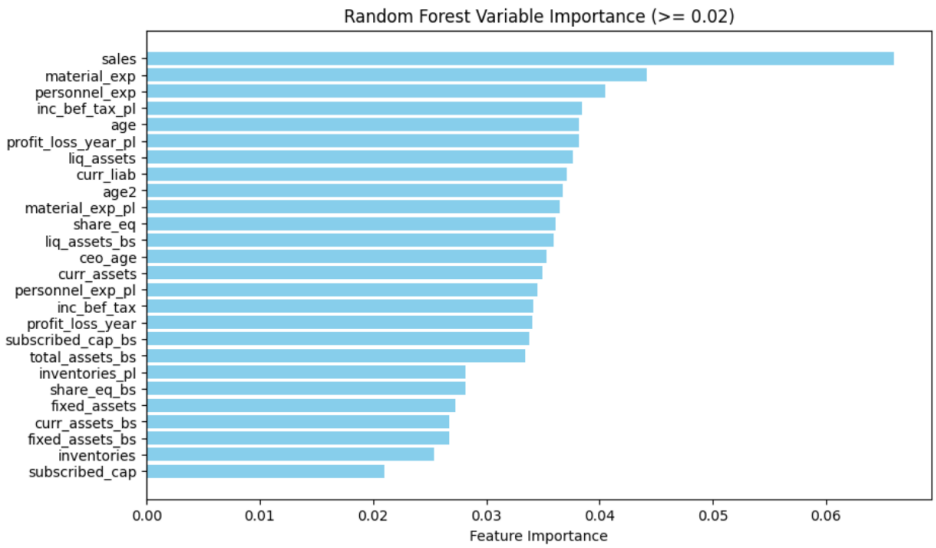
Random Forest's AUC on the holdout set is ~0.703, while XGBoost is ~0.697—both models show decent discriminative ability.



However, Random Forest's cross-validated expected loss was lower, so it remains the preferred model under our cost structure.

Feature Importance (Random Forest):

Sales, material_exp, and personnel_exp stand out as top features. This aligns with corporate finance theory: revenue and key operating expenses can signal future scalability and operational efficiency. Variables like inc_bef_tax_pl, profit_loss_year_pl, and curr_liab also appear among the more important predictors.



Interpretation and Corporate Finance Rationale

From a corporate finance perspective:

- **Revenue & Expense Structure:**
  Firms with higher 2012 sales and well-managed material and personnel expenses may have the **operational capacity** to expand quickly. This supports the notion that controlling costs while growing top-line revenue is critical to sustaining future growth.

- **Missed Opportunity Cost:**
  Setting a higher cost for false negatives aligns with the idea that **missing out** on a truly high-growth firm represents a large opportunity cost. If an investor fails to invest in a firm that ends up in the top 20% of growth, the forgone returns could be substantial.

- **Two-Year Growth Horizon:**
  By examining growth from 2012 to 2014, we capture more stable indicators of the firm's growth trajectory. A one-year measure might be too volatile, while a three- to five-year measure can be too far removed for many investment decisions, especially in fast-moving markets.

Recommendations

1. **Use Random Forest** at a threshold around **0.20–0.30** to flag firms likely to be top 20% growers. This threshold was tuned to minimize the custom loss function where missed opportunities (false negatives) are significantly penalized.

2. **Further Due Diligence:**

- Shortlist firms flagged as high growth and perform deeper financial and strategic analyses to confirm investment suitability.

3. **Expand Feature Set:**
    - Consider adding macroeconomic indicators or more detailed managerial/ownership data. Additional corporate finance measures (e.g., R&D intensity, capital structure ratios) might further refine predictions.
4. **Monitor & Retrain:**
    - As new data arrives (e.g., for 2013 or 2015 financials), retrain the model and re-optimize the threshold. If the cost structure changes (e.g., false negatives become even more expensive), the optimal threshold may shift further.

Conclusion

In summary, our analysis identifies firms with sustainable, high revenue growth using a two-year growth measure (2012–2014). By emphasizing a cost function that penalizes missed high-growth opportunities 2.5 times more than false positives, Random Forest proves to be the optimal model with an average expected loss of approximately 2166 and an AUC of 0.703 on the holdout set. This model provides a robust, cost-sensitive framework for screening firms and guiding subsequent in-depth investment analysis.