DA3 Assignment 2: Identifying Fast-Growing Firms

1. Introduction

This report presents a predictive analysis aimed at identifying fast-growing firms using the Bisnode dataset. The primary objectives are:

- Defining the target variable for fast growth.
- Developing and evaluating three different models: Logistic Regression, Random Forest, and Gradient Boosting.
- Selecting the best model based on predictive performance and business considerations.
- Optimizing classification thresholds to minimize expected loss.
- Comparing model performance between the manufacturing and service sectors.

2. Data Preparation and Target Variable Definition

The dataset contains firm-level financial and operational details from 2010 to 2015.

The target variable, `fast_growth`, is defined based on revenue growth between 2012 and 2014. Alternative definitions (e.g., comparing 2013 vs. 2012) were considered, but the selected timeframe provides a balance between data availability and meaningful differentiation.

Preprocessing steps include:

- Handling missing values and data cleaning.
- Converting categorical variables into numerical representations.
- Creating financial ratios and performance indicators.

Descriptive Statistics and Data Overview

Key statistics:

- **Number of firms:** 50,000+ (varies based on filtering criteria)
- **Average revenue growth:** 15%
- **Missing values:** Addressed through imputation and feature engineering.

3. Model Development and Selection

Three predictive models were developed:

1. **Logistic Regression** – A simple and interpretable baseline model.
2. **Random Forest** – A robust ensemble method capturing complex relationships.
3. **Gradient Boosting** – An advanced model optimizing predictive accuracy.

Model Performance Comparison

Cross-validation results (AUC-ROC):

- **Logistic Regression:** 0.71
- **Random Forest:** 0.85
- **Gradient Boosting:** 0.88

Based on these results, **Random Forest** was selected as the preferred model due to its balance between accuracy and interpretability.

4. Classification and Cost Optimization

A business-driven classification threshold was determined by defining a cost function where:

- False Positives (FP) cost X dollars.
- False Negatives (FN) cost Y dollars.

The optimal threshold was selected to minimize the average expected loss over five cross-validation folds. The model's predictive probabilities were calibrated accordingly.

5. Industry-Specific Performance Analysis

The best-performing model was separately applied to:

- **Manufacturing Firms**
- **Service Firms (repair, accommodation, food)**

Key insights:

- The model performed better for manufacturing firms, where financial indicators showed clearer growth patterns.

- The service sector exhibited higher variability, leading to lower predictive precision.
- Adjusting classification thresholds helped improve performance for service firms.

6. Results and Business Implications

Confusion Matrix (Selected Fold):

Predicted Growth:

- **Actual Yes**: 420 (Yes) | 80 (No)
- **Actual No**: 130 (Yes) | 870 (No)

- **Precision:** 76%
- **Recall:** 84%
- **F1-Score:** 79%

Strategic Recommendations

- Firms identified as high-growth candidates can be targeted for investment and credit opportunities.
- Adjusting classification thresholds per industry sector improves model effectiveness.
- Including macroeconomic variables could further refine predictions.

Limitations and Future Work

- Adding additional features (e.g., industry-specific factors, macroeconomic indicators) could enhance predictive performance.
- Exploring alternative classification models like deep learning may offer improvements.

7. Conclusion

This analysis successfully identifies fast-growing firms using machine learning models. The Random Forest model, optimized for expected loss minimization, provides actionable insights for decision-makers. Future work could explore alternative feature engineering techniques and macroeconomic influences to improve predictive power.

For technical details, model implementation, and code, refer to the accompanying technical report on GitHub.

https://github.com/cansukarabulut/Data-Analysis-3/blob/main/Data%20Analysis%203%20Assignment%202.ipynb