Summary Report - Link to Code in Github

Finding Fast Growing Firms

Muhammad Hamza Dhich 2304086 & Zunaira Pasha 2203204

Introduction

Data from the OSF website contained information about European Union firms from 2005-2016 in three industries: auto manufacturing, equipment manufacturing, and hotels and restaurants. It was constructed from publicly available sources by Bisnode for educational purposes. This report presents a predictive analysis for identifying fast-growing firms using financial and operational data. The objective is to develop machine learning models that accurately classify firms based on their growth potential. The dataset includes multiple financial indicators spanning several years, and the models aim to provide actionable insights for decision-makers.

This project aims to build models that predict fast-growing firms. A fast-growing firm is typically defined as a company that experiences a high compound annual growth rate (CAGR) over a certain period (2012-2014 in our case). The threshold of **20% CAGR** is used to classify firms into two groups: **fast-growing** and **not fast-growing**.

A total of seven models were used for the prediction: five logistic regression models (logit), LASSO regression, and Random Forest.

Data Preparation

The original dataset **"cs_bisnode_panel.csv"** includes **287,829 observations and 48 variables** covering the period **2005-2016**.

Only firms with sales between 1,000 and 10,000,000 were retained. Additionally, firms missing Compound Annual Growth Rate (CAGR) data or those exceeding the 99th percentile (~200% CAGR) were dropped. The **fast_growth** variable was created to classify firms based on the **20% CAGR threshold**. The firms were classified as 'fast-growing' if their CAGR for 2012-2014 was higher than 20%. A total of 118 variables (HR, quality indicators, balance sheet variables, sales etc) were created using multiple feature engineering and label engineering.

After preprocessing, the final dataset consisted of **10,462 observations and 118 variables**.

Dataset Stage	Observations	Variables	Period
Original Dataset	287,829	48	2005–2016
Cleaned Dataset	10,462	118	2012

The table below shows the number of variables that fell into each category:

Fast Growth Category	Number of Firms
0	7996
1	2497

Methodology:

Statistical Models Used

- Logistic Regression (Logit Models) – A set of five models with increasing complexity, incorporating various predictor variables.
- LASSO Regression – Used for feature selection and reducing overfitting.
- Random Forest (RF) – An ensemble learning method aimed at improving prediction accuracy.

Model Evaluation Metrics

- Root Mean Squared Error (RMSE) – Measures the average difference between predicted and actual values.
- Area Under the Curve (AUC) – Evaluates classification performance.
- Expected Loss Value – Evaluates on expected loss using a certain threshold/loss function.

1)Logistic regression models:

Five logistic regression models were constructed, increasing in complexity from Model 1 being the simplest one included only 12 predictors and Model 5 being the most complex contained 153 predictors:

Model	Variables Included
M1	Sales, Sales², Sales change, Profit/loss, Industry
M2	M1 + Fixed assets, Equity, Liabilities, Liabilities flags, Age, Foreign management
M3	M2 + Firm details, Financial & operational variables
M4	M3 + HR, Quality indicators
M5	M4 + Interaction terms

Model	Predictors	CV RMSE	CV AUC
M1	12	0.418	0.602
M2	19	0.416	0.623
M3	36	0.413	0.645
M4	79	0.413	0.648
M5	153	0.414	0.646

Based on evaluation metrics, choose **M4** from Logistic model for its superior performance, as it consistently achieves the lowest RMSE across folds, offering the best predictive accuracy. While slightly more complex than **M3**, its performance gains justify the added complexity, making it a more reliable choice than **M5**.

2)Logit LASSO Model

Taking the most complex model of logit – M5, having 153 predictors and putting it in LASSO Model. It shrinks the number of predictors from 153 to 57.

Model	**Number of Coefficients**	**CV RMSE**	**CV AUC**
M5	153	0.414	0.646
LASSO	57	0.412	0.653

As seen from the table above, LASSO performs marginally better than Logit Models of M5 and M4 with slightly improved RMSE of 0.412 and higher AUC of 0.653. In terms of prediction accuracy, LASSO marginally outperforms logit models.

3)Random Forest Model

Using the best parameters (parameters with lowest RMSE and highest AUC – 6 maximum features and 16 minimum samples splits) through cross validation, Random Forest was run, results of which are below:

Model	**CV RMSE**	**CV AUC**
RF	0.412199	0.654405

It can be seen that **Random Forest** showed better results for both RMSE and AUC among all the models including 5 logit models and LASSO. **Random Forest** has the best predictive performance.

Combined Models' Performance:

To evaluate the performance of each model the Root mean squared error (RMSE) and Area Under the Curve (AUC) were used. The table below displays the results for all models:

	Number of	CV	CV
Model	Coefficients	RMSE	AUC
M1	12	0.4184	0.6022
M2	19	0.4163	0.6232
M3	36	0.4135	0.6450
M4	79	0.4127	0.6481
M5	153	0.4141	0.6460
LASSO	57	0.4123	0.6527
RF	n.a.	**0.4122**	**0.6544**

Key Findings:

- **Random Forest achieved the highest CV AUC (0.654405)**, making it the best model for identifying fast-growing firms.
- Logistic regression models showed improved performance with increasing complexity, but Model 4 (M4) was the best-performing among them.
- The LASSO model reduced the number of predictors while maintaining competitive performance.
- **Random Forest outperformed all models in both RMSE and AUC**, making it the preferred choice.

Classification:

Defining Loss Function - Cost of False Negatives vs. False Positives

The business problem is to identify fast-growing firms for investment, where the goal is to target firms with a CAGR greater than 20%. A **10:1** ratio is chosen because missing a fast-growing firm (False Negative) leads to significant revenue loss, while the cost of pursuing a non-growth firm (False Positive) is typically lower and can be managed. This ratio strikes a right balance by ensuring the model focuses on identifying growth firms without being overly cautious, which could lead to missing profitable opportunities, as a higher penalty (like 20:1) might overly constrain the model.

Model Performance Summary:

To evaluate the chosen **Random Forest model's** performance on the holdout set, let's look at the confusion table.

	Predicted no fast growth	Predicted fast growth
Actual no fast growth	83	1498
Actual fast growth	7	511

Based on the table above: Accuracy = **28.3%** Sensitivity = **98.7%** Specificity = **5.2%**

The RF model is excellent at identifying fast-growing firms, with a sensitivity of 98.7%, meaning it correctly identifies almost all firms that are growing fast. However, this comes at the expense of a low specificity (5.2%), which means it has a high rate of false positives. In practical terms, the model labels many firms as fast-growing when they are not, leading to potential inefficiencies or incorrect decisions.

Business Goal Alignment: High sensitivity aligns with the goal of identifying fast-growing investment opportunities. However, Excessive False Positives could at times lead to inefficient investment decisions and wasted resources.

Model Performance by Industry:

The preferred Random Forest model was further evaluated by fitting it on sub datasets of manufacturing and service industry, and the results were evaluated. Using the same loss function of 10:1:

- **Manufacturing Industry:**

	Predicted no fast growth	Predicted fast growth
Actual no fast growth	14	373
Actual fast growth	3	157

In the manufacturing sector, the model demonstrated high sensitivity, correctly identifying **157** fast-growing firms. However, it also exhibited a high false positive rate, misclassifying **373** non-fast-growing firms as fast-growing. This resulted in an accuracy of **31.3%**.

- **Service Industry:**

	Predicted no fast growth	Predicted fast growth
Actual no fast growth	78	1118
Actual fast growth	8	348

Similarly, in the service sector, the model identified 348 fast-growing firms but incorrectly classified 1118 non-fast-growing firms, leading to an accuracy of **27.4%**. This indicates a significant tendency to over-predict growth in this sector as well.

The following table compares RF model's across three datasets:

Metric	Total Dataset	Manufacturing	Service Industry
Accuracy	28.3%	31.3%	27.4%
Sensitivity	98.7%	98.1%	97.8%
Specificity	5.2%	3.6%	6.5%

Business Implications and Final Remarks:

The Random Forest model chosen performs consistently across all three datasets, showing similarly **high sensitivity (~98%)** and **low specificity (3.6%–6.5%)** in each industry. While it effectively identifies fast-growing firms, the high false positive rate lowers overall accuracy (27%–31%).

These results show a tendency for the model to **over-predict growth**, aligning with the **10:1 loss function.** While the model's high sensitivity is beneficial for identifying growth opportunities, the large number of false positives could lead to inefficient resource allocation and increased investment risks. It is crucial to consider the trade-off between capturing all potential growth and minimizing the risk of misclassifications