Data Analysis 3: Assignment 3 - Summary Report

Viet Nguyen, Marcell Magda

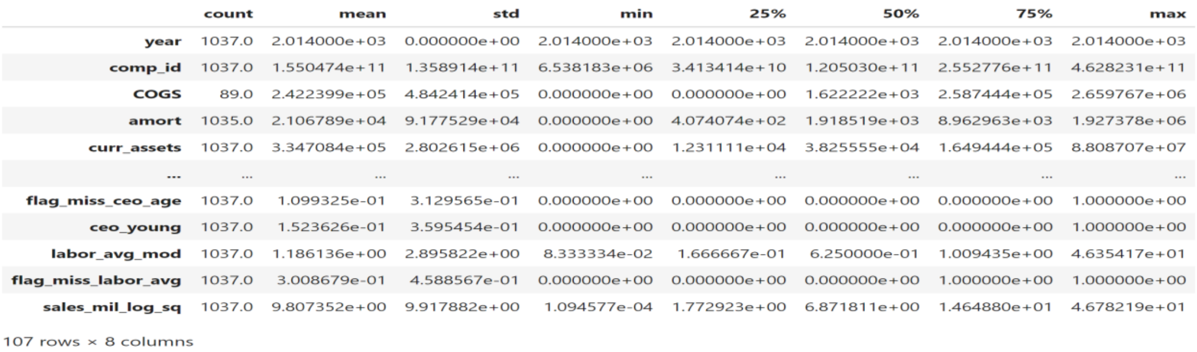
Github Repo

Objective

The core aim of this assignment is to develop a predictive model capable of identifying firms within the "Manufacture of computer, electronic, and optical products" sector that defaulted in the year 2015. Specifically, the model aims to predict defaults among small and medium enterprises (SMEs) with annual sales ranging from 1,000 to 10 million EUR. By analysing historical financial and operational data from 2014, the model seeks to discern patterns and indicators that distinguish firms that ceased operations in 2015. The ultimate goal is to provide stakeholders with a reliable tool for assessing default risk, enabling more informed decision-making in financial planning, risk management, and strategic investment within this industry. This project not only contributes to academic knowledge in predictive modeling and financial risk assessment but also offers practical insights for businesses and policymakers to support the sustainability and growth of SMEs in the high-tech manufacturing sector.

Exploratory Data Analysis

The EDA phase was instrumental in evaluating the financial health and operational status of SMEs in the "Manufacture of computer, electronic, and optical products" sector. It shed light on sales distributions and the prevalence of default among firms, highlighting significant variations and outliers. The analysis revealed a higher default rate among firms at the lower end of the sales spectrum, leading to the strategic removal of extreme outliers to enhance the predictive model's accuracy. This adjustment was pivotal, as initial models without this refinement exhibited significantly higher error rates. Furthermore, the exploration into the industry classification 'ind2 == 26' provided nuanced insights into sector-specific risks and operational characteristics. This detailed examination of categorical variables, including firm size and industry segment, enriched our understanding of default patterns, thereby refining the model's focus and potential applicability to broader or more specific client needs. The inclusion of diverse firm characteristics ensures the model's adaptability to varying investment strategies, including those contemplating diversification into related technological sectors.



	count	mean	std	min	25%	50%	75%	max
year	1037.0	2.014000e+03	0.000000e+00	2.014000e+03	2.014000e+03	2.014000e+03	2.014000e+03	2.014000e+03
comp_id	1037.0	1.550474e+11	1.358914e+11	6.538183e+06	3.413414e+10	1.205030e+11	2.552776e+11	4.628231e+11
COGS	89.0	2.422399e+05	4.842414e+05	0.000000e+00	0.000000e+00	1.622222e+03	2.587444e+05	2.659767e+06
amort	1035.0	2.106789e+04	9.177529e+04	0.000000e+00	4.074074e+02	1.918519e+03	8.962963e+03	1.927378e+06
curr_assets	1037.0	3.347084e+05	2.802615e+06	0.000000e+00	1.231111e+04	3.825555e+04	1.649444e+05	8.808707e+07
...	...	...	...	...	...	...	...	...
flag_miss_ceo_age	1037.0	1.099325e-01	3.129565e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
ceo_young	1037.0	1.523626e-01	3.595454e-01	0.000000e+00	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
labor_avg_mod	1037.0	1.186136e+00	2.895822e+00	8.333334e-02	1.666667e-01	6.250000e-01	1.009435e+00	4.635417e+01
flag_miss_labor_avg	1037.0	3.008679e-01	4.588567e-01	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00	1.000000e+00
sales_mil_log_sq	1037.0	9.807352e+00	9.917882e+00	1.094577e-04	1.772923e+00	6.871811e+00	1.464880e+01	4.678219e+01

107 rows × 8 columns

Feature Engineering

In the feature engineering phase, we streamlined the dataset for modelling. Industry codes were simplified for better interpretability, grouping firms into broader categories for easier analysis. Firm demographics, including squared age, foreign management presence, and regional locations, were refined to highlight their impact on default risk. Negative asset values were corrected, and a comprehensive total assets variable was introduced to ensure data integrity and enhance financial health assessment. Financial ratios derived from profit/loss and balance sheet metrics were standardized, with flags for outliers improving model accuracy. CEO age was normalized with bounds and missing values imputed, alongside the introduction of a young CEO indicator, and labour averages were adjusted. Additionally, key numerical variables were transformed into categorical ones, including industry categories and default flags, facilitating nuanced analysis.

Predictors Used (grouped by function for clarity)

Financial Ratio Related

Financial ratios such as gross profit margin (n_gross_profit_margin), net profit margin (n_net_profit_margin), return on equity (n_return_on_equity), debt-equity ratio (n_debt_equity_ratio), current ratio (n_current_ratio), quick ratio (n_quick_ratio), and return on assets (n_return_on_assets) are crucial in assessing a firm's financial health and operational efficiency. These metrics offer insights into profitability, financial leverage, liquidity, and asset efficiency, serving as key indicators for predicting firm defaults.

Sales Related

Sales metrics encompass original sales figures and their transformed versions, such as logarithmic transformations (n_sales_mil_log, n_sales_mil_logsq), alongside year-over-year growth (n_yoy_growth). These variables capture the revenue generation capacity and growth trajectory of firms, vital for understanding their market position and sustainability.

Firm Function Related

Variables like firm lifespan (n_day_alive and its square n_day_alive_sq) and indicators of asset issues (flag_asset_problem) reflect on the operational aspects and health of a firm. The lifespan captures the maturity and potentially the resilience of a firm, while asset problem flags serve as direct indicators of financial mismanagement or operational challenges, both of which are pivotal in assessing default risk.

Corporate Governance Related

Aspects of corporate governance, such as foreign management presence (f_foreign_management) and the count of CEOs (n_ceo_count), shed light on the firm's stability and decision-making structures. Foreign management involvement can indicate international standards of operation and potentially higher resilience, whereas the number of CEOs could reflect on the firm's governance stability.

Other

Additional variables like CEO age (n_ceo_age), gender representation (n_female, f_gender), tenure (n_inoffice_days), and demographic and location factors (f_origin, f_urban_m, f_region_m) provide a broader context on the demographic characteristics of leadership, employee engagement, and the geographical and cultural positioning of firms.

Model Development and Selection – Running on Training data

Logit

The Logit model utilizes 107 predictors with no regularization, aiming for simplicity and interpretability. Its performance metrics suggest moderate predictive ability with a Cross-Validation (CV) Root Mean Square Error (RMSE) of 0.278230 and an Area Under the Curve (AUC) of 0.652518, indicating fair discrimination ability. The optimal threshold for classification is set at 0.488771, with an expected loss of 0.909374. This model serves as a baseline, highlighting the potential for improvement through more complex models or regularization techniques.

Logit with LASSO

The Logit with LASSO model significantly increases complexity by incorporating 5778 predictors through polynomial features and LASSO regularization to reduce overfitting and enhance feature selection. It shows improved predictive performance with a CV RMSE of 0.232864 and a CV AUC of 0.777904, demonstrating better overall model performance compared to the basic Logit model. The optimal threshold is notably lower at 0.140603, with an expected loss reduced to 0.787068. This model's complexity and improved metrics indicate its effectiveness in handling high-dimensional data and making more nuanced predictions.

Random Forrest

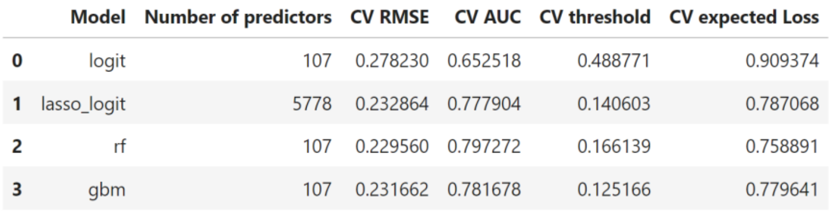
The Random Forest model uses 107 predictors, similar to the basic Logit model, but leverages the ensemble learning technique to improve prediction accuracy and model robustness. It achieves the best performance among the evaluated models with a CV RMSE of 0.229560 and a CV AUC of 0.797272, indicating its superior discriminative power. The optimal threshold for classification decisions is 0.166139, with the lowest expected loss of 0.758891 among the models. This model's balance between simplicity in terms of predictors, high accuracy, and low expected loss makes it particularly appealing for practical applications.

GBM (Generalized Boosted Regression Model)

The GBM model also uses 107 predictors. It optimizes for depth, learning rate, and minimum samples per leaf, achieving a CV RMSE of 0.231662 and a CV AUC of 0.781678. These metrics suggest that GBM is a strong performer, slightly trailing the Random Forest model in terms of AUC but offering competitive predictive power. The optimal threshold for GBM is set at 0.125166, with an expected loss of 0.779641. GBM stands out for its capacity to incrementally improve upon weak learners and adaptively focus on difficult-to-classify observations, making it a robust choice for complex datasets.

Results and Interpretation – Training Data

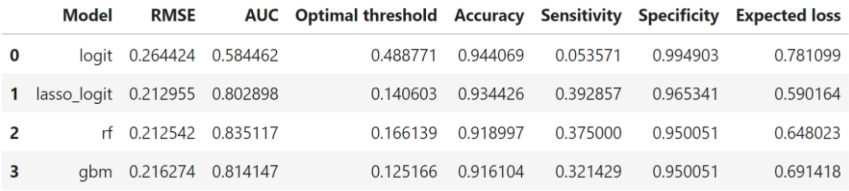
The evaluation of the Logit, Logit with LASSO, Random Forest, and GBM models highlights the Random Forest as the optimal choice for our prediction task. With an AUC of 0.797272 and the lowest expected loss of 0.758891, it outshines the alternatives in predictive power and efficiency. Using only 107 predictors, it balances high accuracy with simplicity, making it the preferred model for achieving effective and efficient outcomes in practical applications.



	Model	Number of predictors	CV RMSE	CV AUC	CV threshold	CV expected Loss
0	logit	107	0.278230	0.652518	0.488771	0.909374
1	lasso_logit	5778	0.232864	0.777904	0.140603	0.787068
2	rf	107	0.229560	0.797272	0.166139	0.758891
3	gbm	107	0.231662	0.781678	0.125166	0.779641

Results and Interpretation - Holdout Sample

The Lasso Logit model emerges as the preferable option based on the validation results. With the lowest expected loss of 0.590164, it demonstrates superior performance. While the Random Forest model boasts the highest AUC of 0.835117, indicating strong discriminative power, its expected loss of 0.648023 is higher compared to Lasso Logit. Lasso Logit also achieves a commendable balance with an AUC of 0.802898, showcasing its robustness in distinguishing between classes, and a notable sensitivity of 39.2857%, indicating its effectiveness in correctly identifying positive cases. Therefore, when prioritizing the minimization of expected loss in the model selection process, Lasso Logit stands out as the most appropriate choice.



	Model	RMSE	AUC	Optimal threshold	Accuracy	Sensitivity	Specificity	Expected loss
0	logit	0.264424	0.584462	0.488771	0.944069	0.053571	0.994903	0.781099
1	lasso_logit	0.212955	0.802898	0.140603	0.934426	0.392857	0.965341	0.590164
2	rf	0.212542	0.835117	0.166139	0.918997	0.375000	0.950051	0.648023
3	gbm	0.216274	0.814147	0.125166	0.916104	0.321429	0.950051	0.691418

Recommendation and Conclusion

Upon evaluating the models across training and holdout sets, the Lasso Logit model is recommended for its lowest expected loss of 0.590164, crucial for minimizing cost and risk. Despite Random Forest's higher AUC, Lasso Logit's balanced sensitivity and specificity, alongside its robust performance metrics, make it suitable for scenarios where managing classification trade-offs is key. This choice underscores the principle that training performance doesn't always predict real-world effectiveness, highlighting the need for validating multiple models. The validation phase is critical, revealing potential strengths or weaknesses not evident during training. Therefore, despite the time and resources required, testing multiple models ensures the selected one aligns with operational objectives, minimizing the risk of poor generalization. This comprehensive approach to model selection, prioritizing expected loss and a thorough validation process, ensures a cost-effective, reliable outcome for predictive tasks.