DA3 - Assignment 3 | GitHub

Summary Report: Predictive Modeling of Defaulted Firms in the Computer, Electronic, and Optical Products Industry

By: Dávid Szabados & Hla Myitzu

Executive Overview

This report synthesizes findings from a comprehensive technical study aimed at predicting defaults among small and medium enterprises (SMEs) within the 'Manufacture of computer, electronic, and optical products' industry in 2015, utilizing 2014 data. Leveraging advanced data analysis and modeling techniques, the study navigated through extensive data preprocessing, exploratory analysis, and the application of statistical and machine learning models. The culmination of these efforts was the development of a predictive model that significantly outperforms traditional analysis methods, with the Random Forest model showcasing superior accuracy and efficiency.

Methodological Framework

The study embarked on a structured data analysis pipeline, beginning with the setup of a Python environment equipped with libraries essential for data manipulation, visualization, and modeling. Notable steps in the process included:

- **Data Preprocessing**: A rigorous phase of cleaning, filtering, and transforming data to ensure a robust foundation for analysis. This involved handling missing values, creating new variables to better capture company dynamics, and adjusting financial figures to facilitate accurate comparisons.
- **Feature Engineering**: The development of new variables that encapsulate the characteristics and behaviors of firms, such as company age adjustments, log-transformed sales features, and financial anomaly flags. These engineered features aimed to enrich the models with nuanced insights into firm performance and risks.
- **Model Development**: The exploration of various predictive models, including logistic regression with cross-validation and LASSO regularization, culminating in the adoption of a Random Forest model. This model was meticulously tuned through hyperparameter optimization and evaluated against key performance metrics such as RMSE (Root Mean Squared Error) and AUC (Area Under the ROC Curve).

Variable Sets

Multiple sets of variables are created to assign them on different models later on. These sets includes raw variables, which are the basic variables freshly from the dataset, quality variables, summary variables (for example *total_assets_bs*), flags, yearly changes, human resource variables (including ceo age, ceo count), firm attributes, and some interactions between these variables for the LASSO model.

Key Findings and Implications

	Number of Coefficients	CV RMSE	CV AUC	CV treshold	CV expected Loss
M1	18.00000	0.23230	0.71975	0.08296	0.43497
M2	25.00000	0.23031	0.75757	0.09764	0.40213
M3	30.00000	0.37454	0.67411	0.41848	0.50185
M4	74.00000	0.37352	0.68202	0.40976	0.48323
M5	89.00000	0.48771	0.65534	0.50000	0.52345
LASSO	31.00000	0.22722	0.77961	0.08248	0.38489
RF	n.a.	0.22356	0.81351	0.11113	0.35892

- **Random Forest Superiority**: The Random Forest model emerged as the most effective tool for predicting firm defaults, achieving a CV RMSE of 0.224, a CV AUC of 0.814, and the lowest expected loss among evaluated models. Its ability to handle complex interactions between variables and robustness against overfitting makes it particularly valuable for risk assessment in the financial domain.
- **Model Performance**: Evaluation metrics revealed the model's strong predictive accuracy and economic impact, with the expected loss metric providing a direct link to financial implications. This underscores the model's potential to guide decision-making processes and risk management strategies effectively.
- **Operational Considerations**: While the Random Forest model offers significant advantages in accuracy and cost-effectiveness, considerations around computational demand, model interpretability, and the necessity for periodic updates must be factored into its deployment strategy.

Conclusion

The study's outcome highlights the power of leveraging advanced data analytics and machine learning to predict firm defaults within the manufacturing sector. The Random Forest model stands out for its exceptional performance, offering a reliable and efficient tool for financial risk assessment. As organizations strive to navigate the complexities of the financial landscape, such predictive capabilities become indispensable in informing strategic decisions and fostering a proactive approach to risk management.