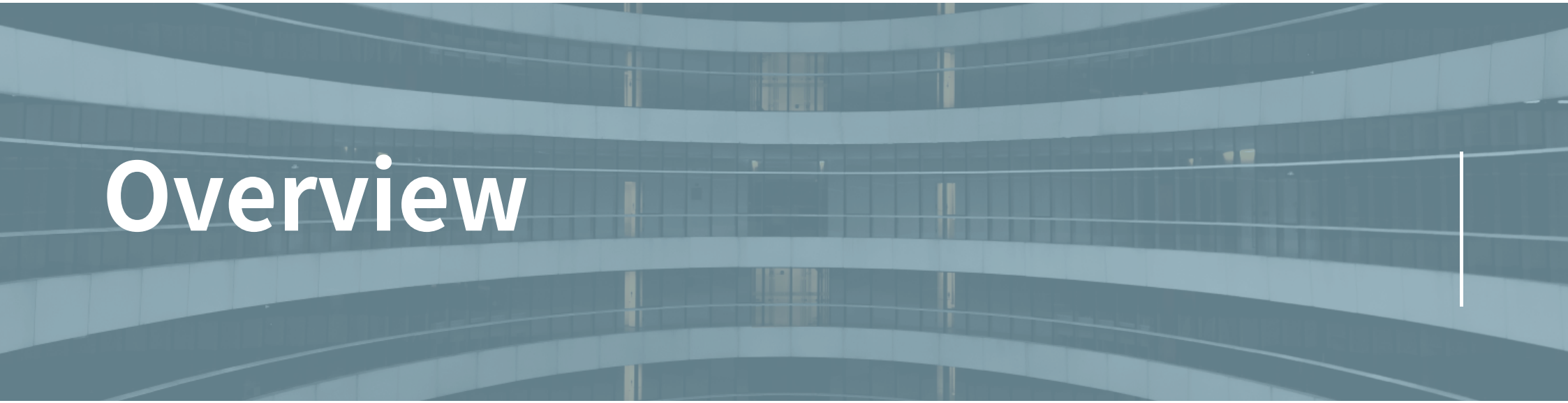
Sep 2022 - Nov 2022

Competition Project

Small and Medium-sized Enterprises Closure Prediction Project

Nayeon Kwon, Younghoon Yoo



Overview

About the competition

The competition is organized by NICE DNB, Korea's largest commercial credit bureau.

Objective

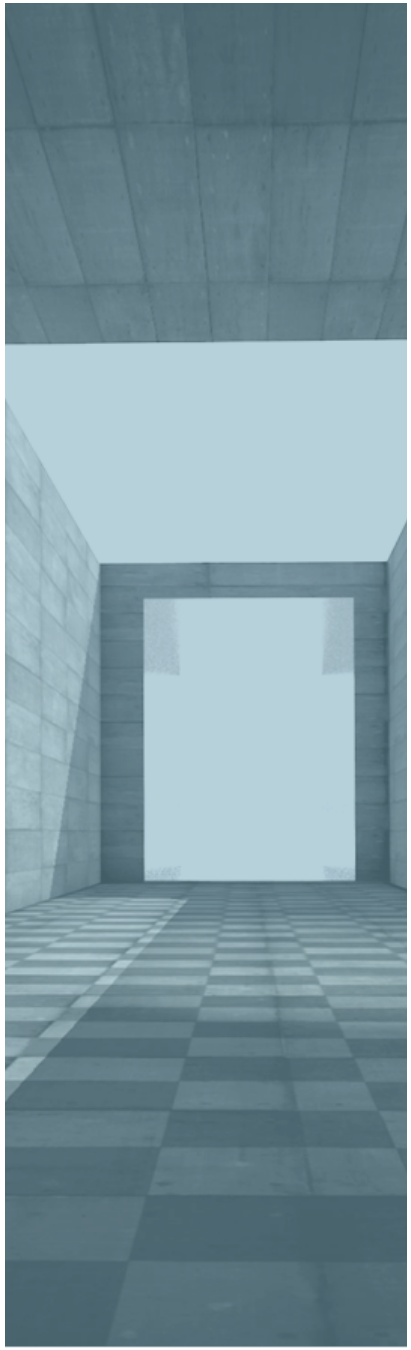
The objective is to construct an SME commercial credit model by developing machine learning models that predict SME closures.

Approach

The approach for this project entails leveraging non-financial data and implementing machine learning models to accurately classify and predict SME closures.

Contribution

My roles in this project include finding appropriate non-financial data, performing data pre-processing, and providing support in building machine learning models.



Background

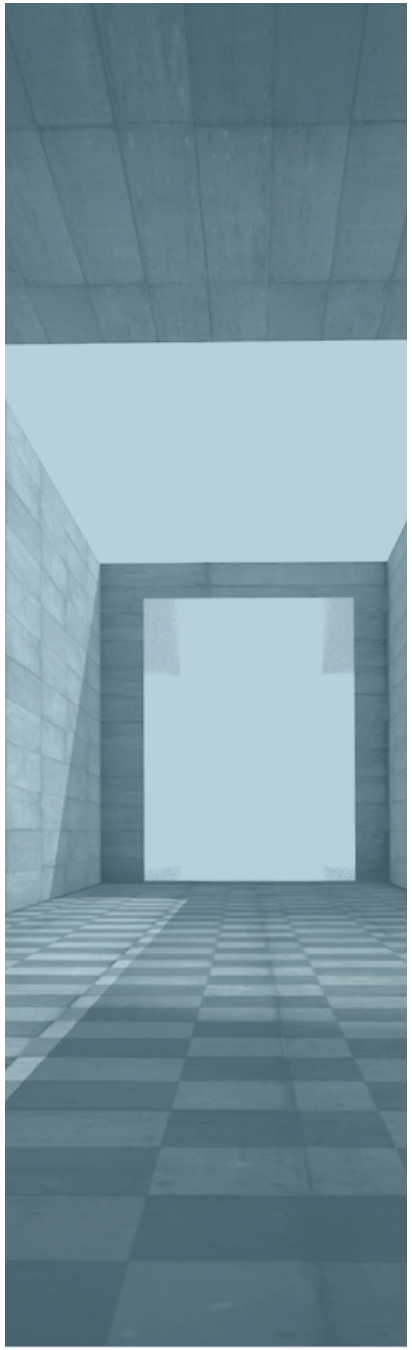
The Importance of Non-Financial Factors in SME Credit Ratings and Stakeholders' Interests in Utilizing Non-Financial Information

Why Non-Financial Factors Should Be Considered in SME Credit Ratings:

- Although financial factors provide objective information, their limited scope hinders accurate credit rating and corporate insolvency prediction.
- Non-financial factors play a crucial role in evaluating small and medium-sized enterprises (SMEs) due to their financial vulnerability and sensitivity to changing conditions, making them less reliable compared to larger enterprises.
- Sample Company Size: Approximately 75% of the companies in the sample have 58 employees or less.

Stakeholders' Interests in Utilizing Non-Financial Information for Diversified Credit Scoring in SMEs:

- Banks: Employing various non-financial information enables systematic measurement and management of borrowers' credit risk.
- SMEs: SMEs with limited assets, but possessing technological capabilities for generating future profits, can actively secure loans through the utilization of non-financial information.

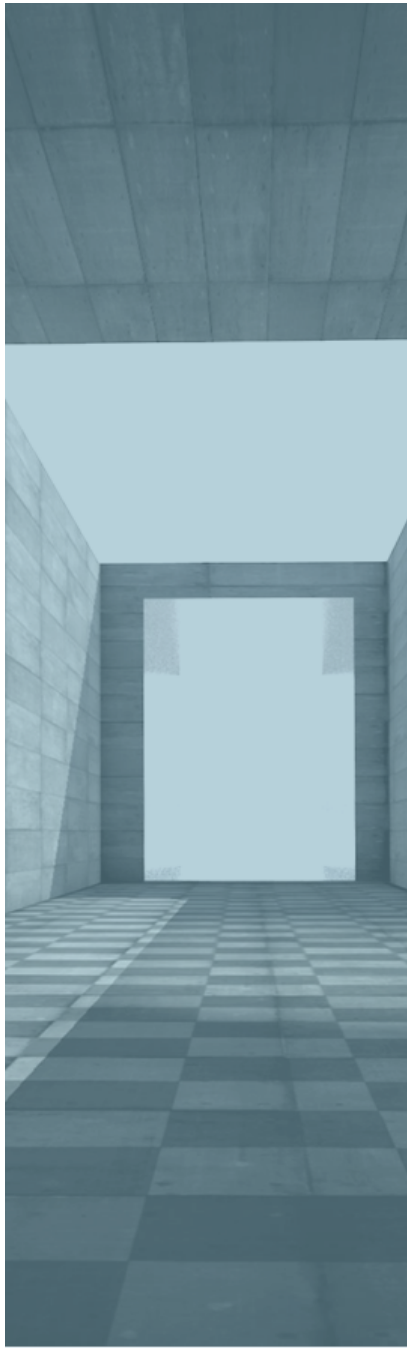


Data

Data Sources for Model Prediction

Data:

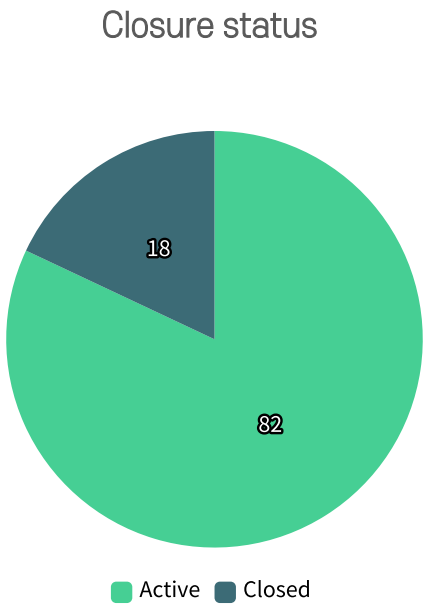
- We utilized monthly Business Trends and Outlook Survey data (covering overall economy, production, domestic sales, etc.) from January 2018 to June 2022, which was provided by the Ministry of SMEs and Startups in South Korea. Specifically, we focused on using the SBHI(Small Business Health Index) for companies categorized under the manufacturing industry.
- SBHI: The Small Business Health Index (SBHI) is calculated by assigning weights to responses on a five-point scale. A value above 100 indicates that more businesses anticipate improvement in their business over the next month, while a value below 100 suggests the opposite. This index helps assess the overall health and outlook of small businesses.
- We also utilized the open data and financial data provided by NICE DNB.

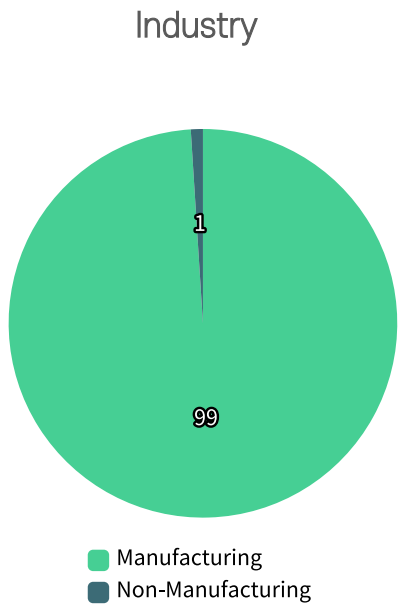


EDA

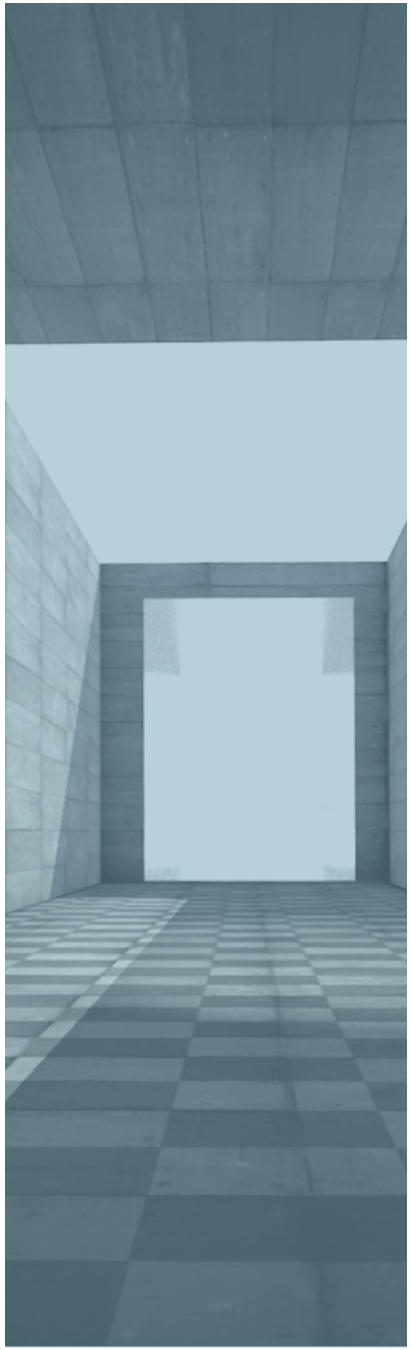
Exploratory Data Analysis: Uncovering Insights and Patterns

Summary Statistics:





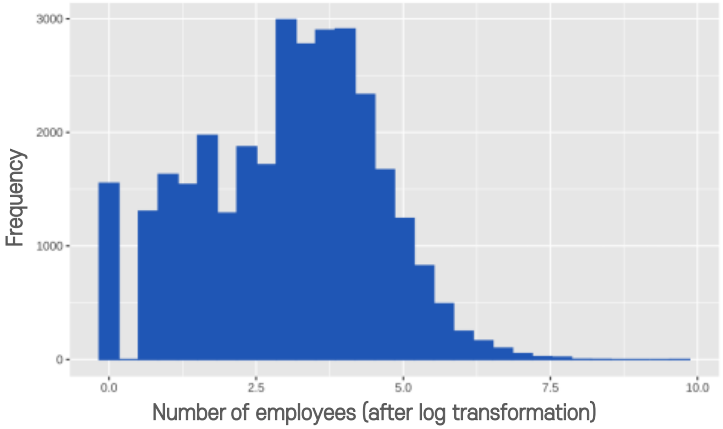
For detailed summary statistics of the variables, please click here: Summary Statistics



EDA

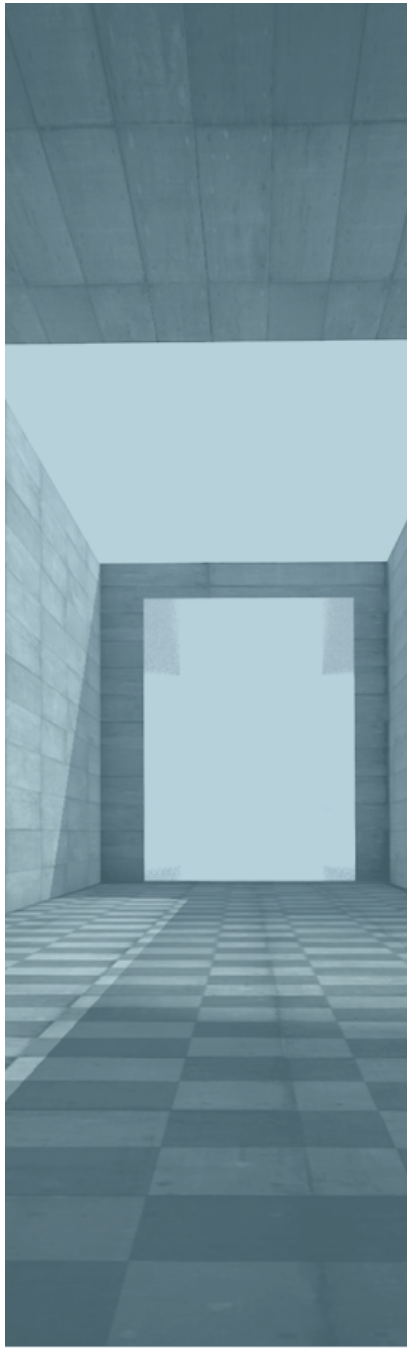
EDA (cont)

Exploratory Data Analysis: Uncovering Insights and Patterns



- The histogram showcases the distribution of the number of employees in a company after log-transformation.
- The histogram exhibits skewness with a long right tail, indicating that most companies fall into the small size category.

- A word cloud was created using the main business content variable from the open data.
- The word cloud highlights the most important words associated with the businesses in our dataset.
- Prominent words in the word cloud: manufacturing, real estate, sales, leasing, and wholesale and retail.

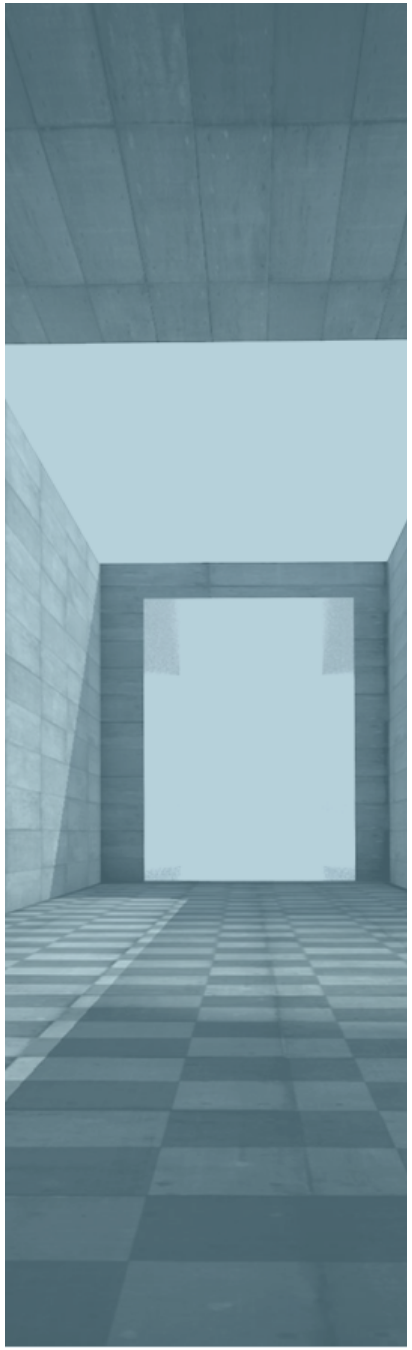


Data Preprocessing

Data Preprocessing: Integration of Open, Financial, and Non-Financial Data for SMEs in the Manufacturing Industry

Data Preprocessing:

- Predictive model applied to: Small and Medium-sized Enterprises (SMEs) in the manufacturing industry.
- Response variable definition: The response variable was defined by dichotomizing the closure status. Active SMEs were assigned a value of 0, while currently closed SMEs were assigned a value of 1, transforming it into a binary classification problem.
- Data merging: We merged three datasets, including the open data and financial data provided by NICE DNB, along with the non-financial data we collected independently.
- Total number of observations: After data preprocessing, we obtained a total of 32,392 observations with 28 variables.



Model Implementation

Model Implementation and Evaluation Metrics Using R Language

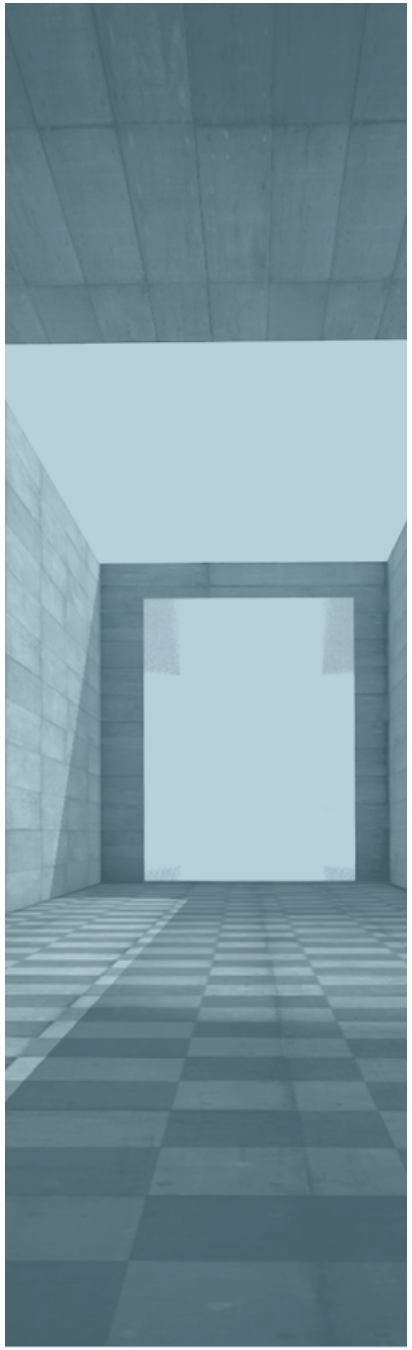
ML Models Utilized (in R language):

- Random Forest
- CatBoost
- BART

Evaluation Metrics:

- AUROC (Area Under the Receiver Operating Characteristic curve)
- Accuracy
- F1 score

The model implementation, conducted in the R language, involved leveraging the Random Forest, CatBoost, and BART packages. The evaluation of model performance was based on metrics such as AUROC, Accuracy, and F1 score.

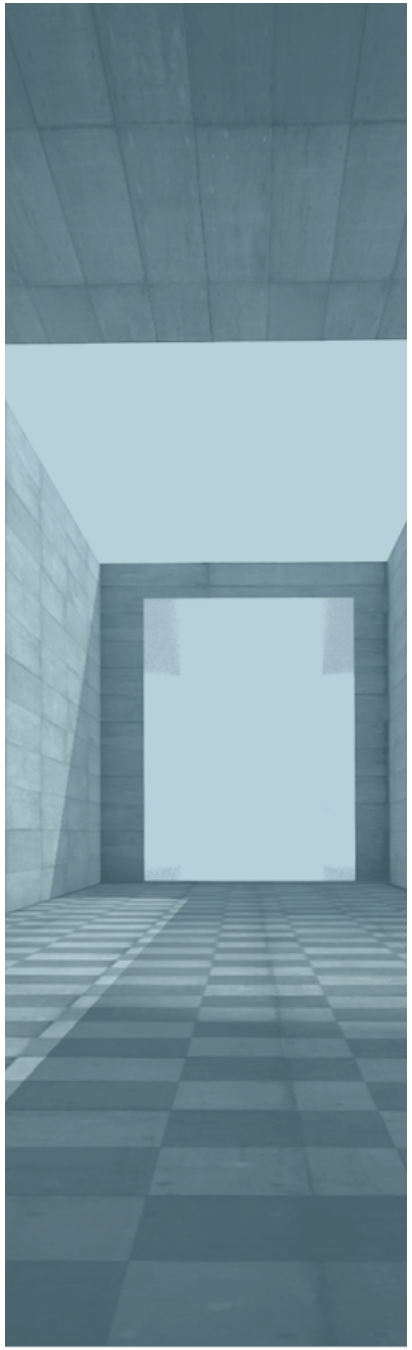


Model Performance

Evaluation of Predictive Model Performance

	AUROC	F1 score	Accuracy
Baseline	0.556	0.964	0.940
RF+	0.511	0.991	0.985
CatBoost	0.504	0.992	0.987
BART	0.536	0.977	0.961

Model Evaluation Results: Performance Metrics of Predictive Models

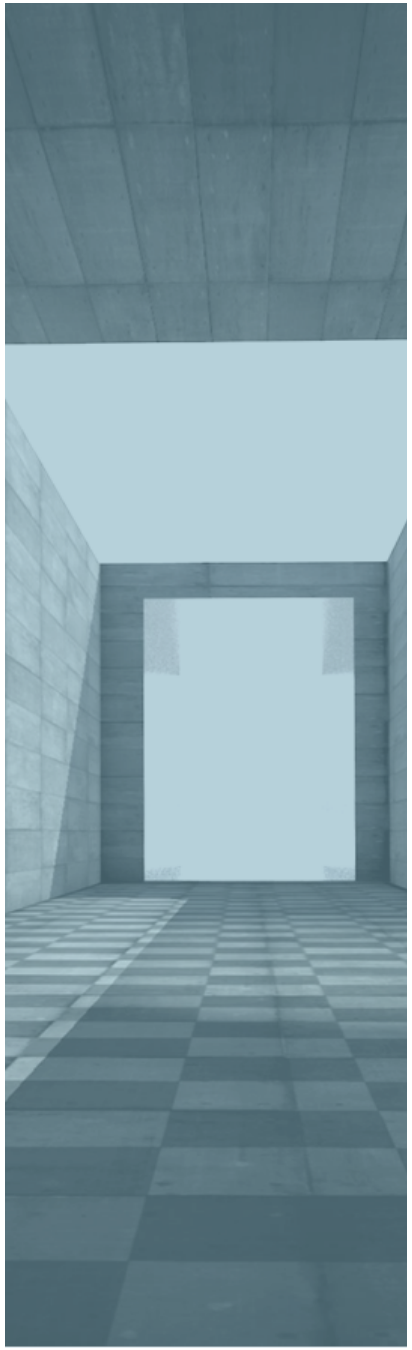


Model Performance (cont)

Evaluation of Predictive Model Performance

Summary:

- AUROC: The predictive models exhibited AUROC values ranging from 0.5 to 0.6, with the baseline model achieving the highest AUROC value of 0.556. While the AUROC does not indicate strong model performance, it should be noted that AUROC values can vary depending on the specific data and task.
- F1 score: Among the models, the CatBoost model achieved the highest F1 score of 0.992, while all models demonstrated F1 scores close to 1. This suggests that the prediction model performs well when evaluated based on the F1 score.
- Accuracy: Similar to the F1 score, the CatBoost model yielded the highest accuracy value of 0.987. Furthermore, all models exhibited an accuracy above 0.9, indicating that the proportion of correctly predicted data points was greater than or equal to 0.9 for each model.

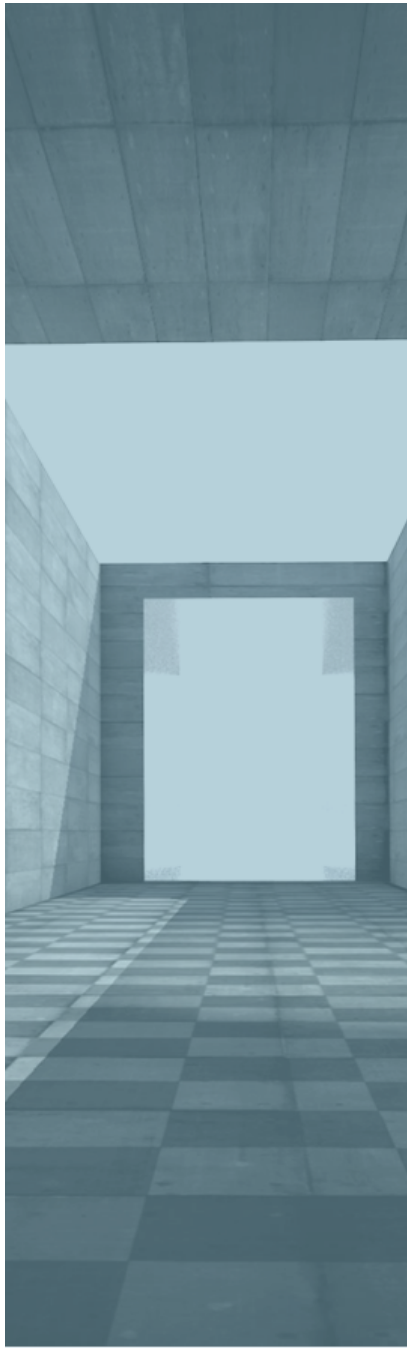


Conclusion

Key Findings, Insights, and Limitations: Concluding Remarks

Key Findings:

- Financial data variables were more influential in predicting SME closures compared to the open data variables.
- Incorporating non-financial data, specifically SBHI, led to improvements in F1 score and Accuracy compared to not using it.
- Among the non-BART models, total assets emerged as the most important variable, whereas in the BART model, employment level performance SBHI took precedence.
- All models demonstrated the significance of at least eight SBHI variables.
- This result suggests the potential utilization of the Ministry of SMEs and Startups' Business Trends and Outlook Survey in SME credit rating models.

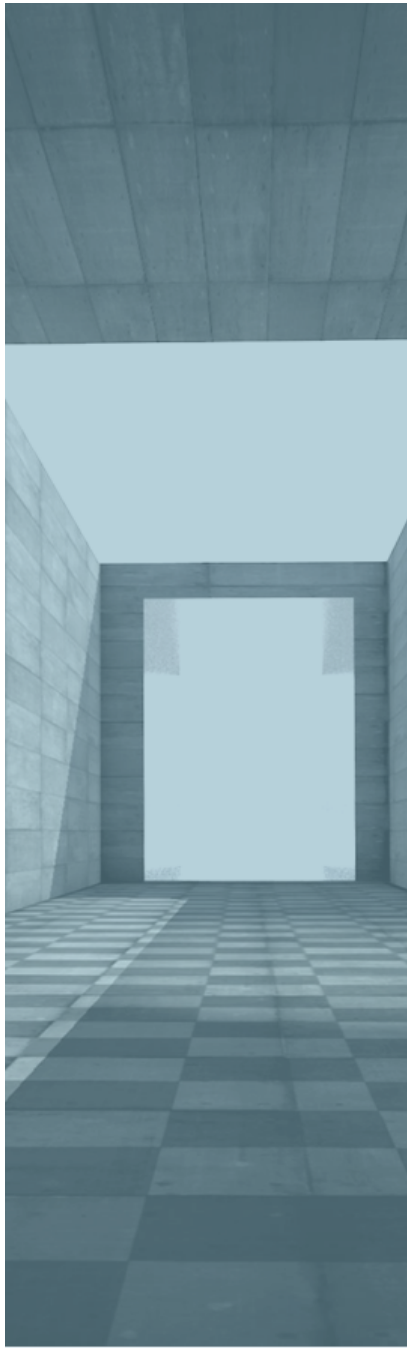


Conclusion (cont)

Key Findings, Insights, and Limitations: Concluding Remarks

Limitations:

- Due to the majority of the sample belonging to the manufacturing industry (99%), generalizing the results to other industries, such as services, may be challenging.
- The merger of non-financial data relied on the 10th Revision of the Korean Standard Industrial Classification (KSIC), which may overlook information in the open data that does not adhere to this standard.
- The collection of non-financial data faced limitations in capturing detailed information related to company management factors (e.g., management expertise, leadership) and technology.



Summary

Project Summary and Key Challenges

Summary:

- Utilized Business Trends and Outlook Survey Data to predict SME closures
- Implemented ML models using the R language, utilizing packages such as Random Forest, CatBoost, and BART
- Evaluated models using AUROC, F1 score, and accuracy metrics
- The project insights highlight the value of incorporating non-financial data in predicting SME closures

Challenge:

- Data: Finding appropriate and reliable open-sourced data, as well as facing limitations in merging them with the data obtained from NICE DNB
- Model performance: Initially, the dataset was not separated into train and test sets considering the difference in the ratio between closed and active companies, leading to a poor AUROC value