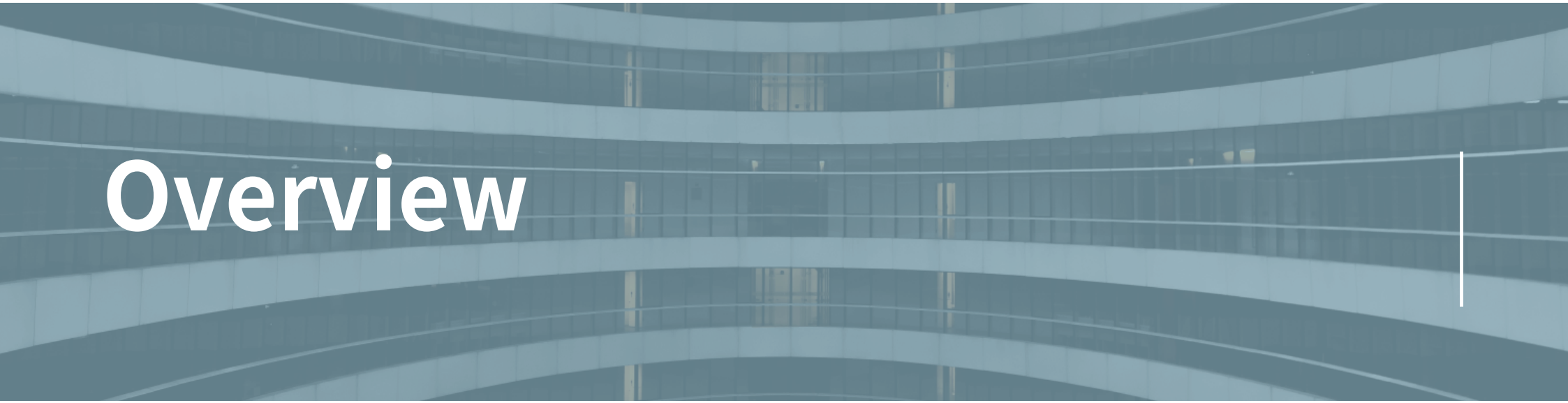
Sep 2022 - Nov 2022

Competition Project

Small and Medium-sized Enterprises Closure Prediction Project

Team members: Nayeon Kwon, Younghoon Yoo

Presentation created by: Nayeon Kwon



Overview

About the Competition

Organized by NICE DNB, Korea's largest commercial credit bureau.

Objective

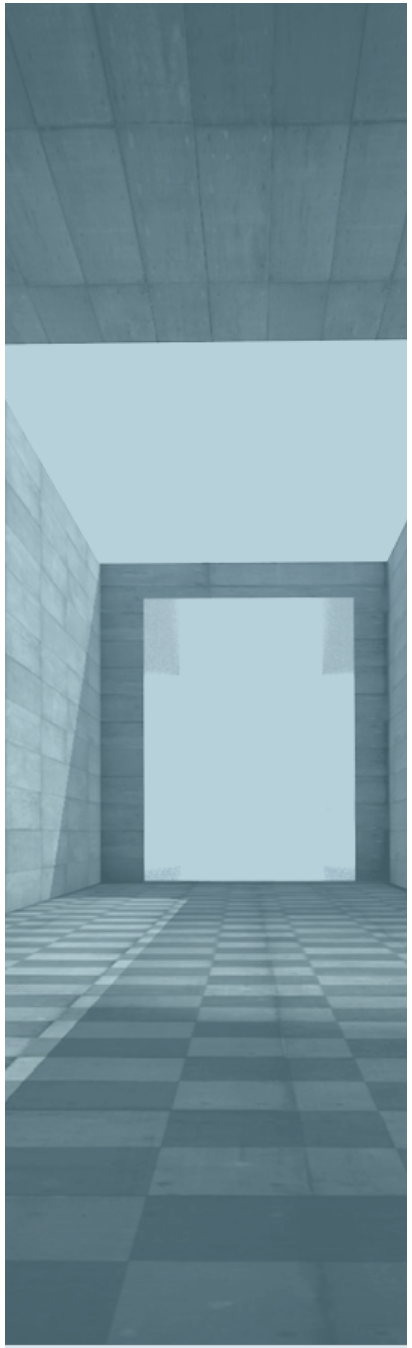
Develop a commercial credit model for SMEs by predicting closures using machine learning models.

Approach

Leverage non-financial data and implement machine learning models to classify and predict SME closures.

Contribution

My roles: Sourcing non-financial data, data preprocessing, supporting building ML models, documentation



Background

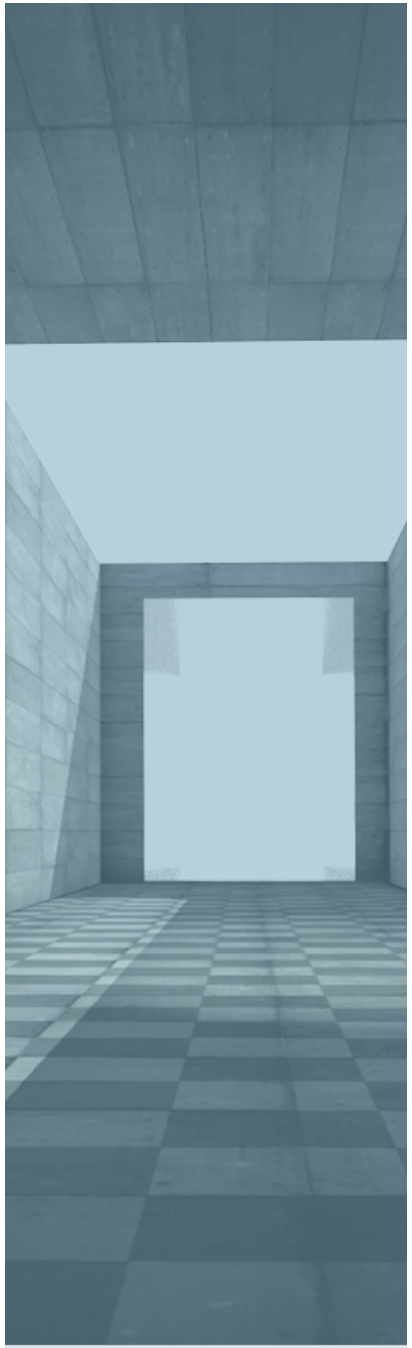
The Significance of Non-Financial Metrics in SME Credit Ratings and Stakeholder Utilization of Non-Financial Information

Importance of Non-Financial Factors:

- Financial factors provide limited and objective information, insufficient for accurate credit ratings and insolvency predictions.
- Non-financial factors are crucial for evaluating SMEs due to their financial vulnerability and sensitivity to changes.
- Sample Company Size: ~75% of companies have 58 employees or fewer.

Stakeholders' Interests:

- Banks: Use non-financial information for systematic credit risk management.
- SMEs: Secure loans through technological capabilities and non-financial information.



Data Description

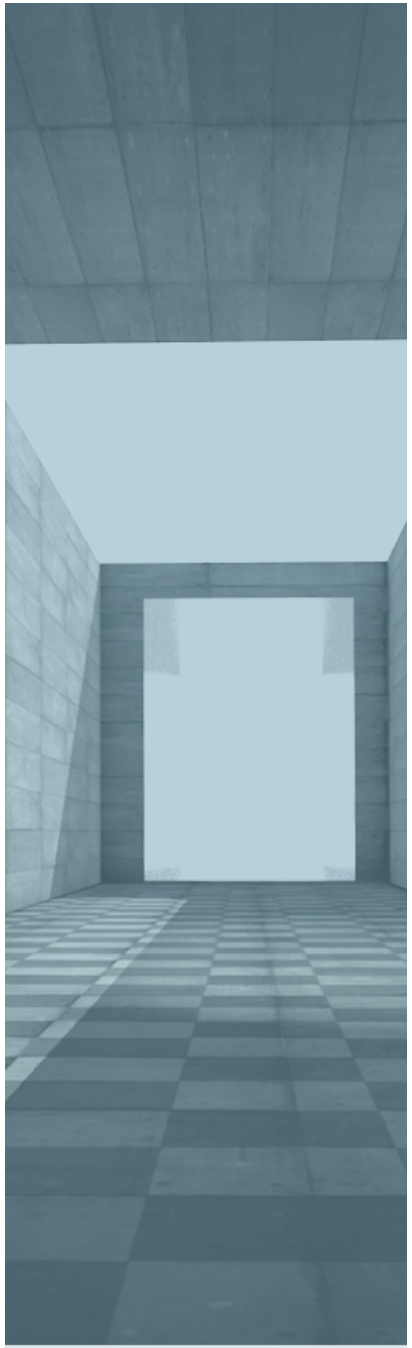
Data Sources for Model Prediction

Data Sources:

- Monthly Business Trends and Outlook Survey data (Jan 2018 - Jun 2022) from the Ministry of SMEs and Startups, South Korea.
- Focus on the SBHI (Small Business Health Index) for manufacturing industry companies.
- Additional open and financial data from NICE DNB.

SBHI Explanation:

- Calculated from five-point scale responses.
- Values > 100 indicate business improvement anticipation; values < 100 indicate the opposite.



Data Preprocessing

Integrating Diverse Data Sources for SMEs

Integration of Data:

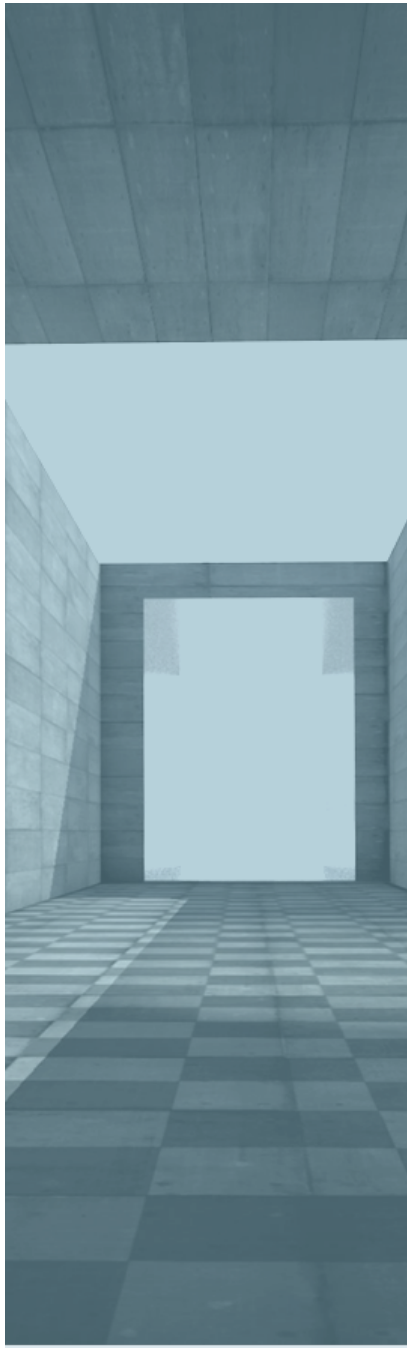
- Merged open, financial, and non-financial data for SMEs in the manufacturing industry.

Response Variable:

- Binary classification: 0 (Active SMEs) and 1 (Closed SMEs).

Total Observations:

- 32,392 observations with 28 variables after preprocessing.

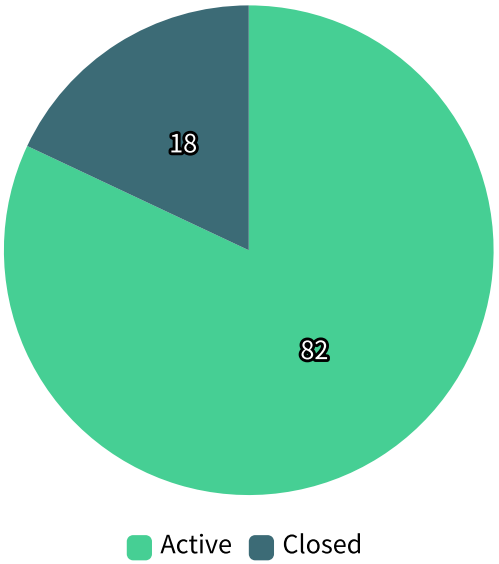


EDA

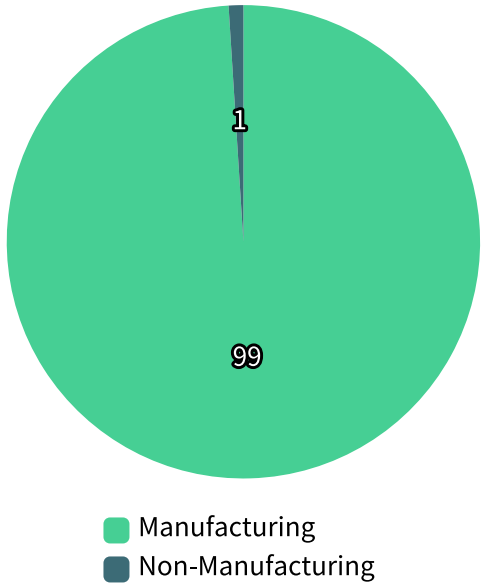
Exploratory Data Analysis

Summary Statistics:

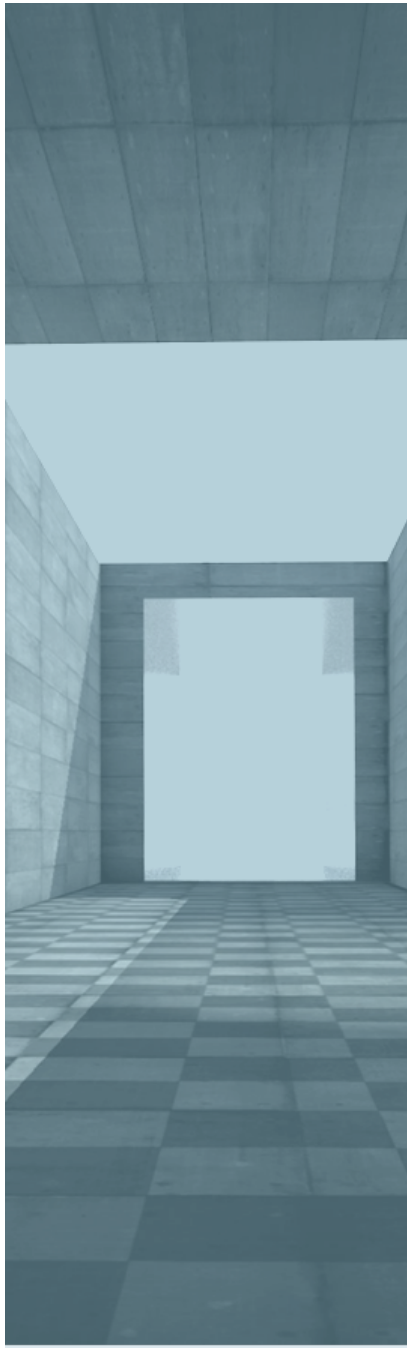
Closure status



Industry



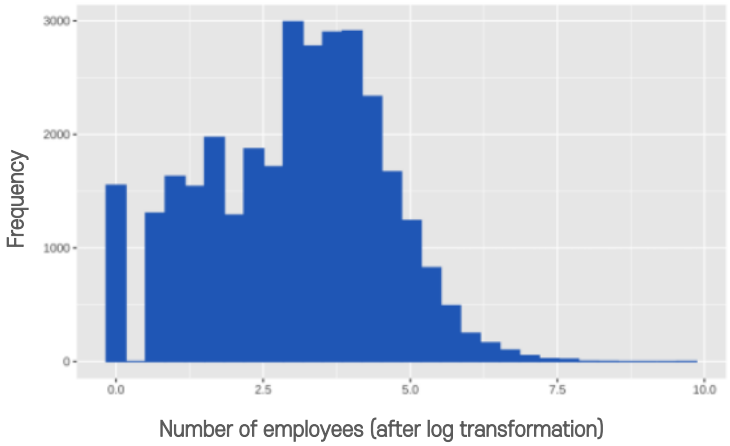
For detailed summary statistics of the variables, please click here: Summary Statistics



EDA

EDA (cont)

Exploratory Data Analysis

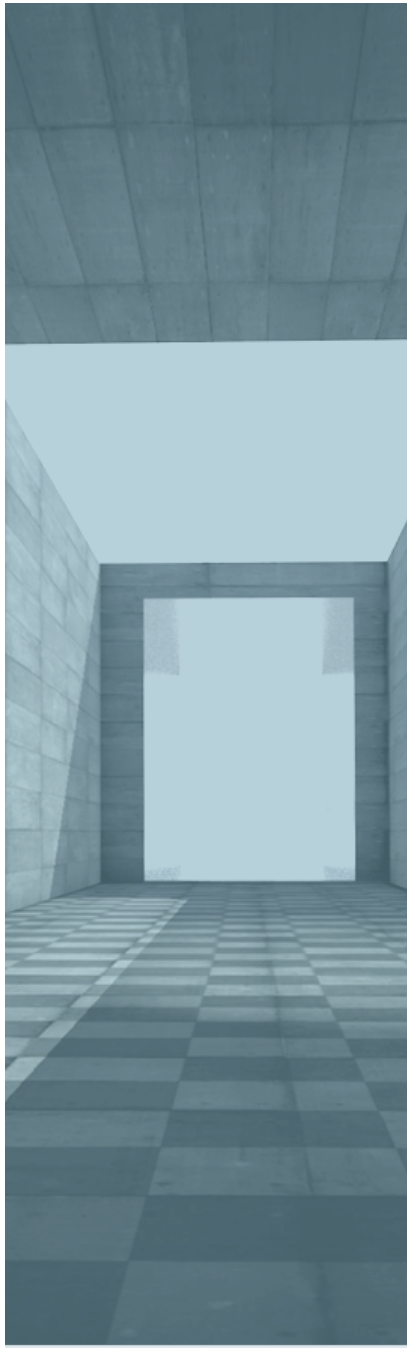


Histogram of Employees:

- Log-transformed, skewed distribution with a long right tail.

Word Cloud Analysis:

- Key words: manufacturing, real estate, sales, leasing, wholesale, and retail.



Model Implementation

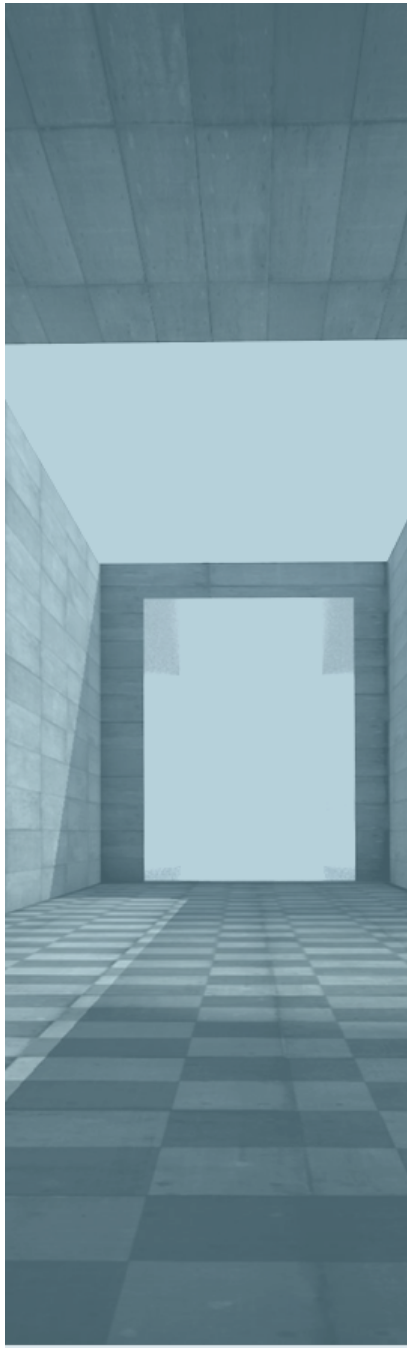
Model Implementation and Evaluation in R

Models Used (in R):

- Random Forest
- CatBoost
- BART

Evaluation Metrics:

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

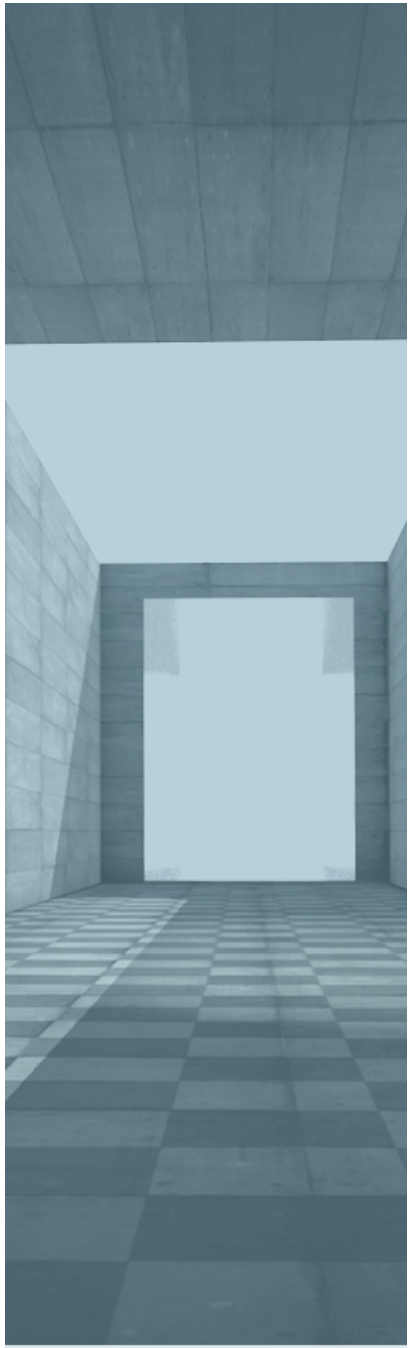
Model Evaluation Results

Model Performance (cont)

Evaluation of Predictive Model Performance

Summary:

- AUROC: Values ranged from 0.5 to 0.6, highest at 0.556.
- F1 Score: CatBoost highest at 0.992, models performed well on this metric.
- Accuracy: CatBoost highest at 0.987, all models > 0.9, indicating high correctness in predictions.



Key Findings

Variable Importance in Models

Influence of Financial Data:

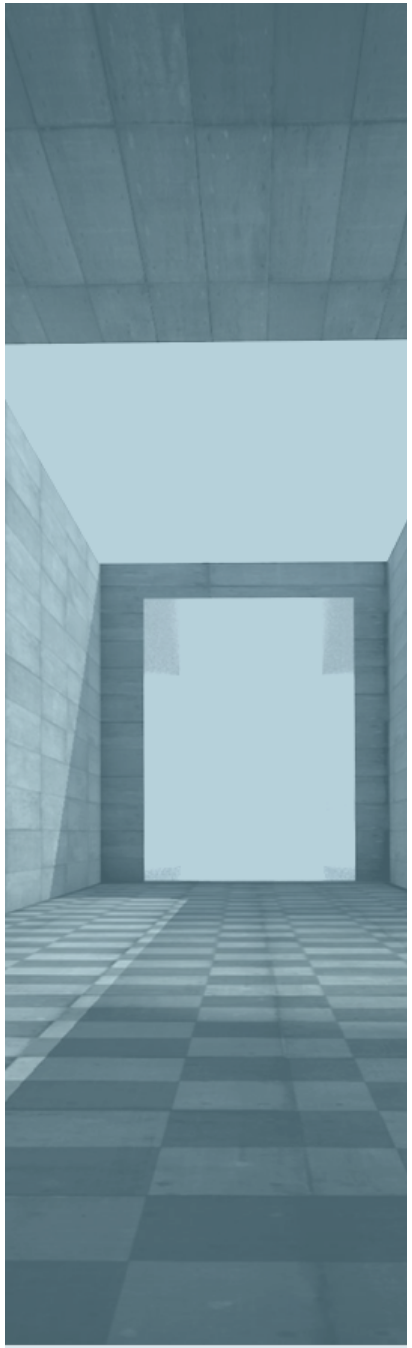
- More influential in predicting SME closures than open data.

Incorporation of Non-Financial Data:

- SBHI improved F1 score and accuracy.

Variable Importance:

- Total assets significant in non-BART models.
- Employment level performance SBHI significant in BART model.
- At least eight SBHI variables significant across models.



Limitations

Generalization Issues

Generalization:

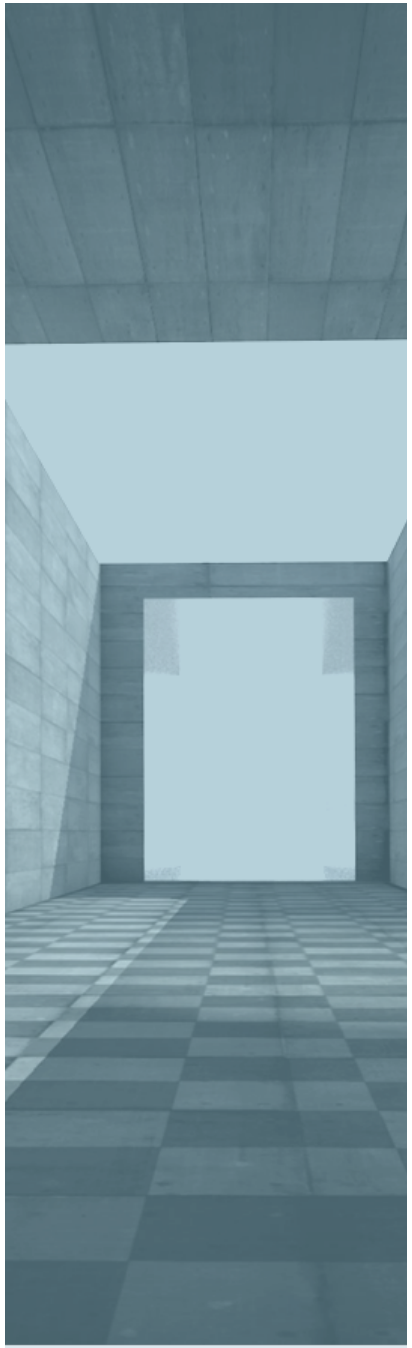
- Majority sample from manufacturing, limited applicability to other industries.

Data Merging Challenges:

- Reliance on 10th Revision of KSIC, potentially missing open data information.

Non-Financial Data Collection:

- Limited detailed company management factors (e.g., expertise, leadership) and technology.



Summary

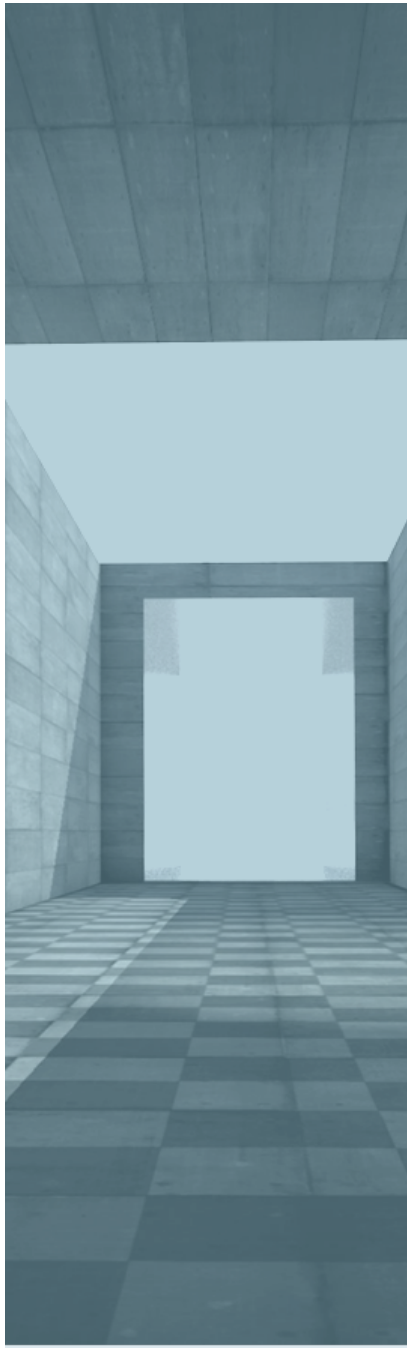
Summary and Challenges

Summary:

- Used Business Trends and Outlook Survey Data to predict SME closures.
- Implemented ML models using R (Random Forest, CatBoost, BART).
- Evaluated using AUROC, F1 score, and accuracy.
- Highlighted the value of non-financial data.

Challenges:

- Data: Finding and merging reliable data sources.
- Model Performance: Initial poor AUROC due to unbalanced dataset.



Conclusion

Key Insights

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.