NUS DATATHON CAT B (CHAMPIONS GROUP) COMPETITION 2025

REPORT FOR SUBMISSION ENTRY

Title: Machine Learning Models to Predict Company Structure

Team No. 24

Team member:

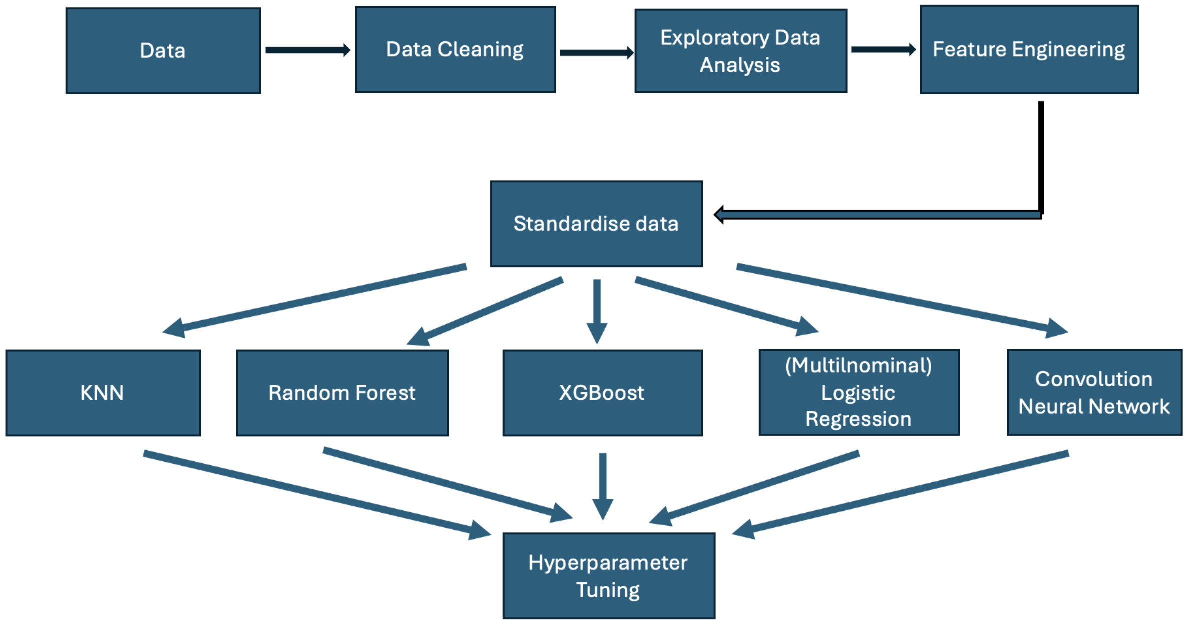
Name	**Email**	**Major/School**
Nguyen My Binh An	e1156700@u.nus.edu	Data Science & Analytics, Faculty of Science
Bui Phuong Linh	e1281054@u.nus.edu	Data Science & Analytics, Faculty of Science
Pham Ngoc Minh	e1249385@u.nus.edu	Data Science & Analytics, Faculty of Science

TABLE OF CONTENT

Introduction

Identifying corporate ownership structures is crucial for businesses, investors, and regulators to assess financial control, decision-making power, and market influence. Our project explores which factors play more important roles in determining whether a company is Domestic Ultimate or Global Ultimate; and thus applies machine learning models to predict corporate ownership structure classifications based on those factors.

Workflow



Dataset Overview

The dataset **"Champions_Group_2025.csv"**. consists of 29,182 records and 24 columns. The primary response variables are "Is Domestic Ultimate" and "Is Global Ultimate", which indicate whether a company is classified as a domestic ultimate, global ultimate, both or neither.

Methodology

To predict the two target variables "Is Domestic Ultimate" and "Is Global Ultimate", we will first explore the dataset to select features that have a (strong) association to the target variables. More importantly, since we are predicting two target variables, we will encode the target variable pair ("Is Domestic Ultimate", "Is Global Ultimate") and transform it to a single output variable "Structure".

Due to high numbers of columns in our dataset, we will perform feature selection to select variables that have strong association to the output variable.

- **For categorical variables**, their association with the output variable is assessed through:

  - **Contingency tables** to examine the frequency distribution of categorical variables across different company structures.
  - **Bayesian (conditional) probability** to evaluate how the probability of a company belonging to a specific structure changes given a categorical variable.
  - **Chi-square tests** to statistically validate the association, where a significant difference in conditional probabilities suggests a strong relationship with the output variable.
- **For numerical variables**, a detailed statistical analysis is performed:
  - The **mean and median** are compared across different company structures.

Boxplots and smoothed density histograms are used to visualize distributions and identify patterns that might contribute to predictive performance.

Following feature selection, we will proceed with feature engineering. This involves applying natural logarithm transformations to address skewness in numerical variables, regrouping categorical values, and creating new variables for more meaningful input variables, guided by business domain knowledge and data distribution characteristics. Once relevant features are selected, feature engineering is applied to enhance the predictive quality of the dataset:

- **Natural logarithm transformations** are used to address skewness in numerical variables.
- **Categorical variable regrouping** is performed to consolidate categories with similar characteristics.
- **New feature creation** is guided by business domain knowledge and data distribution patterns to provide more meaningful input representations.

After that, we have specifically chosen 4 machine learning models to predict the company structure: (Multi-class) Logistic Regression, K Nearest Neighbor, Random Forest, XGBoost, Convoluted Neural Network. After feature selection, we have chosen four machine learning models to predict company structure: **(Multi-class) Logistic Regression, K-Nearest Neighbors (KNN), Random Forest, XGBoost, and Convolutional Neural Network (CNN)**. These models were selected based on their suitability for handling categorical classification tasks and their varying levels of complexity and interpretability:

- **Logistic Regression**: Provides a simple and interpretable baseline for multi-class classification.
- **K-Nearest Neighbors (KNN)**: Captures local patterns in the data and is useful for non-linear relationships.
- **Random Forest**: An ensemble model that improves prediction accuracy by aggregating multiple decision trees and handling feature importance.
- **XGBoost**: A powerful gradient boosting model known for its high predictive accuracy and robustness in structured data.

- **Convolutional Neural Network (CNN)**: Although primarily used in image processing, it is included as an exploratory deep learning approach to assess its potential in capturing complex interactions in structured data.

Exploratory Data Analysis & Feature Selection

1. Target variables

In this study, we have two target variables: "Is Domestic Ultimate" and "Is Global Ultimate". The Domestic Ultimate classification indicates whether a company is the highest-level entity within its home country, while the Global Ultimate classification signifies whether a company is the highest-level entity globally, including subsidiaries in other countries.

	Is Global Ultimate		
Is Domestic Ultimate		0	1
	0	14589	0
	1	7086	7507

Table 1. Contingency table of two target variables “Is Global Ultimate” and “Is Domestic Ultimate”

Observation:

- This pattern results in only three possible combinations:
  - (0,0) for companies that are neither Domestic nor Global Ultimates
  - (1,0) for companies that are Domestic Ultimates but not Global Ultimates
  - (1,1) for companies that are both Domestic and Global Ultimates.
- Notably, the combination (0,1) does not exist, reinforcing that a company cannot be a Global Ultimate without also being a Domestic Ultimate

Therefore:

- If a company is global ultimate (Is Global Ultimate == 1), then it has 100% chance of being domestic ultimate (Is Domestic Ultimate == 1)
- However, if a company is domestic ultimate (Is Domestic Ultimate == 1), then it has around 51% of being global ultimate.

Given this structure, we transformed these two binary target variables into a single multi-class target variable, "Structure" as follows:

Original two target variables		New target variable	Remarks
Is Global Ultimate	Is Domestic	Structure	Encode to

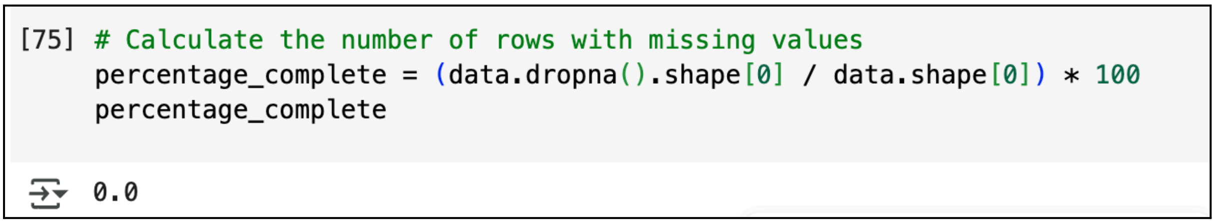
	Ultimate		
1	0	Global	2
1	1		
0	1	Domestic	1
0	0	None	0

Table 2. Create new output variable “Structure” based on the pairing of original target variables “Is Domestic Ultimate” and “Is Global Ultimate”

We further encode 0 represents companies that are neither Domestic nor Global Ultimates, 1 represents Domestic Ultimates that are not Global Ultimates, and 2 represents companies that are both Domestic and Global Ultimates. This transformation simplifies the classification problem while preserving the meaningful distinctions in corporate hierarchy. After that, we removed the “Is Domestic Ultimate" and "Is Global Ultimate" columns.

2. Missing values

First, we explore the percentage of rows that do not have missing values.



```
[75] # Calculate the number of rows with missing values
     percentage_complete = (data.dropna().shape[0] / data.shape[0]) * 100
     percentage_complete

     0.0
```

Figure 1. Our code shows that 0% of rows have non-missing values.

Figure 1 shows a snippet of our code suggesting that all rows in the dataset have missing values. Hence, it is impossible for us to remove all missing values in the dataset. Therefore, we need to investigate the missing values in each column to decide which columns to remove.

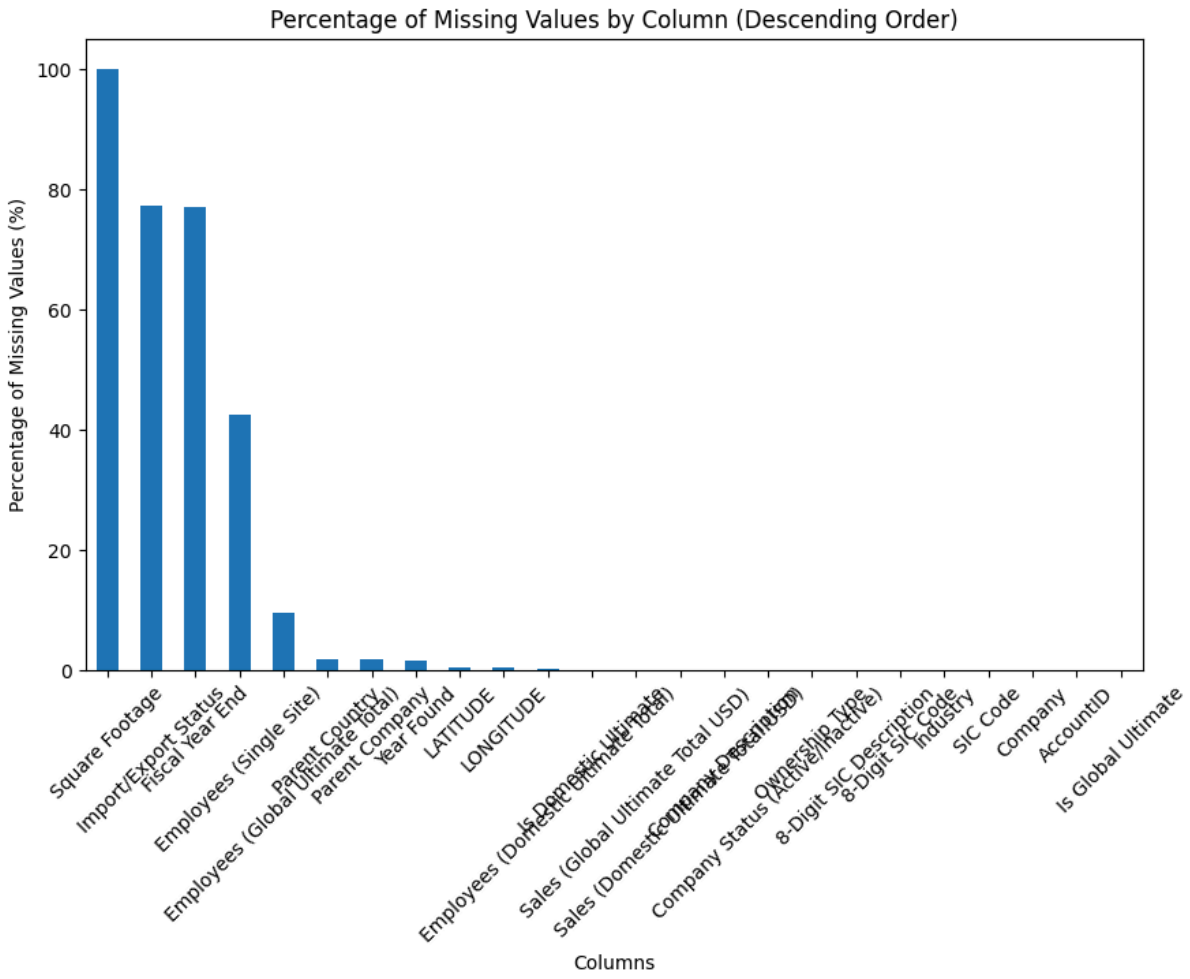


Figure 2. Bar chart to show the percentage of missing values by Column in descending order

The "Square Footage" feature is entirely (100%) missing from the dataset, while "Import/Export Status" and "Fiscal Year End" have approximately 77% missing values. Additionally, "Employees (Single Site)" has around 43% missing values. Due to the high percentage of missing data, we drop these four columns.

3. LATITUDE/LONGITUDE variable

A statistical analysis of the distributions for both "LATITUDE" and "LONGITUDE" indicates that the mean and median values are nearly identical across all three output classes.

Structure	mean_long	mean_lat	mean_long	mean_lat
Domestic	103.8	1.3	103.8	1.3
Global	103.8	1.3	103.8	1.3
None	103.8	1.3	103.8	1.3

Table 3. A statistical summary of mean and median longitude and latitude across 3 different company structures.

Moreover, overall, there appears to be little visual distinction in the geographical distribution (using Longitude and Latitude variable) of companies based on their ownership structure.

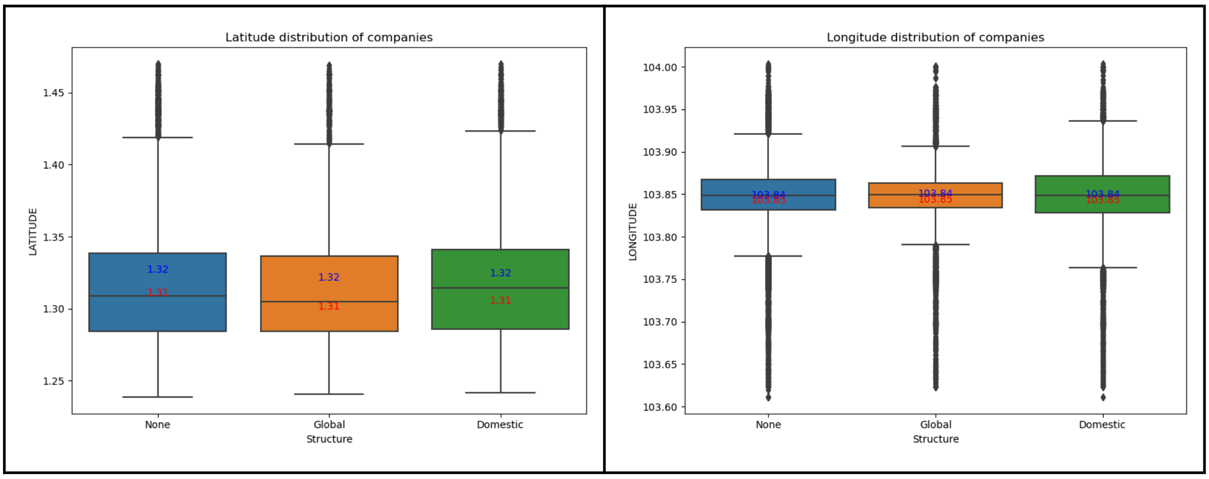


Figure 3. Boxplot of Latitude and Longitude distribution across 3 types of company structures

This suggests that geographic location may not be a significant factor in differentiating between companies with varying organizational structures. We will drop the LONGITUDE and LATITUDE columns.

4. Year Found variable

Structure	Median Founding Year
Domestic	2007
Global	2012
None	2007

Table 4. A statistical summary of median founding year across 3 different company structures

There are significant differences in the mean and median founding year between companies of different structures. The mean founding year for domestic ultimate companies is around 2006, while for non-domestic ultimate companies, it is 2003. The mean founding year for global ultimate companies is 2007, while for non-domestic ultimate companies, it is 2003.

This highlights a difference in founding years between domestic/global ultimate companies and those that are not.

Thus, it is likely that Year Found might have an association to the response variable Structure. We further investigate the distribution of Year Found variable in relation to different company structures.

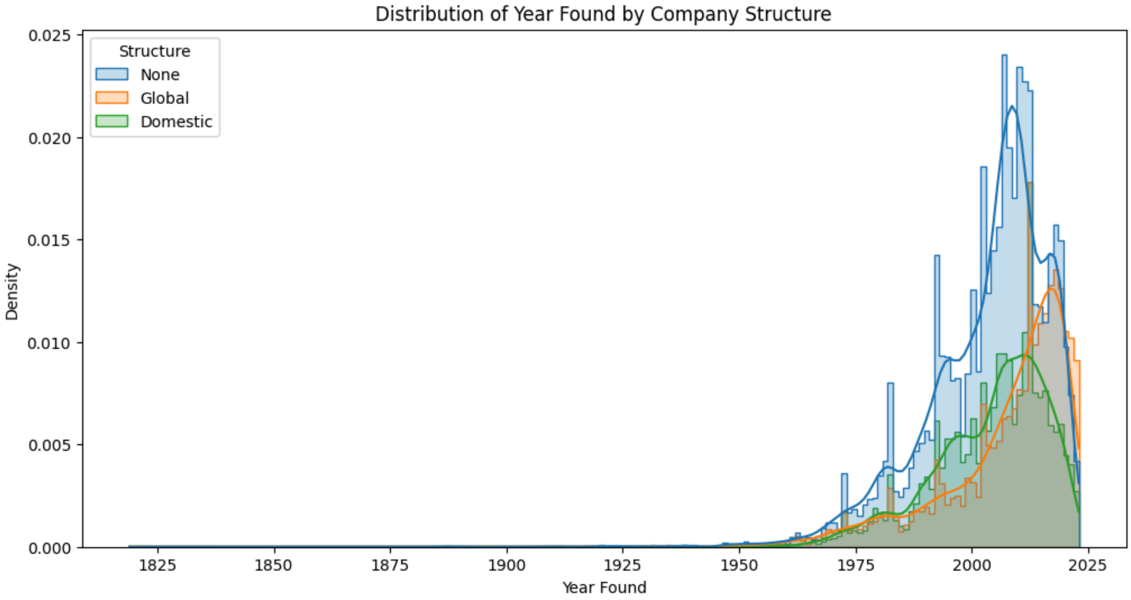
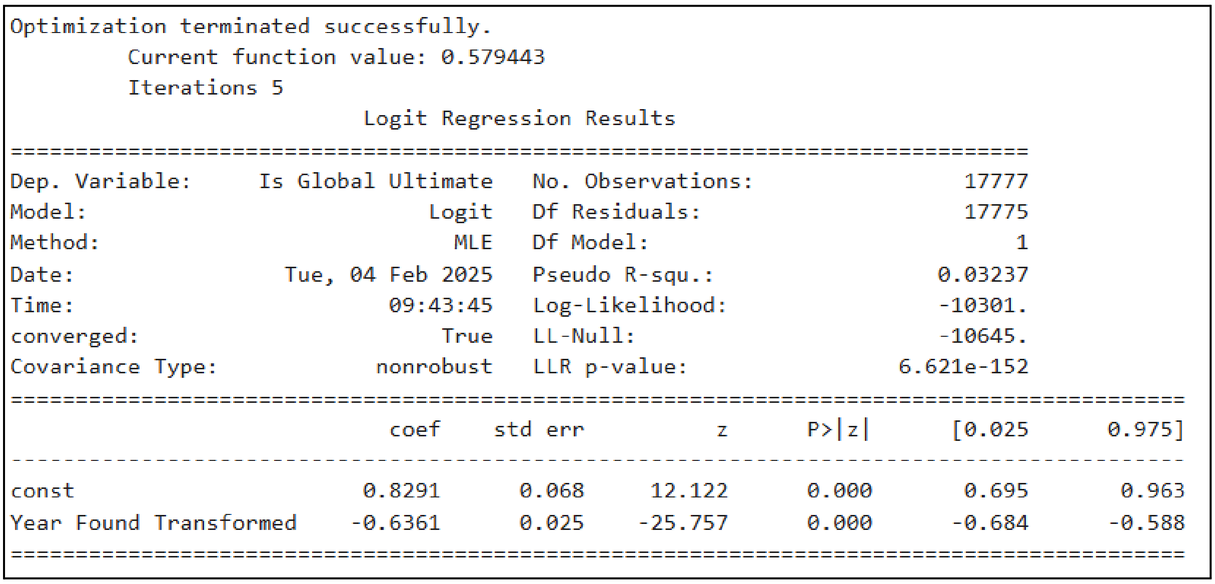


Figure 4. Smoothed distribution plot of Year Found by different Structure

The smoothed distribution shows approximately the same distribution of year found across all types of structures despite a difference in the mode of the distribution. We further carry out Logistic Regression to study the p-value of Year Found to the response variable.



```
Optimization terminated successfully.
         Current function value: 0.579443
         Iterations 5
                           Logit Regression Results
==============================================================================
Dep. Variable:     Is Global Ultimate   No. Observations:                17777
Model:                          Logit   Df Residuals:                    17775
Method:                           MLE   Df Model:                            1
Date:                Tue, 04 Feb 2025   Pseudo R-squ.:                 0.03237
Time:                        09:43:45   Log-Likelihood:                -10301.
converged:                       True   LL-Null:                       -10645.
Covariance Type:            nonrobust   LLR p-value:                6.621e-152
==========================================================================================
                             coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------
const                      0.8291      0.068     12.122      0.000       0.695       0.963
Year Found Transformed    -0.6361      0.025    -25.757      0.000      -0.684      -0.588
==========================================================================================
```

Figure 5. Summary statistics of Logistic Regression shows low p-value that suggests statistical significance of Year Found variable

The p-value for year founded is quite low (lower than the usual threshold of 0.05), indicating that the year founded variable is statistically significant for the model.

5. Employees variable

A statistical analysis indicates a significant difference in the distribution of the number of employees between domestic ultimate sites ("Employees (Domestic Ultimate Total)") and global ultimate sites ("Employees (Global Ultimate Total)").

Structure	mean_employee_global	mean_employee_domestic	median_employee_global	median_employee_domestic
Domestic	25246	90	850	22
Global	668	56	7	7
None	3704	208	15	20

Table 5. A statistical summary of mean and median employees across 3 different company structures

Overall, the employee counts at domestic and global ultimate sites exhibit distinct statistical patterns across different company structures.

There are significant differences in mean and median employee count at domestic and global ultimate sites across company structures.

We further explore the pattern of employee counts at global and domestic sites in each company structure via a scatter plot.

Furthermore, the density plots (Fig 6) indicate that the distribution of employee counts varies significantly across ownership structures. Although the distributions share a similar bimodal shape, the peaks of these distributions differ markedly, highlighting distinct patterns based on ownership structure.

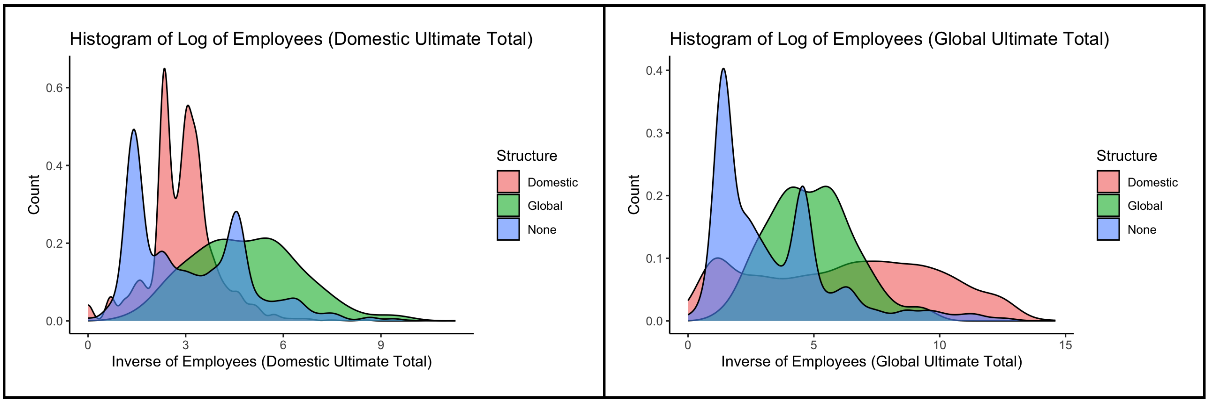


Figure 6. Histogram of Logarithm of Employee variables

Additionally, when a scatter plot of global employees count versus the logarithm of domestic employees is examined, a distinct trend emerges (Fig 7).

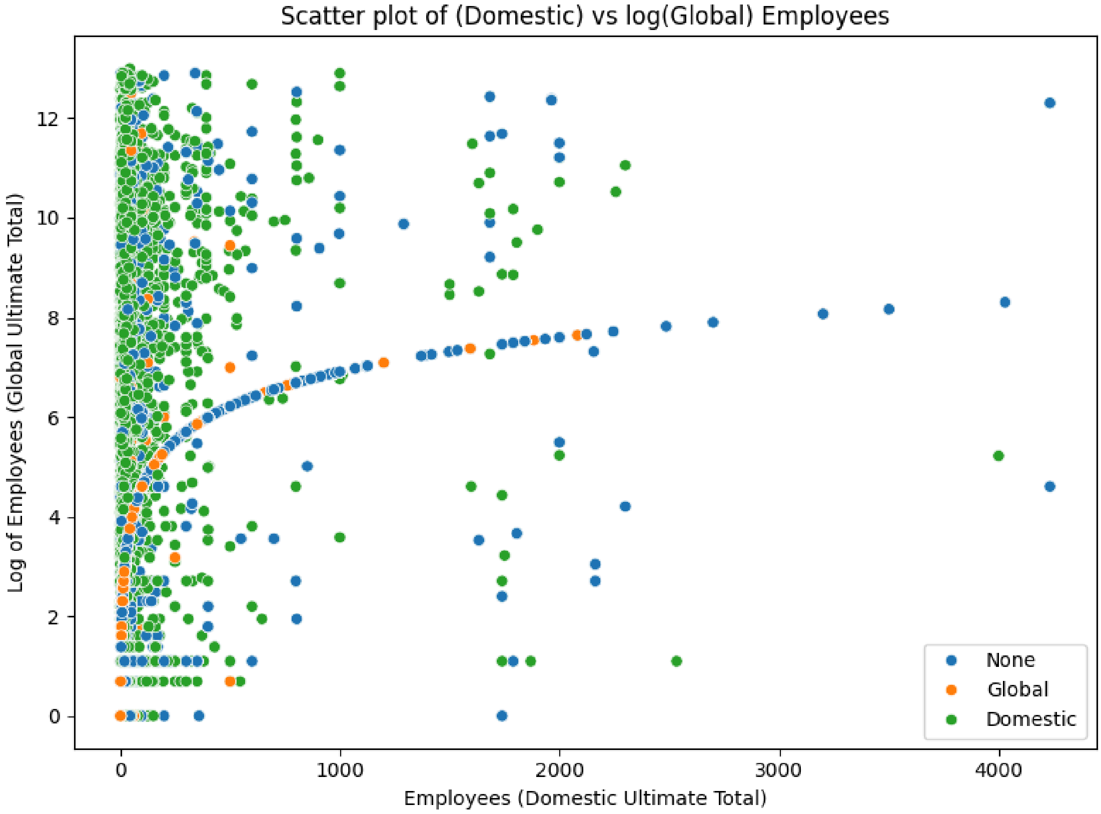


Figure 7. Scatter plot Log of Employee (Global) vs Employee (Domestic)

- The scatter plot reveals a logarithmic-like trend, with most data points along this curve representing companies with a 'None' or 'Global' structure.
- In contrast, companies with a 'Domestic' structure are more dispersed vertically, forming a cluster along the domestic employee count axis, primarily within the range of 0 to 1,000 employees.

Thus:

- Due to (i) the significant statistical distribution of mean and median employee count at domestic and global sites of different company structures and (ii) significant pattern in the employee count distribution in different company structures, the EMPLOYEE variable(s) have a strong association to the output variable

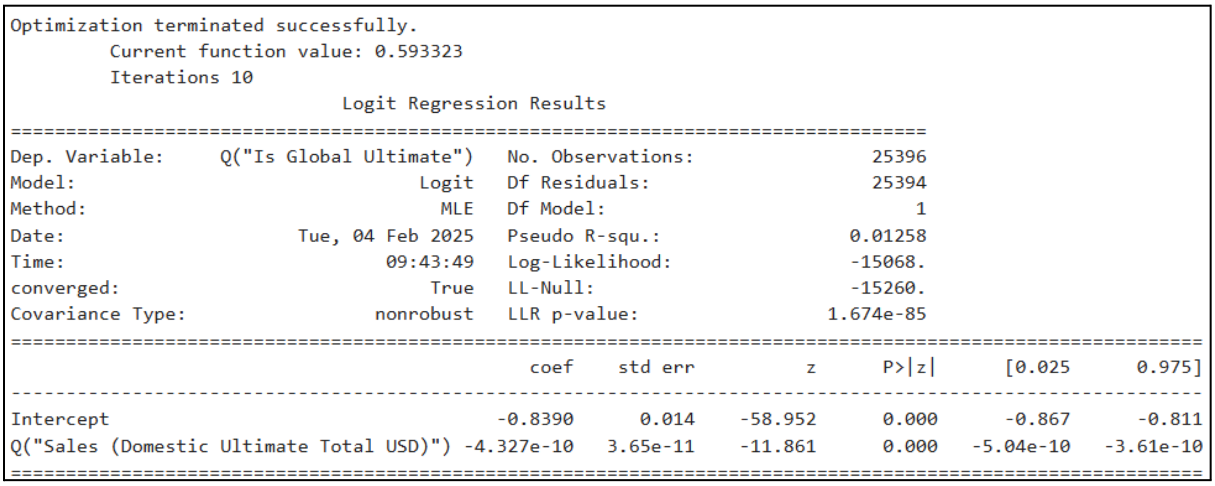
6. Sales variable

Table 7 shows the statistical summary of Sales variables.The mean and median sales at domestic and global sites vary significantly across companies with different structures.

Structure	mean_sales_global	mean_sales_domestic	median_sales_global	median_sales_domestic
Domestic	3.2e8	9.4e9	4514078	177322470.5
Global	7.1e7	3.6e8	1026308	1026308.0
None	1.3e9	3.1e9	4623465	3307860

Table 7. A statistical summary of Sales variables

Moreover, via Logistic Regression, the p-value for sales is quite low of 0 (lower than usual threshold of 0.05), indicating that the sales variable is statistically significant for the model.



```
Optimization terminated successfully.
         Current function value: 0.593323
         Iterations 10
                              Logit Regression Results
===================================================================================
Dep. Variable:     Q("Is Global Ultimate")   No. Observations:                25396
Model:                               Logit   Df Residuals:                    25394
Method:                                MLE   Df Model:                            1
Date:                     Tue, 04 Feb 2025   Pseudo R-squ.:                 0.01258
Time:                             09:43:49   Log-Likelihood:                -15068.
converged:                            True   LL-Null:                       -15260.
Covariance Type:                 nonrobust   LLR p-value:                 1.674e-85
============================================================================================================
                                               coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------------------------------------
Intercept                                   -0.8390      0.014    -58.952      0.000      -0.867      -0.811
Q("Sales (Domestic Ultimate Total USD)") -4.327e-10   3.65e-11    -11.861      0.000   -5.04e-10   -3.61e-10
============================================================================================================
```

Figure 8. A statistical summary from Logistic Regression suggests sales variables are statistically significant

7. SIC Code variable

At first glance, `SIC Code` and `8-Digit SIC Code` appear to represent the same information. Therefore, we aim to evaluate whether these two variables differ and assess their potential association with the response variable.

In a business context, while the 8-Digit SIC Code is not an official code used by the U.S. government, it serves as an extension developed by private agencies and companies. The U.S. Government allows additional subdivisions within specific four-digit industries, as stated in the SIC Code Manual, enabling more detailed industry classification. These extended SIC Codes (6, 7, and 8-digit versions) are continually updated by private data firms to refine industry categorization for marketing and identification purposes. As a result, the 8-Digit SIC Code can be a valuable indicator of whether a company operates within a branched or specialized industry.

Hence, we will explore how the `8-Digit SIC Code` is extended from the `SIC Code` by exploring the extended last 4 digits.

Since the `8-Digit SIC Code` is extended from the `SIC Code`, we hypothesise that the first 4 digits of the `8-Digit SIC Code` are similar from that of `SIC Code`. If any `8-Digit SIC Code` has the first 4 digits matching that of `SIC Code`, we will mark it as `Matched` and `Unmatched` otherwise. We will verify the hypothesis using a pie chart of "Matched" vs "Unmatched" proportion.

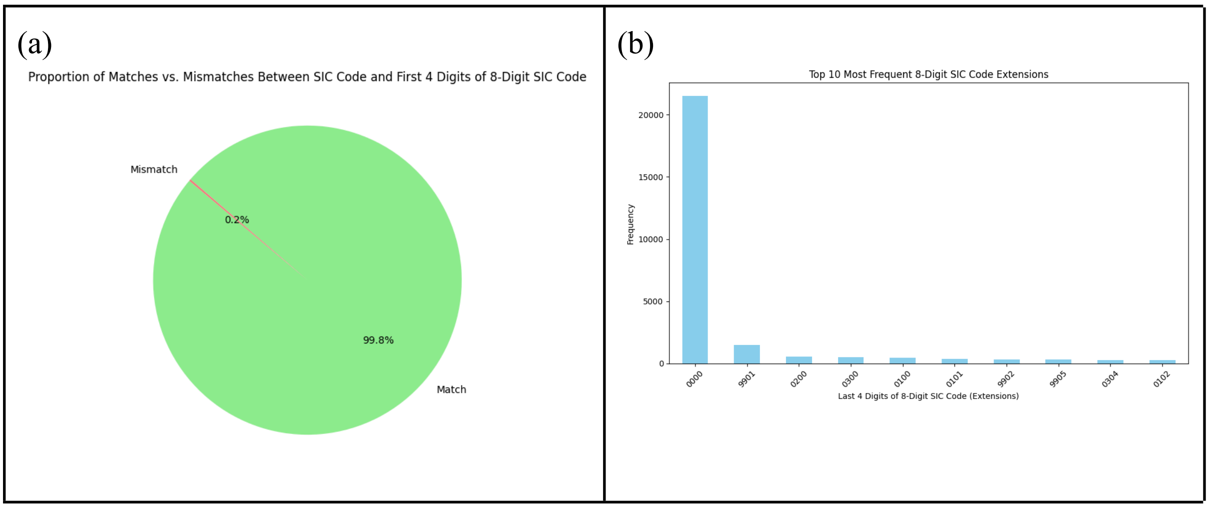


Figure 9. Visualisations to prove our hypothesis that 8-Digit SIC Code is retrieved by extending the SIC Code with ‘0000’. (a) Proportion of matching first 4-digits of 8-Digit SIC Code to SIC Code. (b) Top 10 Most Frequent last 4-digit extensions of 8-Digit SIC Code shows ‘0000’ as the most popular extensions

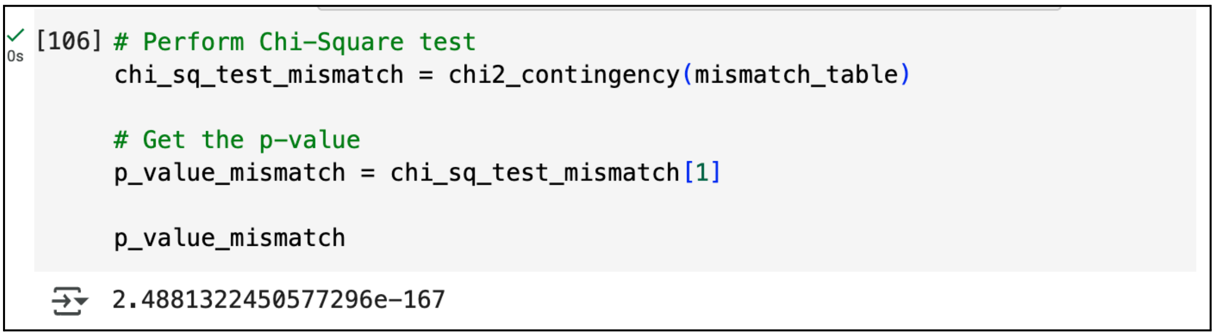
Since 99.8% of the dataset shows that the first four digits of the 8-Digit SIC Code match the SIC Code (Fig 9a), we can confirm that the 8-Digit SIC Code is an extension of the SIC Code with an additional four-digit suffix. Next, we analyze the last four digits of the 8-Digit SIC

Code to understand their patterns and significance (Fig 9b). Since majority of the 4-digit extensions are the code '0000', we can confirm that the 8-Digit SIC Code is generated by appending '0000' to the SIC Code

Thus, we create another variable Mismatch from the SIC Code and 8-Digit SIC Code:

- Mismatch = 'Matched' if the 8-Digit SIC Code = SIC Code + '0000'
- Mismatch = 'Mismatch', otherwise

We will build a contingency table and perform a Chi-square test to check the p-value. This aims to explain whether variable Mismatch has association to the response variable. Since the p-value is much lower than the threshold (Fig 10), it suggests that the Mismatch variable has a statistical significance in predicting the output response. Since we create another variable Mismatch from SIC Code and 8-Digit SIC Code, we will drop these columns.



```
[106] # Perform Chi-Square test
      chi_sq_test_mismatch = chi2_contingency(mismatch_table)

      # Get the p-value
      p_value_mismatch = chi_sq_test_mismatch[1]

      p_value_mismatch

      2.4881322450577296e-167
```

Figure 10. A snippet of our code to test the p-value of new variable “Mismatch” to the response variable

8. Company Status variable

We build a contingency table to explore the count of each category of Company Status in each type of company structure.

Company Status	Structure		
	Domestic	Global	None
Active	7086	7507	14589

Table 8. A contingency table shows that the only category in “Company Status” variable is “Active”

In the “Company Status” variable, all are 'Active'. This suggests that there is no difference in the Company Status regardless of the company structure. Hence, Company Status has no association to the output variable. We will drop the column Company Status.

9. Ownership Type variable

Below is the count summary of each category of Ownership Type for different Ownership structures.

Ownership Type	Structure		
	Domestic	Global	None
Non-corporates	0	6	3
Nonprofit	1	2	1
Partnership	4	31	10
Private	6975	7050	14225
Public	106	417	302
Public Sector	0	1	48

Table 9. A contingency table of "Ownership type" variable in response to output variable

There are 6 categories in the Ownership variable. However, Public and Public Sector might refer to the same thing. Hence, we will regroup those two categories into one common category "Public".

Subsequently, we will compute the Bayesian statistics table to determine the conditional probability of a company belonging to a specific type of company structure based on its ownership type.

Ownership Type	**Structure**		
	Domestic	**Global**	**None**
Non-corporates	0.0	66.7	77.3
Nonprofit	25.0	50.0	25.0
Partnership	8.9	68.9	22.2
Private	24.7	25.0	50.3
Public	12.1	47.8	40.4

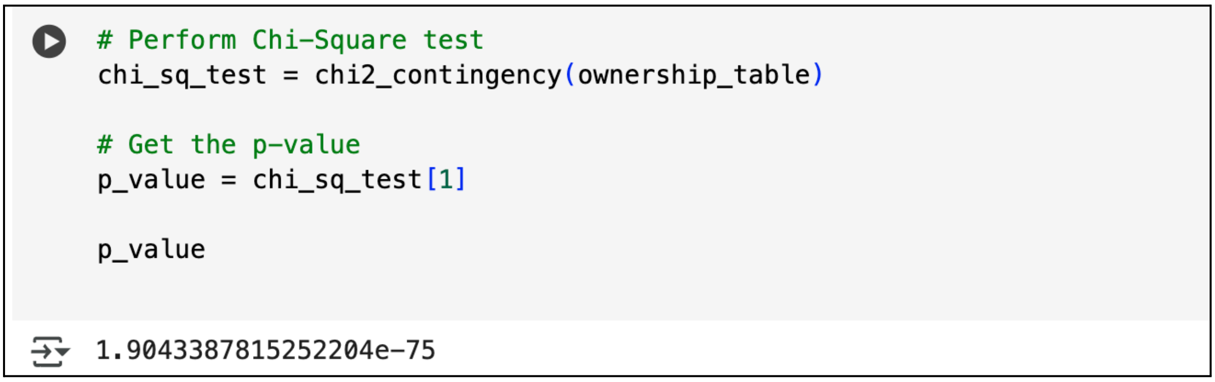
Table 10. The Conditional Probability table of a company belongs to a specific type of company structure given a type of ownership.

By Baysian statistics:

- Given that a company is of "Non-Corporates" ownership type, its structure has 0% chance of being "Domestic", 66.7% of being "Global", and 33.3% of being "None"
- Given that company is of "Nonprofit" type, its structure has 25% chance of being "Domestic", 50% chance of being "Global", and 25% chance of being "None"
- Given that a company is of "Partnership" type, its structure has 8.9% being "Domestic", 68.9% of being "Global", and 22.2% of being "None"
- Given that a company is of "Private" type, its structure has 24.7% of being "Domestic", 25.0% of being "Global", and 50.3% of being "None"
- Given that a company is of "Public" type, its structure has 12.1% of being "Domestic", 47.8% of being "Global", and 40.1% of being "None"

Hence, the conditional probability shows that an ownership type of a company has an association to the ownership structure of that company.

We further test the strength of the association via Chi-square test.



```
# Perform Chi-Square test
chi_sq_test = chi2_contingency(ownership_table)

# Get the p-value
p_value = chi_sq_test[1]

p_value
```

```
1.9043387815252204e-75
```

Figure 11. A snippet of our code in performing Chi-square test suggests that with low p-value, the variable "Ownership Type" is statistically significant.

The p-value is much lower than the usual threshold of 0.05, thus Ownership Type has a strong association and is statistically significant in predicting the output variable.

Feature Engineering

We have identified 5 input variables that show strong association to the response variable:

- SIC Code & 8-Digit SIC Code
- Ownership Type
- Sales
- Employees
- Year Found

For SIC Code & 8-Digit SIC Code, we will transform it to a binary categorical variable "Mismatch" as follows:

- If 8-Digit SIC Code matches the addition of SIC Code and '000', the new variable 'Code' will be 'Match'

- Otherwise, the variable ‘Mismatch’ will be ‘Mismatch’

For Ownership Type, we will re-group the categories to the new variable ‘Ownership’ as follows:

- Ownership = ‘Public’ if ‘Ownership Type’ = ‘Public’ or ‘Ownership Type’ = ‘Public Sector’
- Ownership = ‘Private’ if ‘Ownership Type’ = ‘Private’
- Ownership = ‘Others’ if otherwise

For Year Found, we will transform it to variable ‘Age’ where Age = 2025 - Year Found

For Sales & Employees, we will perform log transformation since the data is highly skewed.

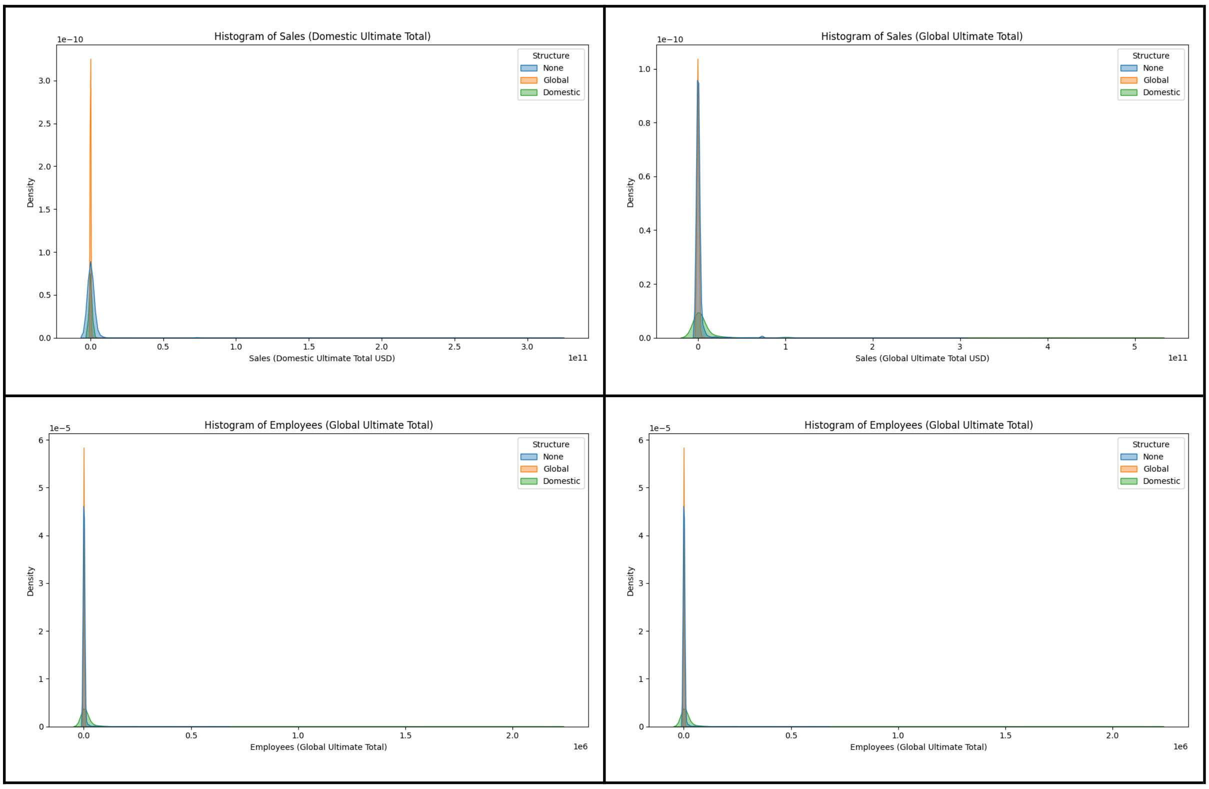


Figure 12. Histogram of Sales & Employees before Log Transformations shows high skewness

Machine Learning Models & Evaluation Metrics

Various Machine Learning models—KNN, Random Forest, Logistic Regression, and XGBoost—were applied to classify companies into either Domestic Ultimate, Global Ultimate, or neither, as each of these algorithms can naturally handle multi-class classification. In addition, a Convolution Neural Network (CNN), representing a Deep Learning approach, was implemented to exploit more complex patterns within the dataset. The performance of all models was assessed using standard classification metrics such as **accuracy, precision, recall, F1-score**, and **confusion matrices**, which together provide a comprehensive view of each model's strengths and weaknesses in predicting corporate ownership structures.

Results

Logistic Regression

Below is the evaluation report after using Logistic Regression to predict.

Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
0.61	0.62	0.62	0.58	0.74
Structure				
Class	Precision	Recall	F1-Score	
0	0.79	0.54	0.64	
1	0.57	0.22	0.32	
2	0.60	0.87	0.71	

Table 11. Evaluation metrics of Logistic Regression

Overall, the Logistic Regression - as a base model - performs relatively well. The Logistic Regression model achieved an accuracy of 0.61, with a precision and recall of 0.62, an F1-score of 0.58, and a ROC-AUC score of 0.74, indicating moderate overall performance. Class-wise, the model performed best on Class 2 with a high recall of 0.87 and an F1-score of 0.71, while struggling with Class 1, which had a low recall of 0.22 and an F1-score of 0.32. Class 0 showed balanced performance with a precision of 0.79 and an F1-score of 0.64. These results suggest the model effectively distinguishes some classes but requires improvement, particularly in identifying Class 1 instances.

K Nearest Neighbour

Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
0.70	0.70	0.70	0.70	0.85
Structure				
Class	Precision	Recall	F1-Score	

0	0.82	0.77	0.80	
1	0.59	0.58	0.58	
2	0.72	0.74	0.73	

Table 12. Evaluation metrics of KNN

The machine learning model achieved an overall accuracy of 0.70, with precision, recall, and F1-score all at 0.70, indicating balanced performance across these metrics. The ROC-AUC score of 0.85 suggests the model has strong discriminatory power in distinguishing between classes. Class-wise, the model performed best on Class 0, with a high precision of 0.82, recall of 0.77, and an F1-score of 0.80, demonstrating its reliability in predicting this class. Class 2 also showed strong results with an F1-score of 0.73, while Class 1 lagged behind, achieving the lowest precision (0.59), recall (0.58), and F1-score (0.58).

While the model performs well overall, particularly in predicting Classes 0 and 2, its relatively lower performance on Class 1 indicates room for improvement. This may be due to class imbalance, overlapping features, or insufficient representation of Class 1 in the training data. To address this, techniques such as resampling (over-sampling or under-sampling), feature engineering, or trying more complex models like ensemble methods could help. Additionally, analyzing the misclassified instances for Class 1 might reveal specific patterns the model struggles with, guiding targeted improvements.

Random Forest

Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
0.74	0.73	0.73	0.73	0.89
		Structure		
Class	Precision	Recall	F1-Score	
0	0.87	0.90	0.89	
1	0.63	0.56	0.59	
2	0.74	0.77	0.75	

Table 13. Evaluation metrics of Random Forest

The Random Forest model demonstrated strong performance with an overall accuracy of 0.74, and balanced precision, recall, and F1-score all at 0.73, indicating consistent predictive capability. The ROC-AUC score of 0.89 reflects the model's excellent ability to distinguish between classes. Class-wise, the model performed exceptionally well on Class 0, achieving a precision of 0.87, recall of 0.90, and an F1-score of 0.89, highlighting its reliability in identifying this class correctly. Class 2 also showed good performance with an F1-score of 0.75, supported by strong precision (0.74) and recall (0.77). However, the model struggled with Class 1, where precision (0.63), recall (0.56), and F1-score (0.59) were significantly lower compared to the other classes.

The Random Forest model showcases robust classification ability, especially for Class 0 and Class 2, but its weaker performance on Class 1 indicates challenges in correctly identifying instances from this class. This could be due to class imbalance, overlapping feature distributions, or insufficient informative features for Class 1. To address this, strategies such as rebalancing the dataset, incorporating feature selection or engineering, or fine-tuning model hyperparameters could improve performance. Additionally, analyzing misclassified samples may provide insights into patterns the model struggles with, helping to refine the approach for better generalization.

XGBoost

Below is the evaluation report after using XGBoost to predict.

Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
0.73	0.73	0.73	0.73	0.89
	Structure			
Class	Precision	Recall	F1-Score	ROC-AUC
0	0.89	0.89	0.89	
1	0.61	0.60	0.61	
2	0.74	0.75	0.75	

Table 14. Evaluation metrics of XGBoost

The XGBoost model achieved an overall accuracy of 0.73, with balanced precision, recall, and F1-score all at 0.73, indicating consistent performance across key metrics. The model's ROC-AUC score of 0.89 reflects strong discriminatory power in distinguishing between classes. Class-wise, the model performed best on Class 0, achieving high precision (0.89), recall (0.89), and F1-score (0.89), demonstrating excellent reliability in predicting this class. Class 2 also showed strong results, with precision of 0.74, recall of 0.75, and an F1-score of 0.75, indicating balanced performance. However, the model struggled with Class 1, with lower precision (0.61), recall (0.60), and F1-score (0.61), reflecting challenges in accurately classifying instances from this class.

While the XGBoost model exhibits robust performance overall, especially for Class 0 and Class 2, its weaker results for Class 1 suggest potential issues with class imbalance, feature relevance, or overlapping class characteristics. To improve performance, techniques such as resampling methods (SMOTE or class weighting), feature engineering, or hyperparameter tuning could be employed. Additionally, analyzing misclassified instances may provide insights into feature patterns that contribute to errors, allowing for targeted improvements to enhance the model's performance on Class 1.

CNN

Accuracy	Precision	Recall	F1-Score	ROC-AUC Score
0.70	0.70	0.70	0.70	0.87
	Structure			
Class	Precision	Recall	F1-Score	
0	0.76	0.94	0.84	
1	0.61	0.51	0.56	
2	0.72	0.72	0.72	

Table 15. Evaluation metrics of CNN

The Convolutional Neural Network (CNN) model achieved an overall accuracy of 0.70, with balanced precision, recall, and F1-score all at 0.70, reflecting consistent performance across these metrics. The ROC-AUC score of 0.87 suggests strong discriminatory power in classifying different categories. Class-wise, the model performed exceptionally well on Class 0, achieving high precision (0.76), recall (0.94), and F1-score (0.84), indicating strong capability in correctly identifying this class with minimal false negatives. Class 2 also showed balanced performance with precision, recall, and F1-score all at 0.72. However, the model struggled with Class 1, where precision (0.61), recall (0.51), and F1-score (0.56) were significantly lower, indicating difficulty in accurately classifying this class.

While the CNN model performs well overall, particularly for Class 0, its weaker performance on Class 1 suggests challenges that could stem from class imbalance, insufficient feature representation, or overlapping class characteristics. To improve performance, techniques such as data augmentation, class weighting, or hyperparameter tuning could be explored. Additionally, analyzing misclassified samples may reveal patterns contributing to errors, providing insights for targeted model improvements. Exploring more complex CNN architectures or combining CNNs with other models might also enhance the model's generalization capabilities.

Comparison between Models & Discussion

Based on the model performance comparison, the Random Forest (RF) and XGBoost (XGB) models stand out as top performers across key evaluation metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. Both models achieve high scores consistently, making them reliable choices, especially in business contexts where balanced performance and robustness are crucial. RF slightly edges out in ROC-AUC, indicating superior capability in distinguishing between classes, which is critical in scenarios like fraud detection or risk assessment. If the business requires high precision (e.g., minimizing false positives in customer targeting), RF and XGB are preferable.

However, if recall is more important (e.g., identifying as many potential risks as possible in medical diagnoses), RF also performs strongly. While K-Nearest Neighbors (KNN) and CNN show decent results, they lag slightly behind, and Logistic Regression (LR) performs the weakest overall, making it less suitable for complex business applications. In summary, **RF** and **XGB** are the recommended models for business-critical tasks due to their balanced and high performance across all relevant metrics.

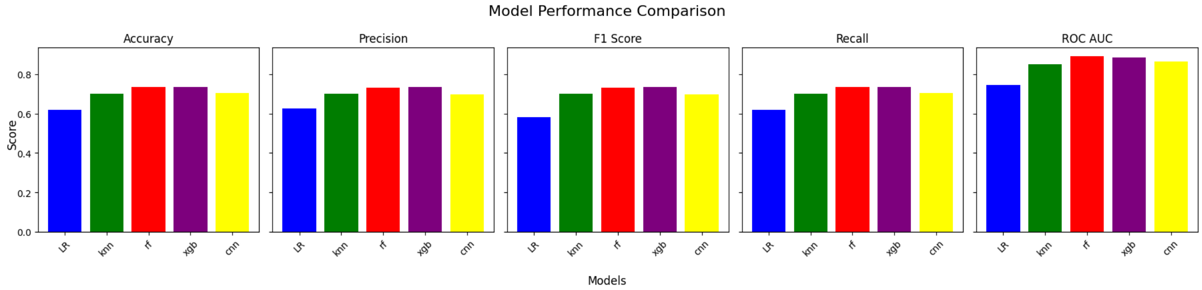
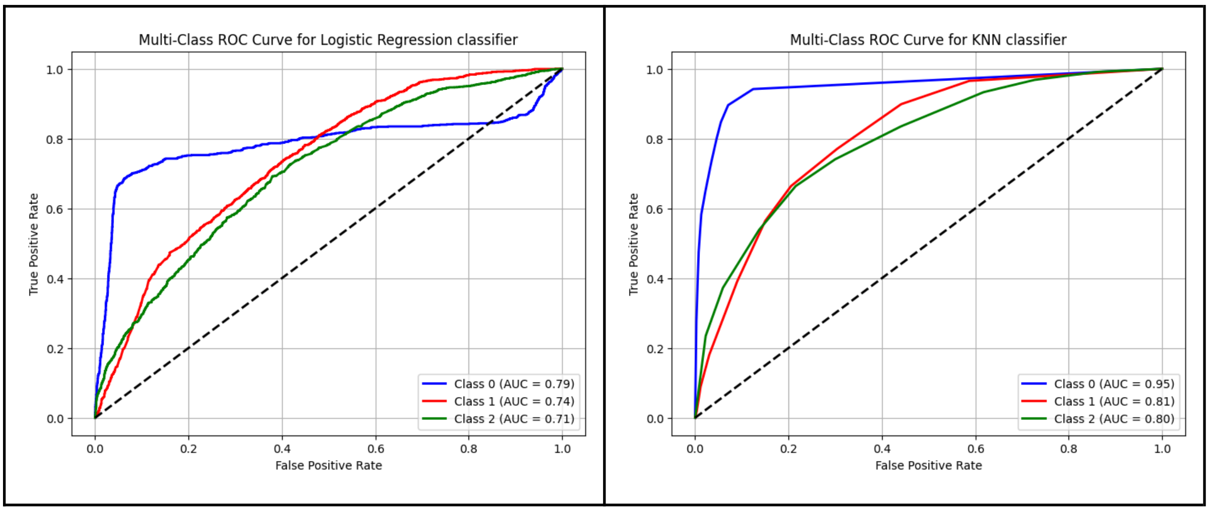


Figure 13. A comparison between evaluation metrics of different models



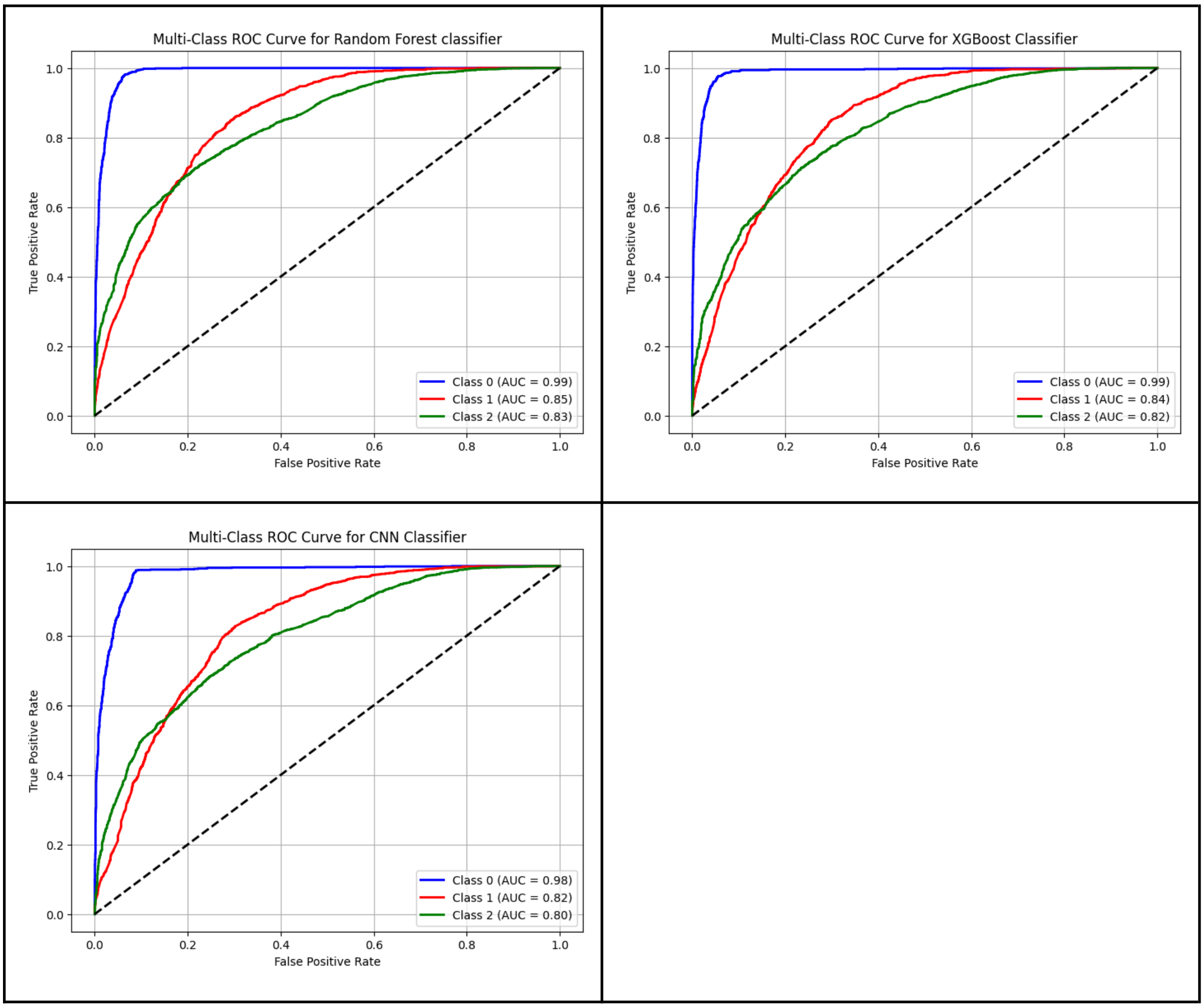


Figure 14. ROC curve and AUC value of 5 models

Insights on the Significance of XGBoost and Random Forest in Predicting Company Structure

The outstanding performance of **XGBoost (XGB)** and **Random Forest (RF)** highlights their significance in accurately predicting the **output structure of companies**—whether classified as **Global Ultimate**, **Domestic Ultimate**, or **None**. Both models excel at handling complex, non-linear relationships, which is critical when dealing with diverse business characteristics influenced by multiple variables such as **SIC Code**, **Ownership Type**, **Sales**, **Employees**, and **Year Found**.

Handling High-Dimensional and Categorical Data

The **SIC Code** and **8-Digit SIC Code** are categorical variables with potentially high cardinality. RF and XGB are particularly effective in dealing with such data without requiring extensive pre-processing. RF's ensemble of decision trees allows it to capture intricate patterns within these codes, identifying industries with specific trends in ownership structure. XGB, with its gradient boosting approach, iteratively improves its predictive performance by focusing on hard-to-classify cases, making it adept at distinguishing nuanced differences between similar industry classifications.

Modeling Business Complexity with Numerical and Ownership Features

For continuous variables like **Sales** and **Employees**, both models excel at capturing non-linear relationships and interactions. Companies with similar sales figures may have different ownership structures depending on employee count, growth trajectory, or industry-specific factors. RF effectively identifies these interactions through random feature selection at each split, while XGB refines these insights through boosting, ensuring that outliers or anomalies (like exceptionally large companies) do not skew the overall performance. Additionally, **Ownership Type**, a critical categorical feature, interacts with these financial metrics in complex ways—patterns that RF and XGB can uncover without manual feature engineering.

Temporal Dynamics and Company Maturity

The variable **Year Found** provides insights into a company's maturity and potential structural evolution. Companies with longer operational histories might exhibit different ownership dynamics compared to newer firms. RF captures these temporal patterns by creating diverse decision paths, while XGB identifies subtle trends, such as how younger tech companies may lean towards being domestically owned, whereas older conglomerates may fall under global structures.

Business Implications

In practical business applications, these predictive capabilities support **strategic decision-making**. For instance, in **credit risk assessment**, knowing the likely ownership structure helps in evaluating financial stability. In **market segmentation**, understanding whether a company is part of a global entity or domestically controlled informs tailored marketing strategies. Moreover, for **mergers and acquisitions**, insights from these models aid in identifying potential acquisition targets based on structural patterns.

In conclusion, **XGBoost and Random Forest not only provide high predictive accuracy but also offer valuable business insights** by effectively modeling complex relationships among diverse input variables. Their robustness, interpretability (especially RF with feature importance measures), and adaptability to both structured and unstructured data make them indispensable tools for predicting company structures in dynamic business environments.

Conclusion

In predicting the company structure (Global Ultimate, Domestic Ultimate, None), XGBoost (XGB) and Random Forest (RF) have proven to be the most effective models, excelling in accuracy, precision, recall, F1-score, and ROC-AUC. Their ability to handle complex, non-linear relationships among key business variables—SIC Code, Ownership Type, Sales, Employees, and Year Found—makes them particularly well-suited for this classification task. RF's ensemble learning approach ensures robust generalization, while XGB's gradient boosting mechanism enhances performance by refining difficult classifications.

From a business perspective, these models provide actionable insights for credit risk assessment, market segmentation, and M&A strategies, enabling informed decision-making based on a company's likely structural classification. Their superior handling of categorical variables, numerical data, and temporal trends ensures they capture nuanced patterns, making them highly valuable in financial and business intelligence applications.

Going forward, further improvements can be explored through feature engineering, hyperparameter tuning, and ensemble stacking to optimize predictive power. However, based on current results, XGB and RF stand out as the most reliable models for accurately predicting company structure, supporting data-driven strategies in a competitive business landscape.