NUS Datathon 2025 | Champions Group
Category B: Team 66
Bryan Kee Tze Ren | Soh Kai Le | Li Qi Ying | Nevin Allen Marian | Chan Wei Wen Kevin

1. Overview

Abstract

This report outlines our approach for predicting corporate ownership classifications using machine learning techniques. The objective is to classify firms into three distinct categories based on their ownership status: **Neither Domestic nor Global Ultimate**, **Domestic Ultimate but not Global Ultimate**, and **Both Domestic & Global Ultimate**. Our methodology involves comprehensive data preprocessing, exploratory data analysis, feature engineering, and model training to enhance predictive accuracy. Our chosen model, **XGBoost**, achieved a **mean accuracy of 91.45%** and a **mean weighted F1-score of 91.48%**, demonstrating good discriminatory ability at classifying corporate ownership structures.

Introduction

Corporate ownership structures play a crucial role in regulatory compliance, risk assessment, and business intelligence. Understanding whether a company is a Domestic Ultimate or a Global Ultimate provides valuable insights into its operational independence, financial backing, and decision-making power—factors that influence competitive strategy, investment decisions, and mergers & acquisitions. As part of NUS's annual Datathon 2025, we worked with a dataset provided by SingLife, containing financial and corporate attributes of various firms. Our goal was to develop a machine learning model capable of accurately predicting corporate ownership classifications. This required careful data preprocessing, feature engineering, and model selection to extract meaningful patterns. Through iterative experimentation and statistical analysis, we refined our approach to achieve high predictive accuracy, enabling more efficient decision-making for stakeholders.

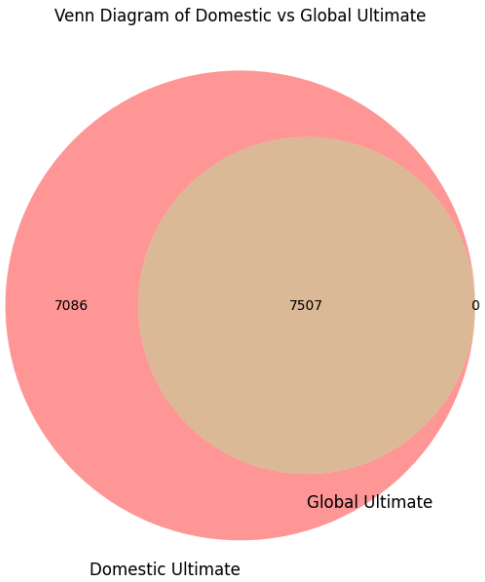
2. Exploratory Data Analysis

Redefining Target Variable

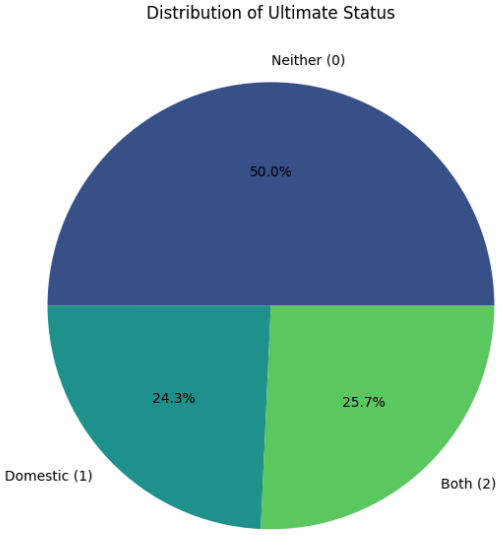
The initial dataset contained two binary response columns: **Is Domestic Ultimate** and **Is Global Ultimate**. A Venn diagram (Figure 1) revealed that all Domestic Ultimates were also Global Ultimates, prompting us to redefine the target variable as **Ultimate_Status**:

- **0** = Neither Domestic nor Global Ultimate
- **1** = Domestic Ultimate but not Global Ultimate
- **2** = Both Domestic & Global Ultimate

This transformation converted the original multi-label classification problem into a simpler multi-class classification task, effectively removing an irrelevant category and streamlining the prediction process. A pie chart (Figure 2) showed a balanced class distribution (2:1:1), reducing concerns of class imbalance. However, class weighting and resampling were considered to further optimize model performance.



(Figure 1: Venn Diagram of Overlapping Domestic & Global Ultimate Response Variables)

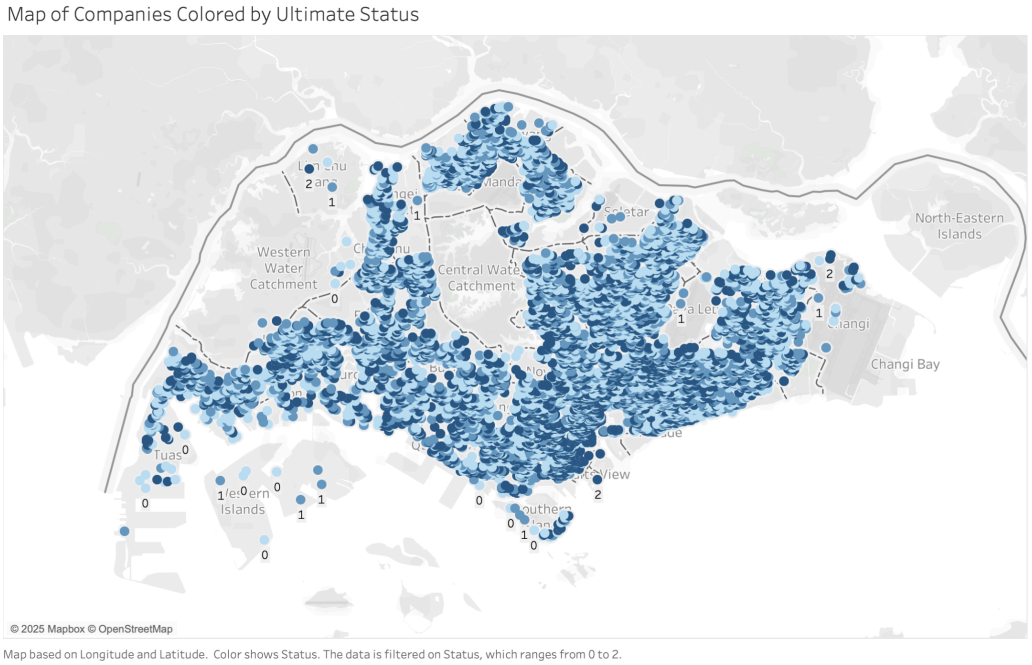


(Figure 2: Pie Chart of Ultimate Status Distribution)

Data Cleaning

Removing Irrelevant Columns: A preliminary analysis of dataset structure and missing value distribution led us to mark certain columns for removal. These columns were identified based on their lack of predictive value, redundancy, excessive missing data, or high cardinality. However, final removal was deferred until the end of data cleaning to ensure no crucial information was lost.

Columns Removed	Rationale for Removal
LATITUDE **LONGITUDE**	Unlikely to contribute to corporate classification since all coordinates correspond to Singapore. A map visualization (Figure 3) confirms this showing that coloring points by the response showed no patterns.
8-Digit SIC Code **8-Digit SIC Description** **Industry**	Contain similar information about a company's industrial sector. We choose to retain SIC Code for subsequent binning
AccountID **Company** **Company Description**	Contained unique or near-unique values for each record, offering no meaningful patterns for prediction
Square Footage **Fiscal Year End**	High percentage of missing values
Company Status (Active/Inactive)	Low variance due to only 1 unique value across the dataset
Parent Country **Parent Company**	Explicitly excluded as per competition rules



(Figure 3: Geospatial plot of coordinates in Singapore with no clear pattern in response variable)

Feature Conversion: To enhance interpretability and model performance, we replaced two features with transformed versions

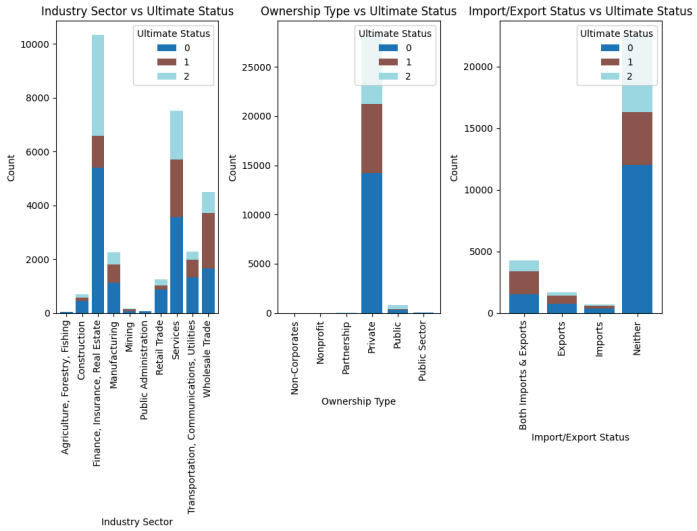
1. **Year Found** was converted to **Company Age**, derived from the current year. This provides a more intuitive measure of company longevity, making comparisons across firms more meaningful.
2. **SIC Code** were grouped into **Industry Sectors** based on predefined classifications. This reduced granularity, allowing the model to capture broader industry patterns instead of treating SIC codes as individual categorical values. (https://en.wikipedia.org/wiki/Standard_Industrial_Classification)

Handling Outliers: We removed impossible values under the assumption that sales and employee counts cannot be negative or zero. Three rows with negative **Sales (Domestic Ultimate Total USD)** were identified and removed due to their rarity and likely erroneous nature.

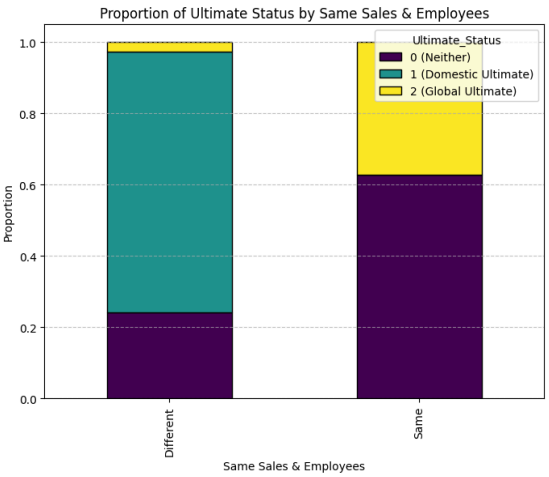
For categorical variables, we visualized distributions (Figure 4) and identified low-frequency categories that could introduce noise into the model. To reduce cardinality while preserving meaningful distinctions, we grouped infrequent industry sectors into an "Others" category, retaining only the most representative ones: Construction, Finance/Insurance/Real Estate, Manufacturing, Retail Trade, Services, Transportation/Utilities, and Wholesale Trade. Additionally, we simplified **Ownership Type** by merging all non-private entities into a single "Non-Private" category.

Handling Duplicates: After applying data cleaning and transformation strategies, we identified 6,761 duplicated rows in the dataset. These duplicates emerged after removing unique identifiers (e.g., **AccountID**, **Company**), indicating that some records were structurally similar but were previously distinguished by these attributes. Rather than removing these duplicates outright, we analyzed their **Ultimate_Status** distribution. The distribution was not uniform, suggesting these records still contained useful patterns rather than pure redundancy. To preserve class balance and prevent information loss, we chose to retain them, ensuring the model learns from the full range of patterns in the dataset.

Additionally, we discovered a large subset (67%) of companies with identical domestic and global sales/employee counts. Almost none of these companies were labeled as Domestic Ultimates, suggesting they were more likely owned by Global Ultimates. While we lacked domain expertise to ratoinalize this, it provided a valuable distinction for classification. To capture this pattern, we created a new feature, **SameSalesEmployees**, which flags companies where domestic and global sales and employee counts are identical. This feature may help the model differentiate between Domestic and Global Ultimates more effectively, potentially improving classification accuracy (Figure 5).



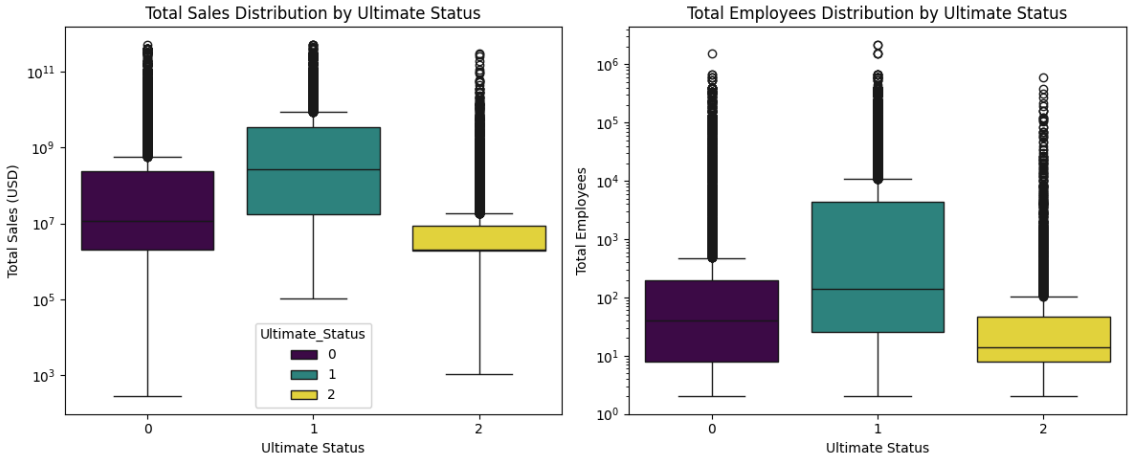
(Figure 4: Stacked bar plot of categorical features, color-coded by Ultimate Status proportions)



(Figure 5: Stacked bar plot of the Same Sales & Employees indicator, color-coded by Ultimate Status proportions.)

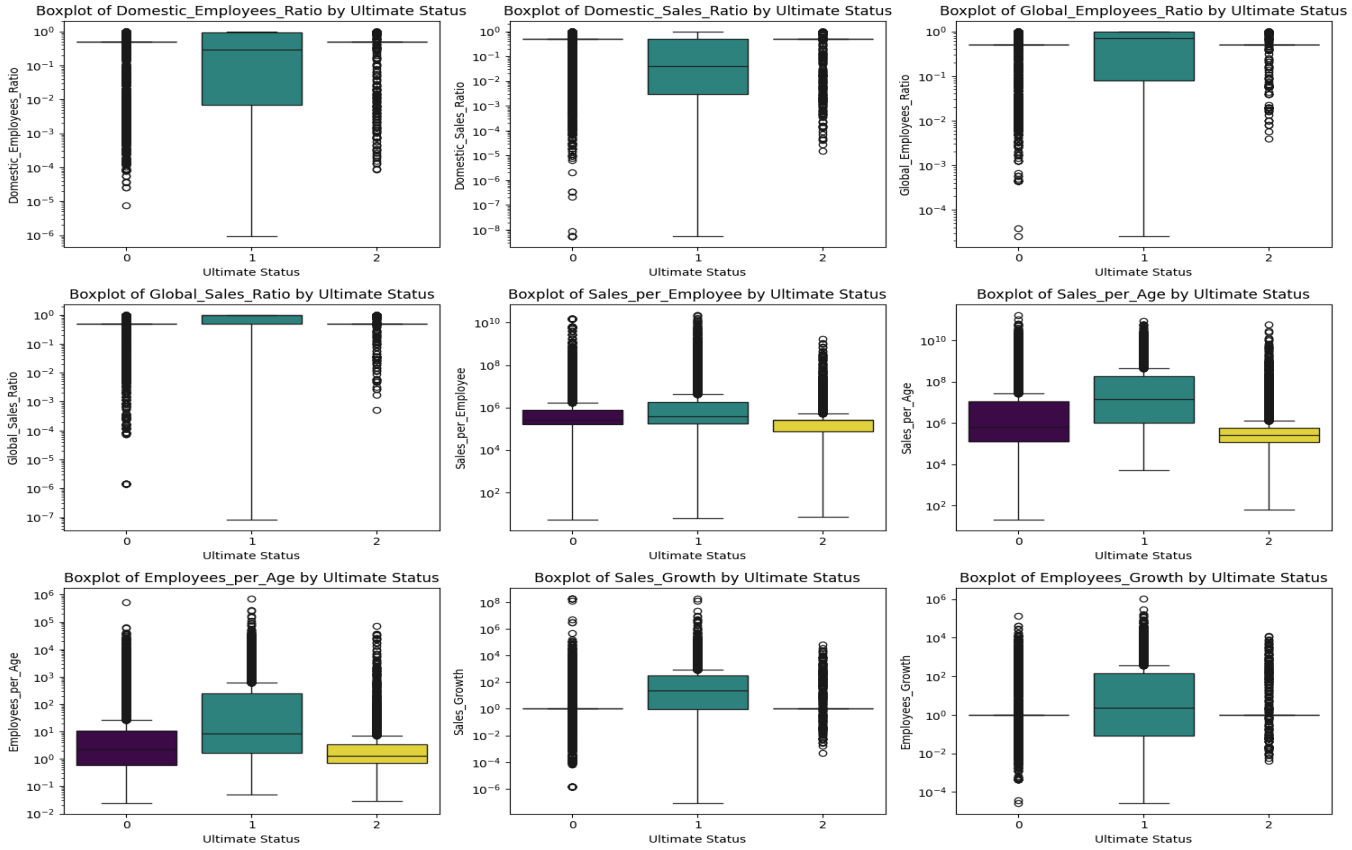
Feature Engineering:

Addition of Variables: Through iterative experimentation, we observed that our model was not performing well on the raw features alone, prompting us to backtrack and engineer additional features to improve predictive accuracy. One key enhancement was the creation of **Total_Sales** and **Total_Employees**, which sum each company's domestic and global sales and employee counts to capture overall company scale. A box plot (Figure 6) revealed that Domestic Ultimates tend to have the highest median sales and employees, distinguishing them from other ownership types. This insight suggests that total company size may be an important predictive factor, allowing our model to better differentiate between ownership classifications.



(Figure 6: Box plots of Total Employees and Total Sales, each split by Ultimate Status)

Ratio of Variables: To enhance predictive power, we engineered ratio-based features that capture relative relationships between sales, employees, company age, and growth patterns. These transformations provided a scale-invariant representation of company characteristics, helping the model generalize better across firms of different sizes. The following nine box plots (Figure 7) illustrate these features across **Ultimate_Status** classes. While not all features appeared significantly distinct, they collectively contributed to improving model performance, so we opted to retain them.



(Figure 7: Box plots of various features interactions, split by Ultimate Status)

3. Model Development

Model Choice

We considered various models, including distance-based models, tree-based models, and linear models. However, due to the skewed distribution of the data, distance-based models and linear models underperformed, even after applying transformations to address the skew. Tree-based models like Random Forests showed promise but were computationally expensive. Given the time constraints of this competition, we chose XGBoost, a gradient boosting algorithm known for its speed. This allowed us to test a variety of feature combinations quickly, providing a solid baseline model while mitigating overfitting through regularization. XGBoost's efficiency made it an ideal choice for our limited time frame. Additionally, XGBoost has the ability to produce interpretable decision trees that may help in understanding feature importance.

Scoring Choice

We chose to score our model based on the F1 score, primarily because it strikes a balance between precision and recall, which is crucial when we aim to correctly identify as many instances of the three classes as possible without excessively increasing the false positive rate (FPR). Given our limited domain expertise to assess the relative costs of false positives or false negatives (FPR or FNR), we focused on the F1 score and basic accuracy as our primary evaluation metrics. This approach allowed us to assess the model's discriminatory ability while minimizing trade-offs between different types of misclassifications.

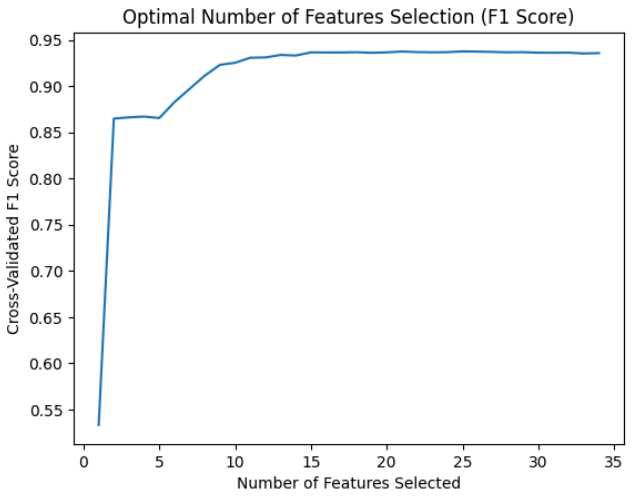
Preprocessing Features

One Hot Encoding: We one-hot encoded the categorical columns: **Ownership Type**, **Industry Sector**, and **Import/Export Status**. This process expanded the dataset by creating new binary columns for each category, resulting in a total of 33 features after the encoding. The original categorical columns were dropped to avoid redundancy.

Train-Test Split, Balancing, and Feature Scaling: Following the 80-20% train-test split as instructed, we applied Synthetic Minority Over-sampling Technique (SMOTE) to the training set. Although the data wasn't highly imbalanced, SMOTE improved model performance by enhancing class balance. While feature scaling isn't necessary for tree-based models, it helped in this case, likely due to the skew and scale of the data. We also used 5-fold cross-validation to validate the model and ensure more reliable performance metrics.

Recursive Feature Elimination

We applied Recursive Feature Elimination with Cross-Validation (RFECV) using an XGBoost model with default parameters to systematically eliminate the least important features from our feature set. At each step, the model evaluated the cross-validated F1 score, and we plotted the results to observe performance trends (Figure 8). Using the elbow method, we identified the optimal number of features where additional feature removal showed diminishing returns. Based on this, we conservatively selected 15 features (Figure 9), which were stored manually for downstream modeling.



(Figure 8: F1 score vs. number of features)

```
selected_features = [
    'Employees (Single Site)',
    'Employees (Domestic Ultimate Total)',
    'missing_single_employees',
    'SameSalesEmployees',
    'Domestic_Employees_Ratio',
    'Domestic_Sales_Ratio',
    'Global_Sales_Ratio',
    'Employees_Growth',
    'Ownership Type_Non-Private',
    'Ownership Type_Private',
    'Industry Sector_Finance, Insurance, Real Estate',
    'Industry Sector_Manufacturing',
    'Industry Sector_Services',
    'Import/Export Status_Both Imports & Exports',
    'Import/Export Status_Neither'
]
```

(Figure 9: Final selected features after RFE)

Hyperparameter Tuning

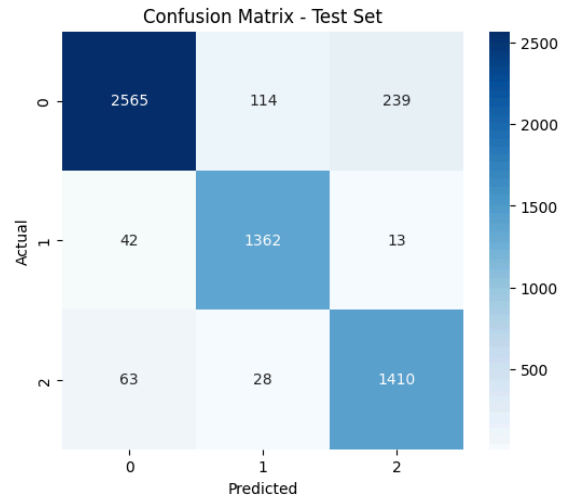
To balance performance and efficiency, we used GridSearchCV with 5-fold cross-validation and a focused parameter grid due to time constraints. We used standard values found in the literature to quickly identify a good starting point. The key parameters tuned were:

Parameter	Description	Range of Testing	Best Parameter
n_estimators	The number of boosting rounds (trees)	[100, 200]	100
learning_rate	Step size in gradient descent	[0.01, 0.1]	0.1
max_depth	Limits the depth of trees	[5, 10]	10
subsample	Fraction of training samples per tree	[0.7, 0.9]	0.9
colsample_bytree	Fraction of features used per tree	[0.7, 0.9]	0.9

3. Results

Model Performance

Our final XGBoost model, trained with the optimized hyperparameters and selected features, delivered strong performance on the test set. The model achieved an **accuracy of 91.45%** and a **weighted F1-score of 91.48%**, reflecting its ability to balance precision and recall across all classes. The confusion matrix (Figure 10) demonstrated that the majority of predictions occurred along the diagonals, indicating strong discriminatory power. The classification report indicates that all three ownership classes were well-predicted, with minimal misclassification. This demonstrates that the model effectively differentiates between corporate ownership types in unseen data, making it a robust solution for real-world applications.

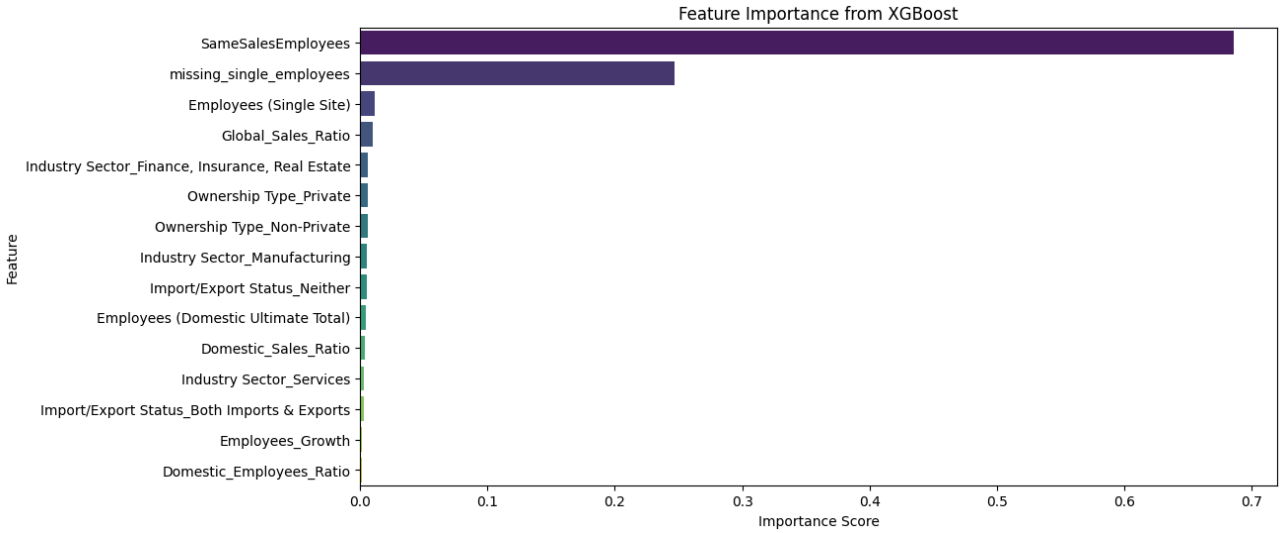


(Figure 10: Confusion matrix with color gradient)

Class	Precision	Recall	F1-Score
0 = Neither	0.96	0.88	0.92
1 = Domestic Ultimate	0.91	0.96	0.93
2 = Global Ultimate	0.85	0.94	0.89
Macro Average	0.9	0.93	0.91
Weighted Average	0.92	0.91	0.91
Accuracy	0.9145		

Key Insights on Feature Importance

To understand the drivers behind our model's predictions, we analyzed feature importance scores from the trained XGBoost model. These scores, ranked in descending order (Figure 11), highlight the most influential features in distinguishing corporate ownership structures. The following sections detail the key strategies and insights that significantly improved model performance.



(Figure 11: Bar plot of feature importance for selected features)

Data Anomalies as Predictive Signals: Interestingly, some of the strongest predictors emerged from data inconsistencies and missing values. One notable inconsistency was the presence of identical domestic and global sales/employee figures, despite the data dictionary stating these attributes were exclusive. This contradiction led us to engineer the **SameSalesEmployees** feature, which flagged such cases. Initially suspected as a data entry or interpretation error, it ultimately proved to be a top feature for classification. Similarly, **missing_single_employees**, which flagged missing values in **Employees (Single Site**), played a key role in predictions. While we assumed these missing values indicated zero employees, imputing zeros did not improve performance. Instead, treating them as a binary flag significantly enhanced predictive accuracy, suggesting these missing values captured meaningful structural differences. Another highly predictive feature was **Import Status**, which had many missing values. Rather than imputing them, we assumed missing values indicated companies that neither imported nor exported. Surprisingly, this assumption strengthened the feature's predictive power, suggesting that trade activity is a key factor in corporate ownership classification. Further collaboration with domain experts is warranted to uncover the underlying reasons for these patterns. Nevertheless, these findings highlight the potential value of data anomalies. While often dismissed as detrimental, we successfully leveraged data inconsistencies to uncover hidden patterns that enhanced predictive accuracy.

Feature Binning with Domain Knowledge: Columns with high cardinality, such as **SIC Code** and **Ownership Type**, were initially considered for removal due to their negative impact on model performance. Instead, we grouped SIC codes into broader industry sectors, reducing noise while retaining key distinctions. This transformation improved predictive power, uncovering patterns that were not evident in the raw data. For instance, Finance, Insurance, and Real Estate firms were more likely to be Global Ultimates, a relationship that became apparent only after binning. This underscores the importance of domain knowledge in feature engineering, as structured industry classifications can reveal insights that raw categorical variables fail to capture.

Improving Predictions Using Feature Interactions: Initially, single-variable transformations did not provide enough predictive power. However, interacting features such as summing values or computing ratios led to significant improvements. For example, **Global_Sales_Ratio**, **Domestic_Sales_Ratio**, and **Employees_Growth** captured important relationships between ownership structures and financial scale. These interactions exposed deeper structural patterns that isolated features could not, reinforcing the importance of feature engineering in driving model performance.

Response Variable Transformation:
A key breakthrough was refining our target variables. Initially, we treated **Is Domestic Ultimate** and **Is Global Ultimate** as independent binary targets, training separate models for each. This made EDA cumbersome, increased training complexity, and allowed the possibility of an impossible class. Restructuring the problem into a single three-class classification task simplified analysis and improved predictive performance. More importantly, it reinforced how problem understanding and domain knowledge are critical in designing effective machine learning solutions.

4. Limitations and Future Improvements

Feature Engineering & Data Quality: Proper domain expertise was needed to refine feature engineering and address missing or problematic data. Many assumptions could have been validated by consulting industry professionals, particularly regarding term definitions, useful transformations, and the meaning of missing values.

Collinear Features: Some features, such as Global Sales Ratio and Domestic Sales Ratio, were highly correlated and could have been removed earlier without impacting model accuracy. Identifying and addressing collinearity sooner would have improved model efficiency but due to time constraints, this refinement was not fully implemented.

Model Selection & Hyperparameter Tuning: Due to time constraints, extensive hyperparameter tuning and model testing were limited. XGBoost was chosen for its speed, but exploring linear models, distance-based methods, margin-based models, or neural networks could offer valuable comparisons. Additionally, a larger parameter grid could further optimize our model.

Model Interpretability: The model's complexity made it difficult to visualize decision trees. Feature importance alone wasn't enough, and methods like Partial Dependence Plots (PDP) and Local Interpretable Model-agnostic Explanations (LIME) could have provided better explanations for stakeholders.

5. Conclusion

Overall, this study highlights the importance of iterative feature engineering, strategic handling of missing data, and model optimization in corporate analytics for predictive modeling of ownership classification. Our project successfully developed an XGBoost-based model to classify corporate ownership structures, achieving a 91.45% accuracy and a 91.48% weighted F1-score. Through feature engineering, recursive feature elimination (RFE), and hyperparameter tuning, we optimized model performance, demonstrating that corporate ownership classification can be effectively predicted using financial, structural, and operational attributes.