1. Introduction

Purpose: Based on the operational, financial, and structural characteristics of the companies, a well-suited machine learning model is selected to accurately predict the **"Is Domestic Ultimate"** and **"Is Global Ultimate"** – provide valuable insights into a company's hierarchical and operational structure.

Context: In the modern global economy, it is essential for businesses, investors, and regulators to understand corporate ownership and influence. Organizations function within intricate structures, often comprising entities at both domestic and international levels.These classifications are essential for various applications, such as financial risk assessment, corporate governance analysis, and market research to help companies better navigate in this complex world.

Structure: This report outlines the *systematic approach undertaken to address this challenge.* It begins by exploring the dataset's characteristics and preprocessing requirements, followed by methodology used, model selection and results. Finally, it will end off with insights that Champion Group can utilise as part of their decision making processes.

2. Dataset Overview

2.1 Exploring Dataset

In the *Champions_Group_2025.csv* dataset, there are **29182 rows** and **22 columns**.

The **input variables** are:

'Latitude'	'Longitude'	'AccountID'	'Company'	'SIC Code'
'Industry'	'8-Digit SIC Code'	'8-Digit SIC Description'	'Year Found'	'Ownership Type'
'Company Description'	'Square Footage'	'Company Status (Active/Inactive)'	'Employees (Single Site)'	'Employees (Domestic Ultimate Total)'
'Employees (Global Ultimate Total)'	'Sales (Domestic Ultimate Total USD)'	'Sales (Global Ultimate Total USD)'	'Import/Export Status'	'Fiscal Year End'

*Yellow columns are part of the training data

The **test variables** (included in test data) are:

'Is Domestic Ultimate'	**'Is Global Ultimate'**

'Is Domestic Ultimate' signifies whether a particular entity or company is the ultimate or highest-level company within a corporate structure based in its home country.

'Is Global Ultimate' signifies whether a particular entity or company holds the status of being the ultimate or highest-level company within a corporate structure on a global scale.

2.2 Data Cleaning

2.2.1 Dropping Redundant Columns

'AccountID', 'Company', 'SIC Code', '8-Digit SIC Code','8-Digit SIC Description', 'Company Description' are removed because there are not determining factors of predicting whether the companies are 'Is Domestic Ultimate' or 'Is Global Ultimate'.

2.2.2 Settling Columns with large proportion of missing values

- **Deletion**

Since there is a high percentage of about **77%** missing values for `'Import/Export Status'` and `'Fiscal Year End'` and **42%** missing values for `'Employees (Single Site)'`, it is reasonable to drop these columns entirely as they provide insignificant insights to the data analysis. There is no information about `'Square Footage'`, which will be dropped as well.

- **Preservation**

`'Year Found'` represents a **historical fact** rather than a continuous variable that fluctuates with other company attributes. Since imputation could introduce artificial data, missing values for `'Year Found'` are dropped entirely to preserve data integrity.

- **Imputation**

`'Employees (Global Ultimate Total)'` is likely to be more dependent on other variables such as global sales. Thus, **KNN imputation** is used to impute missing values.

`'Employees (Domestic Ultimate Total)'` does not seem to have a significant relationship with other variables. Thus, **median imputation** is used to impute missing values.

`'Latitude'` and `'Longitude'` are *geographical features* that influence whether a company is classified as a global or domestic ultimate. Instead of dropping missing values, they are **imputed using the mean coordinates of companies in the same industry** to maintain geographic context as certain industries exhibit geographic clustering due to regional business hubs, supply chain dependencies, or regulatory factors.

2.2.3 Dropping Rows with missing values

There are **550 rows with missing values** for the columns `'Latitude', 'Longitude','Year Found',` which are removed from the dataset.

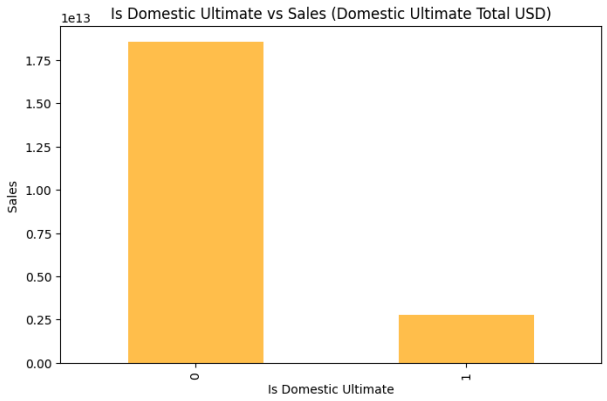
2.2.4 Dropping Rows with outliers

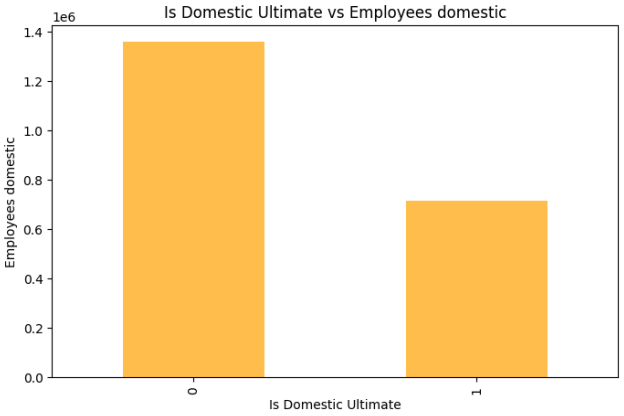
For numerical columns `'Employees (Domestic Ultimate Total)', 'Employees (Global Ultimate Total)', 'Sales (Domestic Ultimate Total USD)', 'Sales (Global Ultimate Total USD)'` which all spans across a large range of values, **logarithm transformation** is employed to normalise the data and make it easier to remove outliers.

Next, **z-score** is used to find outliers for the aforementioned numerical columns - a z-score above 3 or below -3 would be considered to be an outlier. In total, **256 rows with existing outliers** are removed.

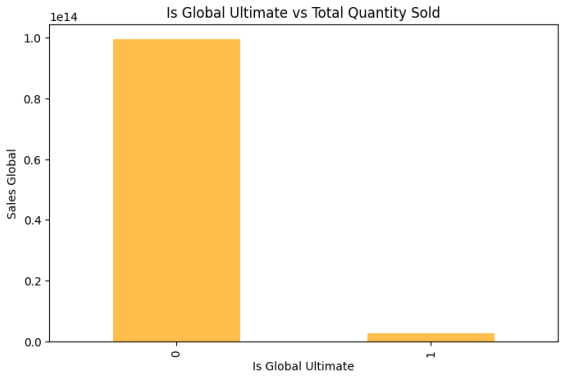
2.3 Finding the relationship between input variables and test variables

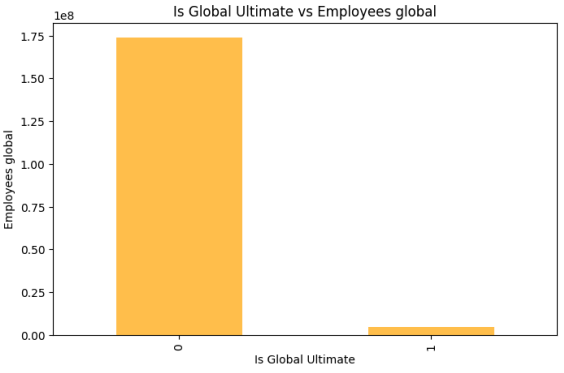
Using **Bar Plot**, it is found that there is a negative correlation between `'Employees (Domestic Ultimate Total)'` and `'Is Domestic Ultimate', 'Sales (Domestic Ultimate Total USD)'` and `'Is Domestic Ultimate'`.





There is also a negative correlation between `'Sales (Global Ultimate Total USD)'` and `'Is Global Ultimate', 'Employees (Global Ultimate Total)'` and `'Is Global Ultimate'`.





2.4 Data Preprocessing

2.4.1 Train-Test Split Validation

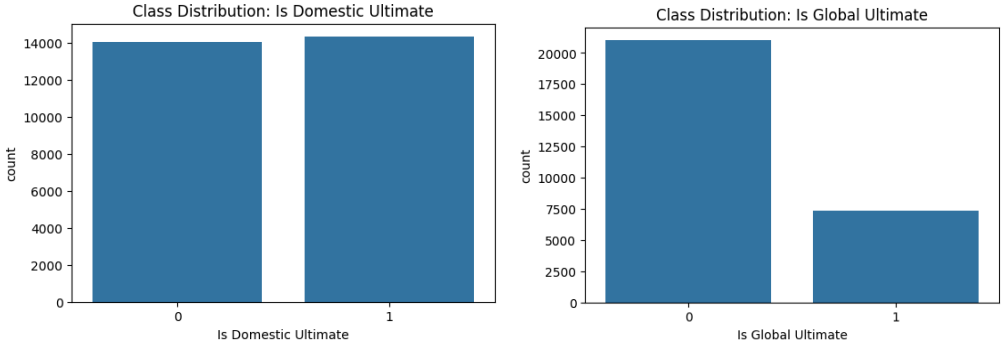
Dataset is divided into two parts - 80% training set and 20% testing set for model development in machine learning.

2.4.2 Scaling numerical data and encoding categorical data

The numerical data is **standardised** - transformed to have a mean of 0 and a standard deviation of 1. The categorical data is encoded using **One Hot Encode**r - converted into a numerical format by creating new binary columns for each unique category within a feature, with a "1" indicating the presence of that category and a "0" indicating its absence. This allows the machine learning model to interpret the data easily.

2.4.3 Checking Class Balance of "Is Domestic Ultimate" and "Is Global Ultimate"

The class distribution for `'Is Domestic Ultimate'` is balanced while the class distribution for `'Is Global Ultimate'` is imbalanced - class "0" outnumbering class "1" significantly.



Hence, **class weights** are used to balance the classes only for the `'Is Global Ultimate'` models and separate models are used.

3. Methodology

3.1 Machine Learning Model - Random Forest Tree

Random Forest is an **ensemble learning algorithm** that combines multiple decision trees to improve accuracy and reduce overfitting.

Advantages	Disadvantages
● High accuracy ● Reduce overfitting ● Works well with large datasets ● Provides feature importance	● Computationally expensive ● Consumes more memory

3.2 Approach and Frameworks to train the models:

We mainly made use of the "`scikit-learn`" package, to access the functions that were needed for our model training and evaluation.

3.2.1 Train-Test Split Validation

Dataset is divided into two parts - training set (80%) and testing set (20%) for model development in machine learning.

3.2.2 Data Preprocessing

Prior to training the model, the dataset was preprocessed to ensure compatibility with the Random Forest classifier. The features were divided into numerical and categorical variables. Numerical features were standardized, while categorical features were one-hot encoded using a preprocessing pipeline. This transformation was essential for ensuring that categorical variables could be effectively utilized within the model.

3.2.3 Model Selection and Training

A Random Forest Classifier was selected for classification tasks. Two separate models were trained:

1. `Is Domestic Ultimate` Classifier: A Random Forest model was trained using `n_estimators=200` and a `random_state=42` to ensure reproducibility.
2. `Is Global Ultimate` Classifier: Another Random Forest model was trained with the additional parameter `class_weight="balanced"` to address potential class imbalances. Additional hyperparameters such as:
   - `max_features=None`
   - `min_samples_leaf=4`
   - `min_samples_split=5`

   were incorporated to improve model performance.

The models were trained using the `fit()` method on the preprocessed training data and their respective target labels.

3.2.4 Model Evaluation:

1. **Classification Report**

The `classification_report` function from `sklearn.metrics` was used to assess precision, recall, and F1-score.

2. **Confusion Matrix**

Confusion Matrix is a table used to define the performance of the Random Forest model.

Accuracy can be calculated using **true positive (TP)** , **true negative (TN)** , **false positive (FP)** and **false negative (FN)** values.

$$\text{Accuracy} = \frac{\text{TN+TP}}{\text{TN} + \text{FP} + \text{FN} + \text{TP}}$$

3. ROC Curve and AUC Score

The Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC) were computed using the `roc_curve` and `auc` functions. This analysis enabled a comprehensive understanding of the model's ability to distinguish between classes.

4. Results

4.1 Classification Report from Random Forest (Is Domestic Ultimate)

Based on the dataset given, the Random Forest model has predicted `Is Domestic Ultimate = 0`.

Classification Report:

	precision	recall	f1-score	support
0	0.81	0.76	0.79	2809
1	0.78	0.82	0.80	2868
accuracy			0.79	5677
macro avg	0.80	0.79	0.79	5677
weighted avg	0.80	0.79	0.79	5677

Classification Report for `'Is Domestic Ultimate'`

Classification Accuracy:

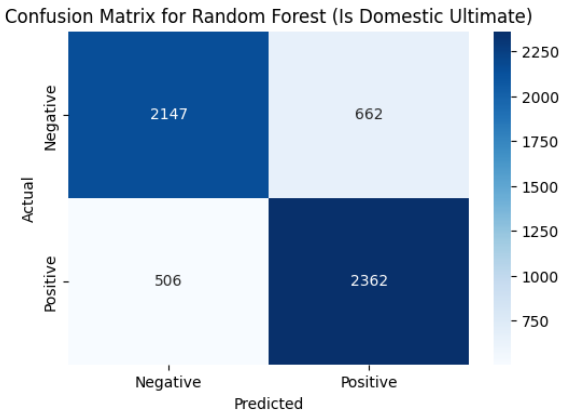
- The model achieved an overall accuracy of **79%**, correctly predicting the class for most instances.

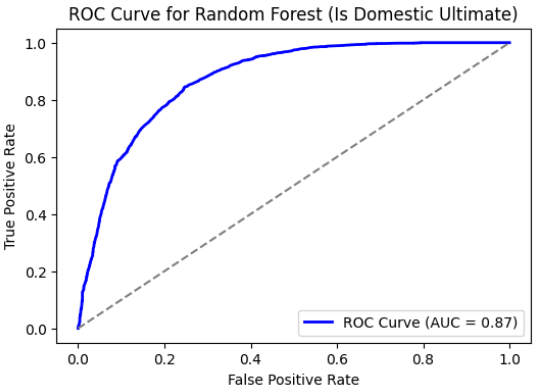
Class-specific Performance:

- For class 0 (not Domestic Ultimate), the model achieved a precision of **81%** and a recall of **76%**, resulting in an F1-score of **79%**.
- For class 1 (Domestic Ultimate), the model had a precision of **78%** and a recall of **82%**, with an F1-score of **80%**. This shows strong ability to identify Domestic Ultimate cases, though a few false positives are present.

Key Observations:

- The model performs slightly better in identifying class 1 (Domestic Ultimate) than class 0. However, there is some room for improvement in reducing false negatives for class 0.





ROC Curve

- The Receiver Operating Characteristic (ROC) curve illustrates the model's performance across different classification thresholds by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR).
- The Area Under the Curve (AUC) is 0.87, indicating that the model performs well in distinguishing between the positive (Domestic Ultimate) and negative (Not Domestic Ultimate) classes. An AUC of 0.87 suggests that the model has a high capability to rank predictions correctly, as a perfect model would have an AUC of 1.0.

Confusion Matrix

- The confusion matrix shows the distribution of actual vs. predicted classifications:
    - **True Negatives (2147):** The model correctly identified 2147 instances as not being Domestic Ultimate.
    - **False Positives (662):** These are instances incorrectly predicted as Domestic Ultimate when they were not.
    - **True Positives (2362):** The model correctly classified 2362 instances as Domestic Ultimate.
    - **False Negatives (506):** These are instances that were incorrectly classified as not being Domestic Ultimate.
- Overall, the model's strong true positive and true negative rates indicate a good balance between sensitivity (recall) and specificity.

4.2 Classification Report from Random Forest (Is Global Ultimate)

The Random Forest model has predicted `Is Global Ultimate = 0`.

```
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.84      0.88      4201
           1       0.65      0.82      0.72      1476

    accuracy                           0.84      5677
   macro avg       0.79      0.83      0.80      5677
weighted avg       0.86      0.84      0.84      5677
```

Classification Report for `'Is Global Ultimate'`

Classification Accuracy:

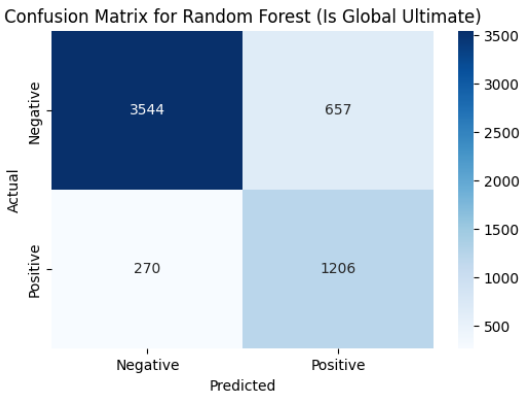
- The model achieved an accuracy of **84%**, indicating that it correctly predicted the class for a majority of the instances.

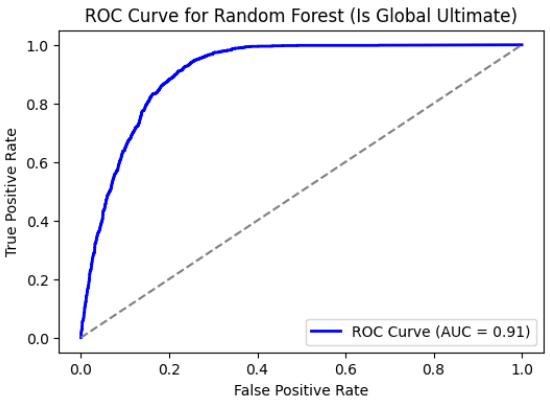
Class-specific Performance:

- For class 0 (not Global Ultimate), the model demonstrated high precision (**93%**) and a strong F1-score (**88%**), indicating its reliability in predicting non-Global Ultimate cases.
- For class 1 (Global Ultimate), the model showed moderate precision (**65%**) but maintained a good recall (**82%**), resulting in an F1-score of **72%**. This suggests the model effectively identifies most Global Ultimate cases, though there are some false positives.

Key Observations:

- While the model excels at identifying non-Global Ultimate cases (class 0), it struggles slightly with precision for class 1 (Global Ultimate). This may result in some false positives, where non-Global Ultimate cases are incorrectly classified as Global Ultimate.





Confusion Matrix

- The confusion matrix highlights the performance of the model on the test data:
  - **True Negatives (3544):** The model correctly predicted 3544 instances as not being Global Ultimate.
  - **False Positives (657):** These are instances incorrectly predicted as Global Ultimate when they were not.
  - **True Positives (1206):** The model correctly classified 1206 instances as Global Ultimate.
  - **False Negatives (270):** These are instances that were incorrectly classified as not being Global Ultimate.
- The results indicate strong performance with a high number of correct classifications (true positives and true negatives). However, there are still some misclassifications (657 false positives and 270 false negatives), which could potentially be reduced with further optimization.

ROC Curve and AUC Score

- The ROC curve illustrates the model's ability to differentiate between the positive and negative classes at various threshold settings.
- The Area Under the Curve (AUC) is 0.91, which is an excellent indicator of the model's performance. An AUC score close to 1.0 signifies that the model is highly effective at distinguishing between Global Ultimate and not Global Ultimate classes.

4.2 Feature Importance

Feature importance is used to determine which variable in the dataset affects the prediction of the test case.

As depicted in the figure below, the top three variables that have the highest level of importance are `Industry`, `Sales (Domestic Ultimate Total USD)`, `Employees (Global Ultimate Total USD)` in predicting `Is Domestic Ultimate`.

	Feature	Importance
0	Industry	0.223262
1	Sales (Domestic Ultimate Total USD)	0.117743
2	Employees (Global Ultimate Total)	0.114243
3	Latitude	0.112067
4	Employees (Domestic Ultimate Total)	0.110022
5	Longitude	0.108445
6	Year Found	0.104177
7	Sales (Global Ultimate Total USD)	0.099452
8	Ownership Type	0.010589
9	Company Status (Active/Inactive)	0.000000

Feature Importance for `Is Domestic Ultimate`

For the prediction of `Is Global Ultimate`, the top three variables that have the highest level of importance are `Employees (Global Ultimate Total USD)`,`Sales (Domestic Ultimate Total USD)`, `Industry`.

	Feature	Importance
0	Employees (Global Ultimate Total)	0.293205
1	Sales (Domestic Ultimate Total USD)	0.126384
2	Industry	0.111122
3	Sales (Global Ultimate Total USD)	0.105164
4	Latitude	0.092344
5	Year Found	0.090837
6	Employees (Domestic Ultimate Total)	0.076268
7	Longitude	0.071141
8	Ownership Type	0.033535
9	Company Status (Active/Inactive)	0.000000

Classification Report for `Is Global Ultimate`

5. Insights

5.1 Findings

From the results in the above section, the **type of industry**, **domestic sales revenue** and **employees across the global operations** of the company plays a significant role in determining whether the company is the highest-level company within a corporate structure on a global scale or in their respective home country.

5.2 Implications

For the company to be regarded as high level, it needs to **identify the emerging industries** within its home country. Furthermore, there needs to be **effective marketing strategies** to appeal to local customers who will generate greater revenue for the company. Lastly, a multinational company who has operations globally has greater likelihood to achieve the highest level, which means **merger and acquisition tactics** are effective to help companies expand their operations.

5.3 Limitations

Developing a separate model to address class imbalance for `'Is Global Ultimate'` increased both space and time complexity. This trade-off was necessary to achieve better class balance but resulted in a less streamlined model.

Due to time constraints, hyperparameter tuning was not thoroughly optimized. As a result, the model's performance may not have reached its full potential.

While the random forest model performed well, its precision is not flawless, which may result in occasional misclassifications that could affect the reliability of the results.

6. Conclusion

Our analysis showed that the Random Forest Model has outperformed other classifiers, achieving an accuracy of 91% in predicting the '`Is Global Ultimate`' category. Feature importance analysis revealed that '`Industry`' type and '`Employees (Global Ultimate Total)`' were the most significant factors. Our future work could explore other deep learning techniques or ensemble methods to further enhance accuracy. Better tuning of hyperparameters could be further done to our model to achieve a better precision. Additionally, expanding the dataset to include more companies across industries could improve generalization.