NUS SDS DATATHON 2025

CATEGORY B (Group 65)

Name:

Ang Wei Cheng

Goh Eng Ee, Koen

Han Dexun

Tan Wee En (Mavis)

Keira Low Wei-Qi

1. Introduction

A Domestic Ultimate is the highest-level entity within its home country that is not controlled by any other domestic company, while a Global Ultimate is the highest-level entity worldwide with no parent company anywhere in the world. Accurately predicting these classifications enables better competitive analysis, investment decisions, and merger and acquisition strategies, allowing stakeholders to assess a company's influence and financial backing.

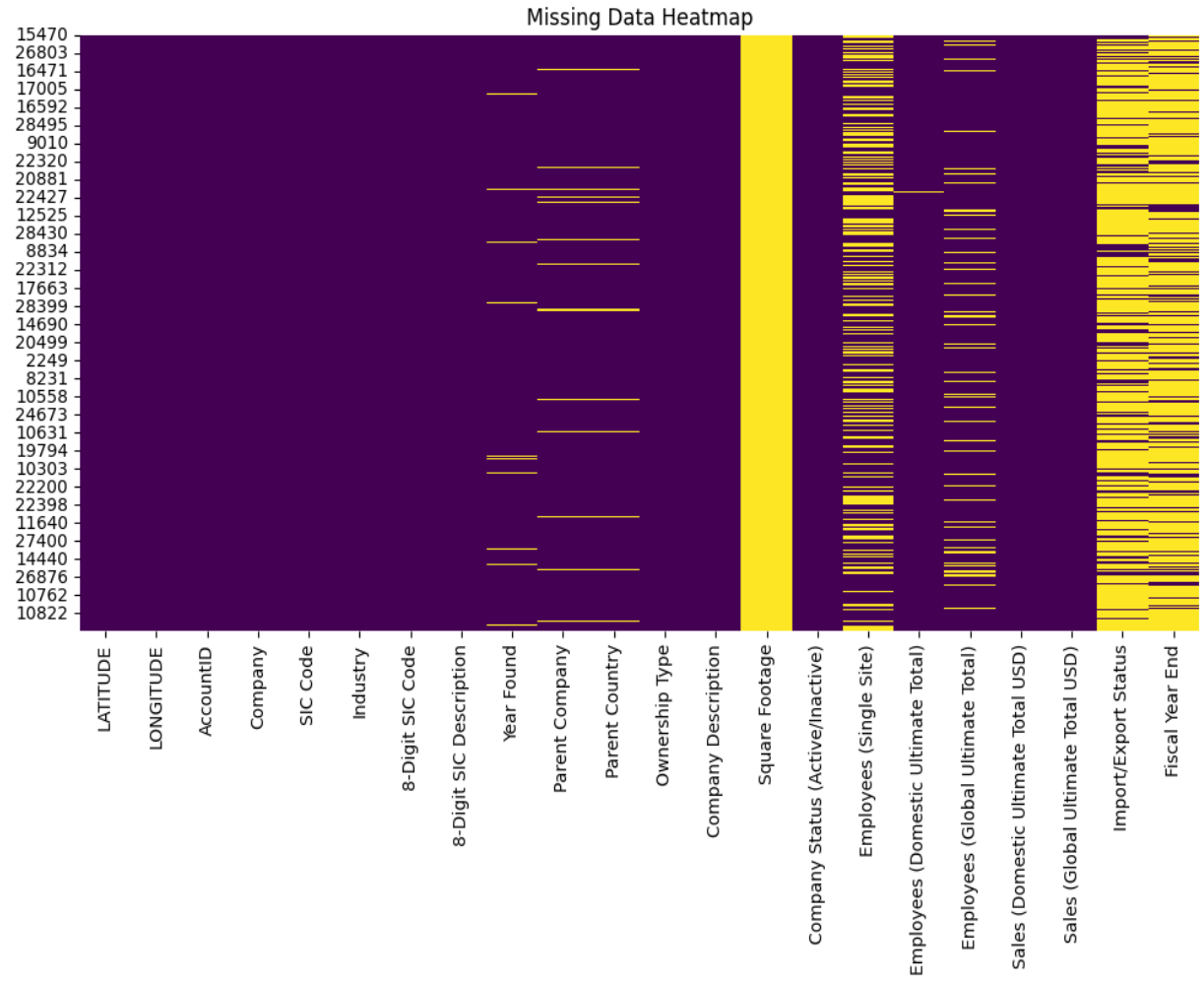
The objective of this project is to develop a machine learning model that can predict whether a company is a Domestic Ultimate or Global Ultimate based on various operational, financial, and structural characteristics.

This report explores data-cleaning techniques, the training of machine learning models and performance analysis.

2. Dataset Overview

The dataset includes key attributes such as industry classification, year found, total sales, number of employees and ownership status. However, the uncleaned datasets presented issues due to missing values. To counter this problem, we generated a heatmap to find which variables had a significant percentage of missing values.

Figure 2.1 Missing Data Heatmap



3. Methodology

3.1 Data Cleaning

Firstly, we removed columns with excessive missing values to prevent potential biases and maintain the data's integrity. The following columns were excluded: Square Footage, Parent Country, Parent Company, Import/Export Status, Fiscal Year End, Employees (Single Site), AccountID, Company, Company Description, and Company Status (Active/Inactive). These attributes either contained a high proportion of missing values or were deemed irrelevant to the analysis.

For numerical variables with missing values, we applied different imputation strategies. 'Year Found' was imputed using the median value within each Industry group, preserving industry-specific trends.

For missing geographical coordinates ('LATITUDE' and 'LONGITUDE'), we implemented the K-Nearest Neighbors (KNN) imputation with five nearest neighbors. This approach leverages existing geographical information to predict missing coordinates based on similar entities.

For missing employee count variables ('Employees (Global Ultimate Total)' and 'Employees (Domestic Ultimate Total)'), we used a supervised machine learning approach. Since these values are numerical and essential for downstream analysis, we trained a Random Forest Regressor using the available data to predict missing values.

Before model training, categorical variables such as 'Industry' and 'Ownership Type' were encoded using Label Encoding to convert categorical values into numerical representations. The model was trained using three key features: 'Industry', 'Ownership Type', 'Sales (Domestic Ultimate Total USD)'.

Once trained, the model was used to predict and fill in the missing values for employee-related attributes, ensuring a more robust dataset.

After handling the missing values, we converted relevant columns into categorical data types for efficient storage and processing. These columns included ‘SIC Code’, ‘8-Digit SIC Code’, ‘8-Digit SIC Description’, ‘Is Domestic Ultimate’, and ‘Is Global Ultimate’.

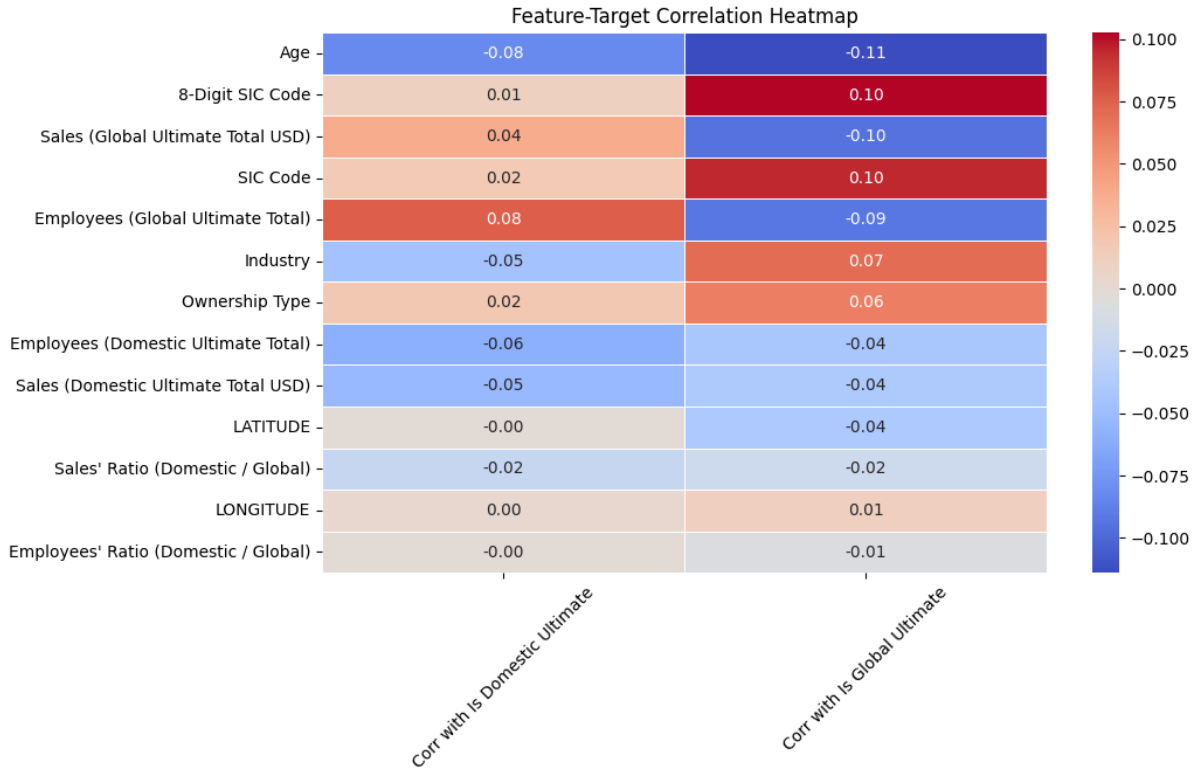
We also created new columns to enhance the dataset's predictive capabilities by calculating ‘Employees' Ratio’ and ’Sales' Ratio’ by taking the proportion of domestic sales to global sales, and ‘Age’, which is the age of the company by subtracting ‘Year Found’ by 2025.

Lastly, we separated the dataset into features (X) and target variables (Y), which are the variables ‘Is Domestic Ultimate’ and ‘Is Global Ultimate’.

3.2 Finding Correlation between Features and Target Variables

For each feature in the dataset, we calculated the Pearson correlation coefficients with the binary target variables. These coefficients allow us to analyse the strength and direction of the linear relationship between the features and the target variables.

Next, we use a heatmap to plot the correlation coefficients between the features and the target variables. This visualisation provides an intuitive representation of how the features are associated with the target variables. The diverging color scale (from cool to warm tones) helps distinguish between negative, positive, and weak correlations, with a legend on the right.



3.3 Data Modeling

Before training the models, we split the dataset into a training set (80%) and a testing set (20%) using stratified sampling to maintain class distribution across both splits. The target variables were converted into binary integer labels (0 or 1). Features were standardized using StandardScaler to ensure all features contributed equally to the model.

To balance class distribution, SMOTE was applied to the training set to oversample the minority class. Additionally, Tomek Links were used to remove noisy data and overlapping points between classes, enhancing data quality. Both of these methods significantly improved class balance, resulting in better model generalization.

In total, we used three different models: Logistical Regression, Random Forest Classifier, and XGBoost Classifier.

For each model, we performed hyperparameter optimization using Random Search Cross-Validation. Specifically, we defined a parameter grid and randomly sampled 10 different parameter combinations. For each combination, we applied 3-fold cross-validation to evaluate model performance. The results are then compared using the F1 score, allowing us to identify the optimal hyperparameters for each model.

3.3.1 Logistical Regression

We started by preparing the dataset through feature standardization. A logistical regression model was then trained on this processed data. To evaluate its performance, we measured accuracy, precision, recall, and the F1 score. This served as our baseline model, allowing us to compare its effectiveness against more complex approaches.

3.3.2 Random Forest Classifier

To capture more intricate patterns in the data, we implemented a Random Forest Classifier model. We optimized its parameters—such as the number of trees and their depth—using RandomSearchCV. To handle multiple target variables simultaneously, we wrapped the model within a MultiOutputClassifier. Performance was assessed using accuracy and F1 scores, with confusion matrices providing further insights into classification performance.

3.3.3 XGBoost (Extreme Gradient Boosting) Classifier

To address class imbalances, we applied SMOTE and Tomek Links before training. We then configured an XGBoost classifier, fine-tuning parameters such as learning rate, tree depth, and the number of estimators. After training on the balanced dataset, we evaluated the model's effectiveness using accuracy metrics and detailed classification reports for each target variable.

By following this approach, we aimed to identify the most effective model for predicting our target variables, balancing complexity with performance.

4. Results

After running all 3 models, we obtained various performance metrics. The tables below summarise the performances of the models.

4.1 Logistical Regression

Table 4.1 Score Matrix for Logistical Regression

	Is Domestic Ultimate	Is Global Ultimate
Accuracy	0.6872	0.7846
Precision	0.6629	0.7068
Recall	0.7619	0.2785
F1 Score	0.7089	0.3995
PR AUC	0.7441	0.5892

4.2 Random Forest Classifier

Table 4.2 Score Matrix for Random Forest Classifier

	Is Domestic Ultimate	Is Global Ultimate
Accuracy	0.7824	0.8131
Precision	0.7942	0.8086
Recall	0.7625	0.3587
F1 Score	0.7780	0.4969
PR AUC	0.8624	0.7610

4.3 XGBoost Classifier

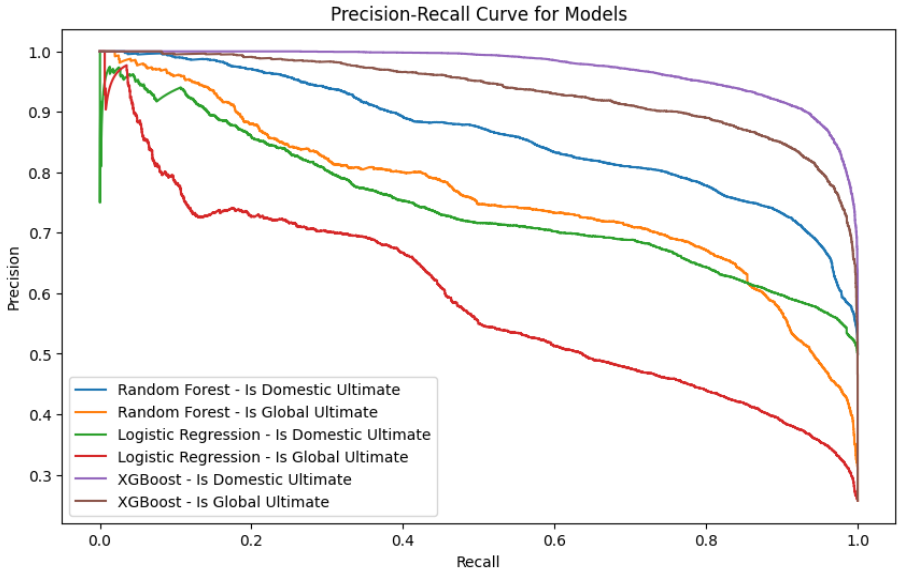
Table 4.3 Score Matrix for XGBoost Classifier

	Is Domestic Ultimate	Is Global Ultimate
Accuracy	0.9132	0.9329
Precision	0.8921	0.8503
Recall	0.9400	0.8973
F1 Score	0.9155	0.8732
PR AUC	0.9721	0.9331

We decided to analyse mainly the Precision Recall Area Under Curve (PR AUC) score further to determine which model was the best-performing one. The PR AUC score XGBoost model was the highest at 0.9721 for the 'Is Domestic Ultimate' model and 0.9331 for the 'Is Global Ultimate' model. This is followed by the Random Forest model with the second highest PR AUC score of 0.8624 for the 'Is Domestic Ultimate' model and 0.7610 for the 'Is Global Ultimate' model. Lastly, the logistic regression model has the lowest PR AUC score at 0.7441 for the 'Is Domestic Ultimate' model and 0.5892 for the 'Is Global Ultimate' model.

The PR AUC curve plots the precision (how many predicted positives are actually correct?) against the recall (how many actual positives are correctly identified?) rate. We chose the PR AUC score as a key determining factor of the accuracy of the model because the dataset given is imbalanced and the score focuses on the minority (positive) class rather than overall classification performance. PR AUC captures the trade-off between Precision and Recall, which is critical for rare event detection.

Figure 4.3 Precision-Recall Curve for Models



While we did keep track of a variety of performance metrics, due to the nature of the imbalanced dataset, our main performance indicator used was still the PR AUC score. It is evident that the scores mostly follow the trend of the XGBoost being the highest, followed by the random forest and then the logistic regression model. This allowed us to ensure that the performance of the different models measured by different metrics are not too far off from each other, increasing the reliability of our results. Hence, based on the PR AUC score, the XG Boost model is the best.

5. Insights

XGBoost has better performance than logistic regression. This is because logistic regression is a linear model, meaning it assumes a linear relationship between features and target. It struggles with complex patterns and non-linearity in data. In the case of the dataset given, there are many interacting factors and complex relationships between the variables. For example, not all high revenue companies are Global Ultimates. Categorical variables such as industry and ownership type also play an important role in determining the company's status. Logistic regression will be unable to capture such non-linear, complex relationships.

XGBoost outperforms random forest because it uses boosting whereas random forest uses bagging. Random forest trains multiple trees independently on random subsets of data, meaning that errors from earlier trees do not get corrected. However, XGBoost trains trees sequentially, where each tree corrects the mistakes of the previous trees. Trees focus on the hard-to-classify examples by assigning higher weights to misclassified examples, often leading to a better fit and higher accuracy.

Lastly, XGBoost conducts efficient feature selection, prioritising the more relevant variables and ignoring the less important ones. However, other models do not possess this feature, resulting in XGBoost being the highest-performing one out of the three.

However, we do concede that our model may not be the most accurate model in terms of classifying a company into 'Domestic Ultimate' or 'Global Ultimate'. This is because the dataset may not capture all of the relevant features that determine a company's corporate structure. Other factors such as economic conditions or industry-specific indicators may also play a huge part in predicting the company's status. For example, factors like market strategy, nature of global competitors and government policies could also have an impact on whether the company is a 'Domestic Ultimate' or 'Global Ultimate'. Hence, our model does have its limitations as not all of the relevant factors are captured in the data provided.

6. Conclusion

In this report, we explored the development of a machine learning model to classify companies as either a ‘Domestic Ultimate’ or ‘Global Ultimate’ based on operational, financial, and structural attributes. Our approach involved extensive data preprocessing, feature engineering, and the application of multiple machine learning models, including Logistic Regression, Random Forest, and XGBoost.

Among the 3 models tested, XGBoost emerged as the best-performing model, achieving the highest PR AUC scores of 0.9721 and 0.9331 for the Domestic Ultimate and Global Ultimate classifications, respectively. This result highlights the model’s ability to balance precision and recall, making it particularly effective in handling imbalanced datasets. The superior performance of XGBoost can be attributed to its boosting mechanism, which iteratively corrects classification errors, and its built-in feature selection, which prioritizes the most relevant attributes for prediction.

Ultimately, this study highlights the potential of machine learning in corporate classification and business analytics, providing valuable insights for stakeholders involved in investment decisions and strategic planning. By leveraging advanced algorithms and data-driven techniques, businesses can gain a deeper understanding of corporate structures and their implications on global markets.