README - NUS SDS DATATHON 2025 (CATEGORY B - Group 65)

Project Title: Predicting Domestic Ultimate and Global Ultimate Companies Using Machine Learning

1. Introduction

This project was developed as part of the **NUS SDS Datathon 2025 Hackathon**, where we aimed to build an **innovative, high-performing, and scalable machine learning model** to classify companies as **Domestic Ultimate** or **Global Ultimate**.

Accurate classification enables superior competitive analysis, strategic investment decisions, and targeted merger and acquisition strategies. We focused on designing a **practical, well-structured, and high-impact solution** that can provide meaningful insights into corporate ownership structures.

---

2. Environment Setup

Programming Language & Version

- **Python 3.10+**

Dependencies

- Jupyter Notebook

  Required Libraries:

  ```
  pip install pandas numpy scikit-learn xgboost seaborn matplotlib imbalanced-learn
  ```

- **GPU-enabled environment recommended** for faster model training.

Hardware Requirements

- **CPU-only execution is feasible**, but a **CUDA-enabled GPU (8GB+ VRAM recommended)** significantly speeds up model training.
- **Minimum 8GB RAM recommended**.

---

3. Running the Notebook

Main Notebook File Name

- CAT_B_65.ipynb

Execution Instructions

1. Clone the repository or download the project files.
2. Navigate to the project directory and open the Jupyter Notebook:
   ```
   jupyter notebook CAT_B_65.ipynb
   ```
3. Run the notebook **sequentially from top to bottom**.

Dataset Requirements

- **Dataset Location:** Ensure the dataset is available in the same directory as the main notebook.
- **Dataset Structure:** CSV file with the following columns:
  - Industry Classification
  - Year Founded
  - Total Sales
  - Number of Employees
  - Ownership Status
- **Expected Runtime:**
  - **Training:** ~10-20 minutes (depending on hardware).
  - **Inference:** ~1-2 minutes.

---

4. Model Execution Instructions

Models Used

1. **Logistic Regression** (Baseline Model)
2. **Random Forest Classifier**
3. **XGBoost Classifier** (Best Performing Model)

Testing the Model

- **Before running the model,** ensure that the test data is processed using the preprocess_test_data(test_data, preprocessor) method.
- Once the data is preprocessed, use the model to generate predictions.

---

5. Key Insights & Findings

Problem Addressed

- Helps businesses and investors **understand corporate ownership structures**.
- Supports **competitive analysis, investment decisions, and M&A strategies**.

Performance Comparison

Model Comparison Table for Is Domestic Ultimate

Model	Accuracy	Precision	Recall	F1 Score	PR AUC
Logistic Regression	68.72%	66.29%	76.19%	70.89%	74.41%
Random Forest	78.24%	79.42%	76.25%	77.80%	86.24%
XGBoost	**91.32%**	**89.21%**	**94.00%**	**91.55%**	**97.21%**

Model Comparison Table for Is Global Ultimate

Model	Accuracy	Precision	Recall	F1 Score	PR AUC
Logistic Regression	78.46%	70.68%	27.85%	39.95%	58.92%
Random Forest	81.31%	80.86%	35.87%	49.69%	76.10%
XGBoost	**93.29%**	**85.03%**	**89.73%**	**87.32%**	**93.31%**

Key Takeaways

- **XGBoost significantly outperformed other models**, with the highest PR AUC scores.
- **Boosting Mechanism** improves classification accuracy.
- **Feature selection helped reduce noise and improve prediction quality**.

Limitations & Future Enhancements

- **More real-world financial & economic indicators** could improve classification.
- **SHAP values** should be implemented for better model interpretability.

6. Conclusion

- We successfully built a **high-performing, hackathon-ready machine learning pipeline**.
- **XGBoost outperformed other models**, achieving the best trade-off between precision and recall.
- Future improvements should focus on **expanding features, optimising hyperparameters, and enhancing model interpretability**.

7. Team Contributors

This project was developed by **Team 65 - NUS SDS Datathon 2025**:

- **Ang Wei Cheng**
- **Goh Eng Ee, Koen**
- **Han Dexun**
- **Tan Wee En (Mavis)**
- **Keira Low Wei-Qi**

Final Note:

This project is **designed for hackathons**, using cutting-edge techniques to deliver meaningful business insights.

For further inquiries, contact **Group 65 - NUS SDS Datathon 2025**.