

Instructions for setting up the environment

1. Download the project folder “NUS_DATATHON_CAT_B_<66>”
2. Set the project folder as your root directory to ensure all paths work correctly
3. Place the dataset file “Champions_Group_2025.csv” into the project folder.
4. Install the libraries listed in **requirements.txt**

How to run your notebook and reproduce results

1. Open the CAT_B_<66>.ipynb
2. Run all

NOTE: If the **CAT_B_<66>.ipynb** file does not run in your preferred environment due to issues with installing the packages from **requirements.txt**. As a last resort, please open the file in Google Colab, drag and drop the data set under the content folder in google Colab. Then, uncomment the code chunk in **Section 1A** to install the necessary packages before executing **Run All**.

Any specific instructions required for executing the model

1. There is **NO** separate model file provided. All necessary code is included in the notebook
2. Sections 4C and 4D are currently commented out. These sections were used for hyperparameter tuning with recursive feature elimination (RFE), which takes a significant amount of time to execute. We have already completed this process and identified the best parameters for the model. These parameters are used in **Section 4E**
3. There is a known version conflict with some libraries, such as requiring scikit-learn==1.3.2 for running Sections 4C and 4D. Given this issue, we **do not recommend running these sections**
4. To execute the model and obtain results, simply run the entire notebook. Section 4E will output our model’s performance

Key insights and findings from your solution

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.