National College of Ireland

Configuration Manual

MSc Research Project
Research Project

Hayman Kyaw
Student ID: x24158593

School of Computing
National College of Ireland

Supervisor: Anderson Simiscuka

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Hayman Kyaw
Student ID:	x24158593
Programme:	MSc in AI for Business **Year:** 2025
Module:	Practicum Part 2
Supervisor:	Anderson Simiscuka
Submission Due Date:	10 December, 2025
Project Title:	Enhanced Customer Churn Prediction in Telecommunications: A Multi-Dataset Comparative Study of Ensemble Machine Learning Models with Explainable AI
Word Count:	**Page Count:** 5

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Hayman Kyaw

Date: 10 December, 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Hayman Kyaw
Student ID: x24158593

1 Introduction

This Configuration Manual details the hardware, software, and procedural steps required to successfully replicate the machine learning experiments conducted for this thesis. The goal is to ensure the project's reproducibility across the entire workflow, from data preprocessing to model training, evaluation, and interpretability using the LIME framework. The core analysis is conducted across three independent telecom churn datasets using four distinct classification models.

2 System Specifications

The machine learning models used in this study, particularly the optimized gradient boosting machines (XGBoost and LightGBM), are computationally efficient, allowing the research to be conducted on standard consumer-grade hardware.

2.1 Hardware Requirements

Component	Minimum Specification	Recommended Specification
Processor (CPU)	Dual-core, 2.0 GHz	Quad-core (or more), 2.5 GHz+
RAM	8 GB	**16 GB** (Recommended for handling large datasets like Dataset 3)
Disk Space	5 GB Free Space	10 GB Free Space (SSD recommended)
Hardware Used	Standard Local Machine / Laptop	Standard Local Machine / Laptop

2.2 Operating System

Component	Supported Operating Systems
OS	**Windows 10/11, macOS (10.15+), or Linux (Ubuntu 20.04+)**
Environment	Anaconda/Jupyter Notebook

3 Software Requirements and Dependencies

The entire project is built using the Python programming language and its data science ecosystem, managed via the Conda environment.

3.1 Core Software

Software	Version Used (Approx.)	Purpose
Python	3.12.6	Core programming language
Anaconda	Latest Stable Version	Environment and package management
Jupyter Notebook	Latest Stable Version	Interactive code execution and documentation

3.2 Key Libraries and Packages

Library	Version Used (Approx.)	Purpose
Pandas	2.2.3	Data loading, manipulation, and preprocessing
NumPy	2.2.6	Numerical array operations
Scikit-learn	1.5.2	Logistic Regression, Random Forest, Scaling, Cross-Validation, and Metrics
XGBoost	3.0.0	High-performance Gradient Boosting Model
LightGBM	4.6.0	High-performance Gradient Boosting Model (Leaf-wise growth)
LIME	0.2.0.1	Local Interpretability Framework
Matplotlib / Seaborn	3.9.2 / 0.13.2	Data visualization and plotting results (Feature Importance, ROC Curves, LIME Plots)

4 Installation Procedure

The following steps outline how to set up the required Python environment using Conda.

4.1 Environment Setup

1. **Install Conda:** Download and install the latest distribution of **Anaconda** for operating system.
2. **Create Environment:** Open the terminal/command prompt and create a new Python virtual environment to isolate project dependencies:

```
conda create -n thesis_env python=3.12.6
```

3. **Activate Environment:** Activate the newly created environment:

```
conda activate thesis_env
```

4.2 Installing Dependencies

Once the environment is active, install all required packages using pip or conda:

Bash

Install core data science and visualization libraries

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

Install high-performance models

```
pip install xgboost lightgbm
```

Install the interpretability framework and Jupyter

```
pip install lime jupyter notebook
```

4.3 Project File Structure

Ensure the following file structure is maintained. All data files must be in the same directory as their corresponding notebook.

/thesis_project/

├── Dataset 1.ipynb (Code for Dataset 1 analysis)

├── Dataset 2.ipynb (Code for Dataset 2 analysis)

├── Dataset 3.ipynb (Code for Dataset 3 analysis)

├── cell2celltrain.csv (Input Data 1)

├── telecom_customer_churn.csv (Input Data 2)

├── Telecom_customer churn.csv (Input Data 3)

├── [Other supporting files, e.g., output dataframes]

5 Execution Procedure

The analysis is executed by running the three sequential Jupyter Notebooks. Each notebook handles the entire end-to-end analysis for one of the three datasets.

5.1 Starting the Analysis

1. **Navigate to Directory:** In your terminal, navigate to the /thesis_project/ directory:

```
cd path/to/thesis_project
```

2. **Launch Jupyter:** Start the Jupyter Notebook server:

```
jupyter notebook
```

3. **Open Notebooks:** A web browser window will open. Start the analysis by opening the first file in the directory: Dataset 1.ipynb.

5.2 Running the Code

Each notebook must be run completely and sequentially to generate the required outputs for the thesis.

Step	Action	Description	Approximate Training Time per Notebook (depending on the machines)
1	Run Dataset 1.ipynb	Executes full analysis pipeline for the **Cell2Cell** dataset.	5 – 10 minutes
2	Run Dataset 2.ipynb	Executes full analysis pipeline for the **Telecom Customer Churn** dataset.	< 5 minutes
3	Run Dataset 3.ipynb	Executes full analysis pipeline for the third, **large Telecom churn** dataset.	10 – 20 minutes

Note on Time and Resources

The bulk of the time is spent training the boosting models (XGBoost, LightGBM) during the 5-fold cross-validation, which can take **1 to 5 minutes per fold**. The linear model (LR) runs in under 1 second. The total execution time is highly dependent on the system's CPU and RAM.

5.3 Key Output Generations

Upon successful execution of each notebook, the following critical outputs and artifacts are generated:

- **Performance Metrics Table and Plots:** A Pandas DataFrame containing the mean and standard deviation for the 5 key comparison metrics (Accuracy, Precision, Recall, F1, AUC-ROC) for all 4 models.
- **Global Feature Importance:** Bar charts and tables ranking the top features for RF, XGBoost, and LightGBM and coefficients for LR.
- **ROC Curves:** Plots generated for all models demonstrating the AUC-ROC values.

LIME Explanations: Bar charts illustrating the local feature effects for one example churner and one non-churner for the XGBoost and Random Forest models.

6 Configuration Files and Project Inventory

6.1 Input Data Files

The research uses three distinct customer churn datasets, all in standard comma-separated values (CSV) format.

File Name	Size (Approx.)	Number of Customers	Number of Features (Initial)
cell2celltrain.csv (for Dataset 1)	~12 MB	51,047	58
telecom_customer_churn.csv (for Dataset 2)	~1.4 MB	7,042	38
Telecom_customer churn.csv (for Dataset 3)	~45 MB	100,000	100

6.2 Code Files (Jupyter Notebooks)

The entire analytical process is consolidated into the following three notebooks, ensuring a clear and documented workflow for each dataset:

File Name	Purpose and Scope
Dataset 1.ipynb	Contains all code for preprocessing, training (LR, RF, XGBoost, LightGBM) with 5-fold Stratified CV, evaluation, feature importance, and LIME explanations for **Dataset 1**.
Dataset 2.ipynb	Contains all code for preprocessing, training, evaluation, feature importance, and LIME explanations for **Dataset 2**.
Dataset 3.ipynb	Contains all code for preprocessing, training, evaluation, feature importance, and LIME explanations for **Dataset 3**.