Configuration Manual

MSc Research Project
Data Analytics

Basil Varghese
Student ID: X23213574

School of Computing
National College of Ireland

Supervisor: Dr. Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Basil Varghese
Student ID:	X23213574
Programme:	Data Analytics **Year:** 2024-2025
Module:	MSC Research project
Lecturer:	Dr. Noel Cosgrave
Submission Due Date:	12/12/2024
Project Title:	Energy Forecasting in Commercial Buildings Using Property Features and Natural Resources
Word Count:	679 **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: Basil Varghese

Date: 12-12-2024

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

Basil Varghese
X23213574

1 Introduction

This manual demonstrates the implementation of the project "Energy Forecasting in Commercial Buildings Using Property Features and Natural Resources". The project aims to predict energy consumption of buildings using their property features and the natural resources used by the buildings for consumption, utilising machine learning algorithms like Random Forest, Logistic Regression, Ridge Regression, and XGBoost classifier. The manual provides details on system configuration, including hardware and software specifications, libraries required, and used to implement the project.

2 System Configuration

Listed below are the system configuration for the project

2.1 Hardware Specifications

Operating System	MacOS Sequoia 15.1.1
System Processor	Mac M1 Chip
Ram	8.00 GB
Storage	256 GB

2.2 Software Requirements

Programming Language Python	Python 3.9.15 and above
Tools	Jupyter Notebook

3 Project Artefacts

Important libraries that were imported during the project development process are listed in the following table.

3.1 Libraries to Import

Process Flow	Libraries Imported
Suppress warnings	warnings
Data manipulation	pandas, numpy
Visualization	matplotlib.pyplot, seaborn, plotly.express
Statistical analysis	scipy.stats, pearsonr
Data preprocessing	sklearn.preprocessing (StandardScaler, MinMaxScaler)
Dimensionality reduction	sklearn.decomposition.PCA
Feature selection	sklearn.feature_selection (SelectKBest, f_regression)
Model building	sklearn.ensemble.RandomForestRegressor, xgboost.XGBClassifier, xgboost.XGBRegressor, sklearn.linear_model.LinearRegression
Cross-Validation and Hyperparameter Tuning	
Model Evaluation	sklearn.metrics (mean_absolute_error, mean_squared_error, r2_score, mean_absolute_percentage_error)
Configuration utilities	plotly.io, pd.set_option

3.2 Libraries to Install

Library	Purpose
pandas, numpy	manipulation and operations
scikit-learn	Preprocessing
Xg-boost	XG-Boost Classifier

4 Project Artefacts

Figure 1 shows the project folder after it has been downloaded from a zip folder.

Figure 1 in the screenshot is a JSON file created when collecting a dataset from the U.S. Energy Information Administration (EIA) website. It shows the project folder that includes the dataset in CSV format, 'cbecs2018_final_public', and the project code, 'cbecs_Code.'

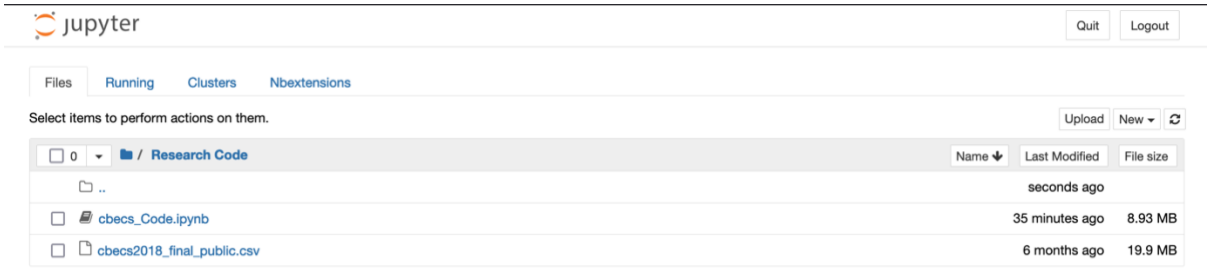

Figure 1.

Figure 2 shows a screenshot of the data set uploaded to the Jupyter Notebook for further analysis

Figure 2.

4.1 Dataset Analysis

The dataset is downloaded from the EIA website -Commercial building energy consumption 2018. It contains 6436 observations and 1249 variables. The figure shows the summary statistics of the top features in the dataset.

	PUBID	REGION	CENDIV	PBA	PUBCLIM	SQFT	SQFTC	WLCNS	RFCNS	RFCOOL	RFTILT	
count	6436.000000	6436.000000	6436.000000	6436.000000	6436.000000	6.436000e+03	6436.000000	6436.000000	6436.000000	6436.000000	6436.000000	643
mean	3218.500000	2.629739	5.105811	12.564947	3.646830	1.691617e+05	5.611871	2.379273	4.355811	1.501554	1.546613	
std	1858.057498	1.006943	2.439221	11.937664	1.701329	2.851251e+05	2.304431	1.741901	2.081375	0.500036	0.745358	
min	1.000000	1.000000	1.000000	1.000000	1.000000	1.001000e+03	2.000000	1.000000	1.000000	1.000000	1.000000	
25%	1609.750000	2.000000	3.000000	5.000000	2.000000	1.125000e+04	4.000000	1.000000	2.000000	1.000000	1.000000	
50%	3218.500000	3.000000	5.000000	13.000000	3.000000	6.900000e+04	6.000000	1.000000	5.000000	2.000000	1.000000	
75%	4827.250000	3.000000	7.000000	16.000000	5.000000	2.100000e+05	8.000000	3.000000	6.000000	2.000000	2.000000	
max	6436.000000	4.000000	9.000000	91.000000	7.000000	2.100000e+06	10.000000	8.000000	8.000000	2.000000	3.000000	1

Figure 3.

4.2 Visualisation

At the beginning of the study, the target variable ELCNS was right skewed (see Fig. 4), indicating that most of the values consumed relatively little energy. For an improved data distribution for modelling purposes, a log operation was performed on the data. The performed transformation of the data (Fig. 5) eliminates skewness and transforms the distribution into a normal distribution, making it more suitable for statistical analysis.

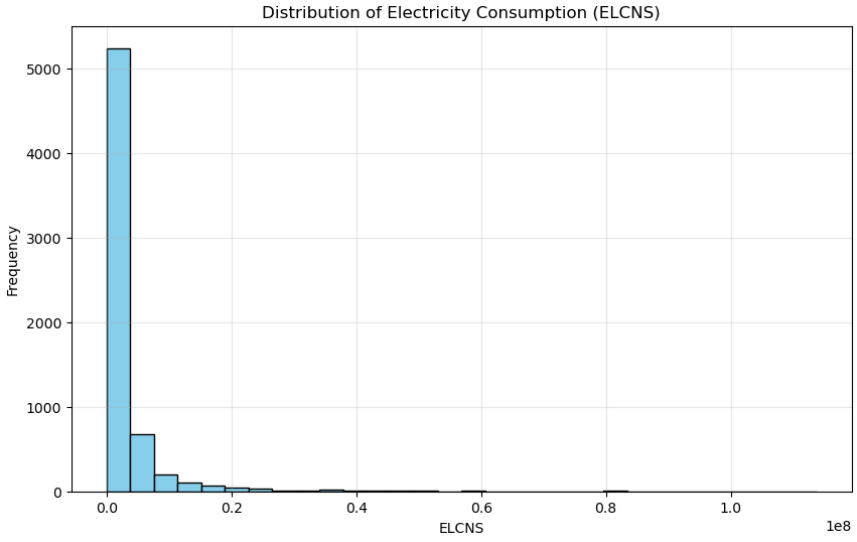

Figure 4

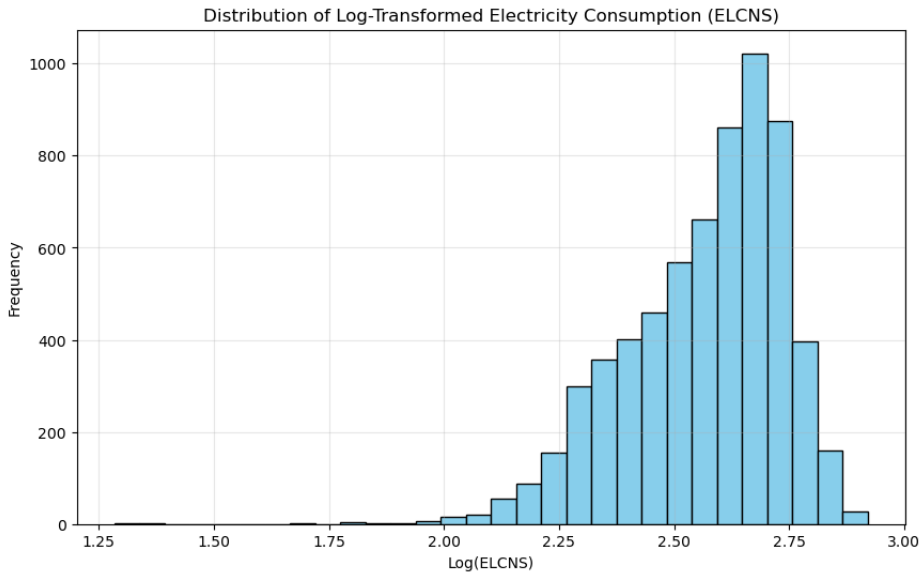
Distribution of Log-Transformed Electricity Consumption (ELCNS)
Frequency
Log(ELCNS)
0
200
400
600
800
1000
1.25
1.50
1.75
2.00
2.25
2.50
2.75
3.00

Figure 5