

Configuration Manual

Comparing Statistical and Machine Learning Models for Forecasting Milk Consumption in Ireland

MSc Research Project
Data Analytics

Utkarsh Arun Satpute
Student ID: x23135760

School of Computing
National College of Ireland

Supervisor: Teerath Kumar Menghwar

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name: Utkarsh Arun Satpute
Student ID: X23135760
Programme: Data Analytics **Year:** 2024-25
Module: Msc Research Project
Lecturer: Teerath Kumar Menghwar
Submission Due Date: 11/08/2025
Project Title: Comparing Statistical and Machine Learning Models for Forecasting Milk Consumption in Ireland
Word Count: 626 **Page Count:** 6

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: Utkarsh Arun Satpute

Date: 11/08/2025

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

Utkarsh Arun Satpute
Student ID: x23135760

1 Environment Setup and Dependencies

This project was developed and tested using the following system specifications and software libraries to ensure consistent results and reproducibility.

System Requirements:

- **Operating System:** Windows 10 / macOS / Linux
- **RAM:** Minimum 8 GB (Recommended 16 GB or more for deep learning models)
- **Processor:** Intel Core i5 or higher / Apple M1 / AMD Ryzen 5 or higher
- **Python Version:** 3.9 or above
- **Development Environment:** Jupyter Notebook (Anaconda distribution) or Google Colab for cloud-based execution

Python Libraries Used:

The following Python libraries were used for data handling, visualization, statistical modelling, and machine learning

- Pandas – Data manipulation and analysis
- NumPy – Numerical computations
- Matplotlib – Data visualization
- Seaborn – Statistical visualizations
- Statsmodels – Implementation of ARIMA, SARIMA, and Holt-Winters models
- Prophet – Time-series forecasting using Facebook Prophet
- Scikit-learn – Preprocessing, metrics calculation, and machine learning utilities
- Xgboost – Gradient boosting for regression tasks
- Tensorflow / keras – Implementation of LSTM deep learning models

All required libraries can be installed via the Python package manager pip using the following command:

2 Data Source and Preprocessing

Data Source:

The datasets used in this project were obtained from reputable Irish government data portals to ensure accuracy and reliability:

1. Central Statistics Office (CSO) – Import & Domestic Milk Intake Data:
 - URL: <https://data.cso.ie/>
 - Contains monthly records of domestic and import milk intake in Ireland, measured in million liters.
2. Data.gov.ie – AKM02: Milk Sales (Dairy) for Human Consumption

- URL:<https://data.gov.ie/dataset/akm02-milk-sales-dairy-for-human-consumption>
- Includes monthly sales data for Whole Milk, Skimmed Milk, and Semi-Skimmed Milk categories from January 1980 to July 2024.

Preprocessing Steps:

To prepare the dataset for forecasting, the following steps were performed.

1. Data Loading & Inspection:

- Imported the CSV files using pandas and examined their structure for column names, data types and missing values.

2. Data Parsing:

- Converted the Month column into proper datetime objects to facilitate time-series analysis.

3. Filtering by Category:

- Extracted separate datasets for:
- Whole Milk
- Skimmed Milk
- Semi-Skimmed Milk
- Domestic Milk Intake

4. Handling Missing Values:

- Applied forward fill (fill) to replace missing values with the last valid observation.
- Removed any rows with missing and unrecoverable data.

5. Sorting and Indexing:

- Sorted the data chronologically and set the Month column as the index.

6. Normalization (For LSTM Models):

- Used MinMaxScaler from scikit-learn to scale values between 0 and 1 for deep learning models, improving training stability.

7. Train-Test Split

- Training set: January 1980 – December 2020
- Testing set: January 2021 – July 2024

Preprocessing Libraries Used:

- pandas – for reading, cleaning, and filtering the datasets
- NumPy – for handling numerical operations
- scikit-learn – for scaling and splitting data

3 Model Execution Guide

This section provides instructions for executing the forecasting models used in this project: ARIMA, SARIMA, Holt-Winters, Prophet, LSTM, and XGBoost. Each model is implemented in separate Python notebooks to ensure modularity and ease of execution.

3.1 Project File Structure

The project files are organized as follows:

```
[ Data Sources ]
milk_consumption.csv
domestic_milk_intake.csv
|
v

[ Processing & Modelling ]
01_arima_forecasting.ipynb
02_sarima_forecasting.ipynb
03_holtwinters_forecasting.ipynb
04_prophet_forecasting.ipynb
05_lstm_forecasting.ipynb
06_xgboost_forecasting.ipynb
07_model_comparison.ipynb
|
v

[ Outputs ]
forecast_plots/
metrics_comparison.csv
```

3.2 Running the Models

Follow these steps to execute the models and generate results

Step 1: Environment Setup

- Install Python (version 3.9 or above).
- Install the required libraries by running.

Step 2: Data Placement

- Place the downloaded datasets (milk_consumption.csv and domestic_milk_intake.csv) in the /data folder.

Step 3: Executing Each Model

- Open Jupyter Notebook or Google Colab.
- Navigate to the /notebooks directory.
- Run the notebooks in the following order.

01_arima_forecasting.ipynb– Implements ARIMA for each milk category.
02_sarima_forecasting.ipynb – Seasonal ARIMA implementation.
03_holtwinters_forecasting.ipynb – Triple Exponential Smoothing model.
04_prophet_forecasting.ipynb – Prophet with parameter tuning.
05_lstm_forecasting.ipynb– LSTM using normalized data.
06_xgboost_forecasting.ipynb – XGBoost regression model.
07_model_comparison.ipynb – Compares performance metrics and generates a ranking chart.

Step 4: Output and Results

- Plots: Each notebook saves forecast graphs in /outputs/forecast_plots/.
- Metrics: RMSE, MAE, and R^2 values are stored in metrics_comparison.csv.
- Model Ranking: A bar chart ranks models based on RMSE scores for quick comparison.

References

Central Statistics Office (CSO), 2024. *Milk Intake and Sales Statistics*. [online] Available at: <https://data.cso.ie/>

Government of Ireland, 2024. *AKM02: Milk Sales (Dairy) for Human Consumption*. [online] Available at: <https://data.gov.ie/dataset/akm02-milk-sales-dairy-for-human-consumption>