National College of Ireland

MSc/PGDip in Data Analytics January 2023 Intake

Release Date: 6th April 2023
Submission Date: 5th May 2023

Statistics for Data Analytics

Terminal Assignment-Based Assessment - Individual Project

PART A – Time Series Analysis

Based on the work of the Climate Institute of the University of East Anglia two datafiles have been provided and are uploaded on Moodle: The datafile ‘nitm18442004.csv’ represents a monthly time series of average temperatures in Armagh from January 1844 to December 2004. The datafile ‘nity18442004.csv’ is a condensed version of the same data giving a timeseries of yearly average temperatures from 1844 to 2004. Both timeseries have different characteristics.

You are required to estimate and report on suitable models for both time series. Your report should contain the following elements:

- A preliminary assessment of the nature and components of the raw time series, using visualisations as appropriate.
- Estimation and discussion of suitable time series models from each of the categories listed below. Appropriate diagnostic tests and checks should be undertaken.
    i. Exponential Smoothing
    ii. ARIMA/SARIMA
    iii. Simple time series models
- Use the data up to and including 2003 as training set to forecast the average temperatures for 2004 and the actual data for 2004 as a test set to evaluate the forecast for the year. Use the monthly temperature data to forecast 12 months and the yearly temperature data to forecast for 1 year. Evaluate the forecasts against the actual data for 2004. Discuss your choice of an ‘optimum’ model for this series, from the above, which you should use to forecast. Provide commentary on the adequacy of your model for forecasting purposes.

PART B – Logistic Regression

The ‘*Diabetes Dataset.csv’* file, uploaded on Moodle, contains details of blood samples of diabetic patients collected in an Iraqi University Hospital in 2020 as published under the reference given below.

The file provided includes 12 relevant variables as follows:

Gender	Male / Female
Age	Patient Age
Urea	A diamine, chief nitrogenous waste product in humans
Cr	Creatinine Ratio, a parameter to assess kidney function
HbA1c	Average blood glucose (sugar) Levels
Chol	Cholesterol, a parameter to assess liver function
TG	Triglycerides a type of fat in the blood used to transport energy
HDL	High-density lipoprotein, the “good” cholesterol
LDL	Low-density lipoprotein, the “bad” cholesterol
VLDL	Very-low-density lipoprotein cholesterol
BMI	Body-Mass-Index
Diabetes	N / Y / P

Ignoring the Diabetes==’P’ data entries, you are required to estimate a binary logistic regression model to facilitate diabetes diagnostic based on blood results. Use the exploratory data analysis to make decisions about transformations of the variables. Split the dataset in a suitable manner into training and test dataset. Evaluate the models on the test dataset using a confusion matrix. Test your final model on the Diabetes==’P’ cases. What is the probability that these ‘prediabetic’ cases are diagnosed as diabetic?

In your report you should:

- Use descriptive statistics and appropriate visualisations to enhance understanding of the variables in the dataset.
- Describe the model-building steps you undertook to arrive at your final logistic regression model. The rationale for rejecting intermediate models should be explained clearly.
- Provide a succinct summary of the parameters of your final model, verify that relevant assumptions are met and discuss model performance and fit.

Reference:

Rashid, Ahlam (2020), “Diabetes Dataset”, Mendeley Data, V1, doi://10.17632/wj9rwkp9c2.1

General Instructions

All work submitted by students for assessment purposes is accepted on the understanding that it is their own work and written in their own words except where explicitly referenced. The report is subject to a maximum page count of 10 pages. Please use the IEEE conference format.

The final delivery consists of three parts which have to be uploaded separately:

1. The report covering part A and B in IEEE format submitted as .pdf file
2. Any files supporting your work on Part A
3. Any files supporting your work on Part B

The supporting files should contain all material required to reproduce the results of your report:

- If you used Jupyter Notebook, submit the notebook file with all the output produced included. Make sure that it reproduces using the "Restart Kernel and run all" option. For any computer-generated graphics you used in the report, insert in the Jupyter notebook a comment referring to the figure number or caption.
- If you used R Studio or similar, submit the source file and make sure that one can run the code sequentially. For any computer-generated graphics you used in the report, insert in the source code a comment referring to the figure number or caption.
- If you used a software package like SPSS upload the generated output and provide a .pdf document with a detailed description of the steps you have taken to obtain the results in your report.

To avoid upload problems with Moodle, package the supporting files in a separate folder and upload a zipped version of that folder on the Moodle Turnitin link provided.

Marks for the assignment will be allocated as follows:

Part A Time series analysis		**35%**
Assessment of the raw time series	(5)	
Investigation of suitable models	(20)	
Forecasting and assessment of the adequacy of the final model	(10)	
Part B Logistic regression modelling		**35%**
Descriptive Statistics and Visualisation	(10)	
Modelling Process and evaluation of intermediate models	(20)	
Discussion of final model performance and fit	(5)	
Supporting Evidence		**10%**
Reproducible Results for Part A	(5)	
Reproducible Results for Part A	(5)	
Report Quality		
Overall structure, flow, professionalism and clarity of the report		**20%**