National College of Ireland

# National College of Ireland

**MSc/PGDip in Data Analytics 2022/23**

**MSCDAD_A, MSCDAD_B, PGDDA_SEPOL**

**Release Date: 2nd December 2022**
**Submission Date: 5th January 2023**

---

**Statistics for Data Analytics**

**Hicham Rifai**

Terminal Assignment-Based Assessment - Individual Project

**PART A – Time Series Analysis**

The 'departure.csv' datafile, uploaded on Moodle, is a monthly time series of number of departures from Ireland via airports, commencing in 2010 to September 2022.

You are required to estimate and report on suitable models for this 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
- Forecast the number of departures in the first 6 months of 2021 and 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 *bank* file, uploaded on Moodle, contains details of a marketing campaign that aims to convince the customer to buy a bank product.

The file provided includes 16 variables as follows:

- Age
- Job
- Marital status
- Education
- Credit
- Housing:Has Mortgage
- Loan: has a personal loan?
- Contact communication type
- Month
- Last contact day of the week
- Duration: last contact duration, in seconds (numeric).
- Campaign: number of contacts performed during this campaign and for this client
- pdays: number of days that passed by after the client was last contacted
- previous: number of contacts performed before this campaign and for this client
- poutcome: outcome of the previous marketing campaign
- y - has the client subscribed for the bank product

Using these data, you are required to estimate a binary logistic regression model to facilitate understanding of the relationships between Marketing campaign characteristics and classification as yes or 'no'. If you deem it useful, you may employ dimension-reduction techniques.

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.

**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 format.

Projects should be uploaded on the Moodle Turnitin link by 17.00 on 5h January 2023. Penalties apply to late submissions in accordance with School of Computing practices.

Marks for the assignment will be allocated as follows:

| | | |
|---|---|---|
| **Time series analysis** | | **40%** |
| Assessment of the raw time series | (5) | |
| Investigation of suitable models | (25) | |
| Forecasting and assessment of the adequacy of the final model | (10) | |
| **Logistic regression modelling** | | **40%** |
| Descriptive Statistics | (5) | |
| Discussion of Modelling Process | (25) | |
| Discussion of final model performance and fit / Summary | (10) | |
| **Overall structure, flow, professionalism and clarity of the submission** | | **20%** |

Reference:

S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014