National College of Ireland

MSc/PGDip in Data Analytics 2021/22

MSCDAD_JAN22A_I, MSCDAD_JAN22B_I, MSCDA_JANOL, PGDDA_JANOL

Release Date: 1st April 2022
Submission Date: 6th May 2022

Statistics for Data Analytics

Terminal Assignment-Based Assessment - Individual Project

PART A – Time Series Analysis

The 'CarRegistrations.csv' datafile, uploaded on Moodle, is a monthly time series of new private car registrations in Ireland from January 1995 to January 2022 inclusive. (Source: Central Statistics Office, Ireland)

You are required to estimate and report on suitable time series 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 candidate time series models from each of the categories listed below. Appropriate diagnostic tests and checks should be undertaken.
    i. Exponential Smoothing / ETS models
    ii. ARIMA/SARIMA models
    iii. Simple time series models
- Discussion on your choice of an 'optimum' model for this series, from the above, which you should use to forecast for six periods ahead with prediction intervals. Provide commentary on the adequacy of your model for forecasting purposes.

PART B – Logistic Regression

The '*Default.csv'* file, uploaded on Moodle, contains details of the characteristics of 2700+ customers of a credit institution and whether they have a loan default on record or not.

In addition to the dichotomous dependent variable [No default on record (0) / Default on record (1)], customer characteristics provided are:

- Gender
  0=Male, 1=Female
- Age in years
- Years of education
- Retired
  0=not retired, 1=retired
- Household income in thousands
- Credit card debt in thousands
- Other debt in thousands
- Marital status
  0=unmarried, 1=married
- Home ownership
  0=rents, 1=owns home

Using these data, you are required to estimate a binary logistic regression model to facilitate understanding of the relationships between the given customer characteristics and classification of default. If you deem it useful, you may employ dimension reduction techniques. In your report you should:

1. Use descriptive statistics and appropriate visualisations to provide a preliminary understanding of the variables in the dataset.
2. Describe the model building steps you undertook in the process of arriving at your final logistic regression model. The rationale for rejecting intermediate models should be explained clearly.
3. Provide a succinct summary of the parameters of your final model, verify that relevant assumptions are met and discuss odds ratios, the confusion matrix and measures of model 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.

Your assignment is subject to a maximum page count of 8 pages.

Please use the fonts, layout and treatment of figures specified in the IEEE format.

Assignments should be uploaded on the Moodle Turnitin link by 17.00 on 6th May 2022. 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 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 / visualisations	(5)	
Discussion of Modelling Process	(20)	
Discussion of the final model and model fit / Summary	(15)	
Overall structure, flow, professionalism and clarity of the submission		**20%**