A picture containing text, clipart

Description automatically generated

National College of Ireland

**MSc/PGDip in Data Analytics January 2024 Intake**

**Release Date: 16th April 2024**

**Submission Date: 6th May 2024**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Statistics for Data Analytics**

Terminal Assignment-Based Assessment - Individual Project

**PART A – Time Series Analysis**

The datafile ‘CocoaPrices.csv’ represents a monthly time series of average cocoa prices as published by the International Cocoa Organisation (<https://www.icco.org>) from October 1995 to March 2024.

You are required to estimate and report on suitable models for the 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.

1. Simple time series models
2. Exponential Smoothing
3. ARIMA/SARIMA

* Use the data up to and including September 2023 as training set to forecast the average prices for the 6 months from October 2023 to March 2024. Evaluate the forecasts against the actual data for the period October 2023 to March 2024. Discuss your choice of an ‘optimum’ model for this series, from the models discussed above or another model of your choice. Provide commentary on the adequacy of the chosen optimal model for forecasting purposes.

**PART B – Logistic Regression**

The ‘*fraud.csv’* file, uploaded on Moodle, contains anonymised and partially normalised data of 283,726 credit card transactions, which are classified into fraud and non-fraud transactions. The dataset is highly imbalanced: there are only 473 fraud transactions.

The file provided includes 31 relevant variables as follows:

|  |  |
| --- | --- |
| Time | Time of transaction (measured in seconds over two days) |
| Amount | Transaction amount in Euro |
| Class | 1=Fraud 0=Non fraud. |
| V1..V28 | Additional parameters that have been normalised to N(0,1) |

Use the exploratory data analysis to make decisions about transformations of the variables. Evaluate your intermediate models using a balanced confusion matrix. Test your final model on the complete dataset using weighted accuracy.

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: Alenzi, H.Z.; Aljehane, N.O. (2020): Fraud Detection in Credit Cards Using Logistic Regression. IJACSA *Int. J. of Advanced Computer Science and Applications*, **11**(12)540-551

<https://thesai.org/Downloads/Volume11No12/Paper_65-Fraud_Detection_in_Credit_Cards.pdf>

**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. Remember that there is a file upload limit of 100MB un compressed data.

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%** |