National College of Ireland

National College of Ireland

MSCDA_JANOL25, MSCDAD_JAN25{A,B,C}_I, PGDDA_JAN25

Release Date: Monday, 3rd March, 2025
Due Date: Monday, 24th March, 2025
Christian Horn, Vladimir Milosavljevic, Hicham Rifai

Statistics and Optimisation

Continuous Assessment (CA) Type: Project

Weight: 35% The assignment will be marked out of 100.

Instructions:

Your task is to find the best suited Logistic Regression model for the readmission of patients undergoing diabetes related treatment in a hospital. The problem was first investigated by B.Strack et al (2014). The attached dataset **hospitaldata.csv** (22.5MB) is derived from the original dataset. It contains 101,763 rows with 47 columns. A description of the variables in the dataset is given below. To analyse the dataset you should use SPSS or Jupyter Notebook with either Python or R.

SUBMISSION DETAILS:
The submission consists of two parts:

1. a report of up to 6 pages in **.pdf** format using the IEEE conference template and
2. the code used as a **.ipynb** file and any supporting files compressed into a single **.zip** file.

Late submissions will not be penalized if the student applied for an extension through NCI360 and it was approved.

In your report you should:

- Describe the statistical model used and the methodology applied
- Use descriptive statistics and appropriate visualisations to demonstrate an understanding of the variables in the dataset
- Document the important relationship between independent variables and the dependent variable.
- Describe the model building steps you undertook in the process of arriving at your final regression model. The rationale for rejecting intermediate models should be explained clearly and details should be provided on the rationale for the chosen predictors, transformations undertaken, treatment of outliers, etc.
- Provide details on the diagnostics undertaken, and
- Provide a succinct summary of the parameters of your final model, details of model performance and fit.

The supporting file should contain intermediate versions of your final report (numbered in sequence) and the code required to reproduce the final results of your report:

- The intermediate versions of your report document your working process. It is your defence in case your final report is flagged as potentially written by AI tools. It is ok to use such tools to improve your use of language as long as you provide the raw version of your submission as you have written it yourself. Make sure that the intermediate versions of your text still have the original timestamps when you created them.
- If you used **Jupyter Notebook**, submit the notebook file with all the output included. Make sure that it works sequentially by using the "Restart Kernel and run all" option and then save the file. 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 **any other programming environment**, make sure to submit a single pdf file that contains all the commands you have executed and the corresponding output in the exact sequence of execution. 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**, provide a .pdf document with a detailed description of the steps you have taken to obtain the results in your report. Make sure that the .pdf file contains screenshots of the relevant interactions.
- If you use code snippets from the classes or other sources, make sure to include in the code a reference to the source and a comment clarifying if and what parts of the code you have modified yourself.

The dataset attached contains the following variables:

Variable	Type	Description
encounter_id	ID	Unique identifier of an encounter no
patient_nbr	ID	Unique identifier of a patient no
race	Categorical	Values: Caucasian, Asian, African American, Hispanic, Other
gender	Categorical	Values: male, female
age	Categorical	Grouped in 10-year intervals: [0, 10), [10, 20),..., [90, 100)
weight	Categorical	Grouped in 25-pounds intervals: [0-25), [25-50), [50-75),…, [175-200), >200 or Unknown
admission_source	Categorical	Values: Referral, Emergency, Other
time_in_hospital	Integer	number of days between admission and discharge
discharge_type	Categorical	Values: Home, Other
medical_specialty		Specialty of the admitting physician. Values: Internal Medicine, Emergency/Trauma, Family/GP, Cardiology, and Other
num_lab_procedures	Integer	Number of lab tests performed during the encounter
num_procedures	Integer	Number of procedures (other than lab tests) performed during the encounter
num_medications	Integer	Number of distinct generic medications administered during the encounter
number_outpatient	Integer	Number of outpatient visits of the patient in the year preceding the encounter
number_emergency	Integer	Number of emergency visits of the patient in the year preceding the encounter
number_inpatient	Integer	Number of inpatient visits of the patient in the year preceding the encounter
number_diagnoses	Integer	Number of diagnoses entered into the system during the encounter
max_glu_serum	Categorical	Indicates the range of the lab test. Values: Norm, >200, >300, and None if not measured
A1CResult	Categorical	Indicates the range of the lab test. Values: Norm, >7, >8, and None if note measured

metformin, epaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide_metformin, glipizide_metformin, glimepiride_pioglitazone, metformin_rosiglitazone, metformin_pioglitazone	Categorical	Indicates whether the drug was prescribed and if there was a change in the dosage. Values: Up, Down, Steady, and No, if the drug was not prescribed
change	Categorical	Indicates if there was a change in diabetic medications (either dosage or name). Values: Yes, and No
diabetesMed	Categorical	Indicates if there was any diabetic medication prescribed. Values: Yes, and No
readmitted	Categorical	Indicates future readmission. Values: No, Within30Days, and After30Days

Please note:

1. The information given in the table above is for you orientation only. With regards to the exact spelling of column names or values, you have to check the data file provided.
2. The target variable **`readmitted`** has three values. To solve a Logistic Regression problem you have to transform the variable into a dichotomous variable by changing the encoding, for example Within30Days and After30Days to Yes.

Academic Integrity

- By submitting your work on Moodle you declare that this is your own work.
- Any material created by others (human or AI) must be properly referenced. Verbatim text copies should be included in quotes.
- Figures not created by yourself should include an acknowledgement detailing the name(s) of the creator(s) and proper references.
- Code and figures copied from class material or other sources should be clearly marked as such and properly referenced. In particular it should not be (directly or implicitly) claimed as your own. Instead a comment should be included in the source code indicating where you obtained it from.
- Students are strongly advised to familiarise themselves with the Guide to Academic Integrity. All submissions will be electronically screened for evidence of academic misconduct, e.g. plagiarism, collusion and misrepresentation.

TURNITIN:

All report submissions will be electronically screened for evidence of academic misconduct (i.e., plagiarism and collusion)

Grade Criterion	H1 (> 70%)	H2.1 (> 60%)	H2.2 (> 50%)	Pass (> 40%)	Fail (< 40%)
Abstract, Keywords, Introduction (10%)	A well written introduction to the problem and the choosen model class and the proposed methodology. The Abstract gives a short summary. A small number of keywords cover the project and the report.	The introduction covers the required elements, but style and language could be improved. Abstract and Keyword are present.	The introduction is covers some of the aspects required. Abstract or Keywords are missing	The introduction is incomplete and contains some technical mistakes. Abstract or Keywords are missing	The Introduction contains major technical errors or is only loosely related to the project.
Explorative Data Analysis (20%)	Descriptive Statistics and Visualisation has been provided for a comprehensive set of independent and dependent variables. The relationship between independent variables and the dependent variable has been investigated and is documented for some essential variables. An interpretation of the relationship has been given.	An attempt on the EDA has been made covering all aspects required for a small, but representative subset of variables.	An attempt on the EDA has been made, covering all aspects for a few arbitrarily selected variables.	A limited attempt on the EDA has been made, covering only some of the aspects.	The EDA appears not related to the project.
Model Development (40%)	There is a clear description of the methodology and the class of models used. The process of incremental improvement of models is clearly documented including a diagnostic of intermediate stages. The proposed final model is evaluated and the its parameters are interpreted.	There is a clear description of the methodology and the class of models used. Some of the intermediate steps appear ad hoc. The proposed final model is evaluated and the its parameters are interpreted.	An argument has been made for the class of models used, but there is only one (final)model given. There is only a limited evaluation of the final model or interpretation of its parameters.	There are no arguments for the proposed model and no interpretation of the model.	The proposed model doesn't fit the problem at hand.
Code (20%)	The code executes flawlessly. The results are reproducible. The code follows the methodology outlined in the report.	There are minor glitches in the code that can be fixed. The results are consistent with the report. The code follows by and large the methodology outlined in the Introduction.	The code doesn't follow the methodology, but is still makes some sense. Some of the results documented in the report can be reproduced.	The code doesn't execute, but there are good elements in the code	It is unclear how the code was derived or what the code is supposed to do.
Presentation (10%)	The text clearly outlines the project and documents the results as they are supported by the code. The document adheres to the template and the page limit. The writing style is excellent. Figures are clearly readable. Language and Grammar are correct.	A good attempt. There are minor deviations between the results in the code and the text. The document follows the template and the page limit. Language and Grammar are good.	There are substantial deviations between the results in the code and the text. The text does not match the template. Some figures are hard to read. Language and Grammar could be improved.	There is a disconnect between the code and the text submitted. There is no logical flow in the text. The text is difficult to read.	There is a major disconnect between the code and the text submitted. The text is littered with typos, there is poor use of English.