Intro to Machine Learning (Report Submission)

Start Assignment

Due Apr 16 by 23:59 **Points** 100 **Submitting** a file upload **File Types** pdf

Course code and name: COSC2753 Machine Learning

Length: See specific instructions below

Type: Individual

Feedback mode: Written feedback

Late work: For assignments 1 to 10 days late, a penalty of 10% (of the marks awarded) per day will apply. For assignments more than 10 days late, a penalty of 100% will apply. Weekend days (Saturday and Sunday) are counted when counting total late days.

Learning Objectives Assessed

- **CLO1:** Understand the fundamental concepts and algorithms of machine learning and applications
- **CLO3:** Set up a machine learning configuration, including processing data and performing feature engineering, for a range of applications
- **CLO4:** Apply machine learning software and toolkits for diverse applications

Ready for Life and Work

Completing this assessment would help you achieve the following employability skills:

- Analyse and model requirements and constraints for the purpose of designing and implementing solutions to a learning challenge
- Evaluate and compare approaches and algorithms on the basis of the nature of the problem/task being addressed
- Interpret abstract theoretical propositions, choose methodologies, justify conclusions and defend professional decisions to both IT and non-IT personnel via technical reports of a professional standard and technical presentations

Assessment Details

This assignment is designed to help you become more confident in applying machine learning.

In this assignment you will explore a real data-set to practice the typical machine learning process which includes:

- Selecting the appropriate ML techniques and applying them to solve a real-world ML problem.
- Analysing the output of the algorithm(s).
- Research how to extend the modelling techniques that are taught in class.
- Providing an ultimate judgement of the final trained model that you would use in a real-world setting.

To complete this assignment, you will require skills and knowledge from lecture and lab material for Weeks 1 to 5 (inclusive). You may find that you will be unable to complete some of the activities until you have completed the relevant lab work. However, you will be able to commence work on some sections. Thus, do the work you can initially, and continue to build in new features as you learn the relevant skills.A machine learning model cannot be developed within a day or two. Therefore, start early.

This assignment has three deliverable:

1. A PDF version of the python notebook. This should include rationale and critical analysis of your approach and ultimate judgement.
2. A set of predictions from your ultimate judgement.
3. Your Python scripts or Jupyter notebooks used to perform your modelling & analysis with instructions on how to run them

Task

Hospitals are constantly challenged to provide timely patient care while maintaining high resource utilization. While this challenge has been around for many years, the recent COVID-19 pandemic has increased its prominence. For a hospitals, the ability to predict length of stay (LOS) of a patient as early as possible (at the admission stage) is very useful in managing its resources.

In this assignment, you will develop a ML model to predict when a patient will be discharged from a hospital (see task below for exact definition) based on several attributes(features) related to diagnostic attributes, patient characteristics and hospital characteristics.

The machine learning task we are interested in is: "Predict if a given patient (i.e. a new born child) will be discharged from the hospital within 3 days (class 0) or will stay in hospital beyond that (class 1)."

- You need to come up with a logistic regression based approach (or a variant of it), where each element of the system is justified using data analysis, performance analysis and/or knowledge from relevant literature.
- As one of the aims of the assignment is to become familiar with the machine learning paradigm, you should evaluate multiple different models (logistic regression or its variants) to determine which one is most appropriate for this task.
- Setup an evaluation framework, including selecting appropriate performance measures, and determining how to split the data into training and validation.
- Finally you need to analyse the model and the results from your model using appropriate techniques and establish how adequate your model is to perform the task in real world and discuss limitation if there are any (ultimate judgement).
- Predict the result for the test set.

Restriction

As the aim of this assignment is to encourage you to learn to explore different approaches,your must NOT explicitly perform manual feature selection. That is, your models should have all features as input (except the "ID" and "Health Service Area" fields).

Data set

The data set for this assignment can be downloaded here .

There are the following files:

- "README.md : Contain the description of the data set.
- "train_data.csv": Contain the train set, attributes and target for each patient. This data is to be used in developing the models. Use this for your own exploration and evaluation of which approach you think is "best" for this prediction task.
- "test_data.csv": Contain the test set, attributes for each patient. You need to make predictions for this data and submit the prediction via canvas. The teaching team will use this data to evaluate the performance of the model you have developed.
- "s1234567_predictions.csv": Shows the expected format for your predictions on the unseen test data. You should organize your predictions in this format. Any deviation from this format will result on zero marks for the results part. Change the number to your student ID.

The original data is from HealthData: Hospital Inpatient Discharges (SPARCS De-Identified) . The data provided is based on this, with some modifications.

Licence agreement: The data set can only be used for the purpose of this assignment. Sharing or distributing this data or using this data for any other commercial or non-commercial purposes is prohibited.

Marking guidelines

A detailed rubric is attached on canvas. In summary:

- Approach 50%;
- Ultimate Judgment & Analysis 20%;
- Performance on test set (Unseen data) 10%;
- Implementation 20%;

Approach: You are required to use a suitable approach to find a predictive model. You may use any form of logistic regression techniques, including: linear, non-linear and regularization techniques. Each element of the approach need to be justified using data analysis, performance analysis and/or published work in literature. *This assignment isn't just about your code or model, but the thought process behind your work.*

The elements of your approach may include:

- Setting up the evaluation framework
- Selecting models, loss function and optimization procedure.
- Hyper-parameter setting and tuning.
- Identify problem specific issues/properties and solutions.
- Analysing model and outputs.

split the data into train/val/test

~~-?~~ hyperparameter

?

All the elements of your approach should be justified and the justifications should be visible in the PDF version of the notebook (inserted as Markdown text). The justifications you provide may include:

- How you formulate the problem and the evaluation framework.
- Modelling techniques you select and why you selected them.
- Parameter settings and other approaches you have tried.
- Limitation and improvements that are required for real-world implantation.

This will allow us to understand your rationale. We encourage you to explore this problem and not just focus on maximising a single performance metric. By the end of your report, we should be convinced that of your ultimate judgement and that you have considered all reasonable aspects in investigating this problem.

Remember that good analysis provides factual statements, evidence and justifications for conclusions that you draw. A statements such as:

"I did xyz because I felt that it was good"

is not analysis. This is an unjustified opinion. Instead, you should aim for statements such as:

"I did xyz because it is more efficient. It is more efficient because..."

Ultimate Judgement & Analysis: You must make an ultimate judgement of the"best" model that you would use and recommend in a real-world setting for this problem.It is up to you to determine the criteria by which you evaluate your model and determine what is means to be "the best model". You need to provide evidence to support your ultimate judgement and discuss limitation of your approach/ultimate model if there are any in the notebook as Markdown text.

Performance on test set (Unseen data): You must use the model chosen in your ultimate judgement to predict the target for unseen testing data (provided in testdata.csv). Your ultimate prediction will be evaluated, and the performance of all of the ultimate judgements will be published.

Implementation: Your implementation needs to be efficient and understandable by the instructor.Should follow good programming practices.

Support Resources

This assessment requires that you meet RMIT's expectations for academic integrity. More information and advice on how to avoid plagiarism are available in the Getting Started module.

Open the academic integrity page.

Additional library and learning resources are available to help with the assessment in this course

Link to Assignment Support.

Submission Instruction

You have to submit all the relevant material as listed below via Canvas.

1. **The PDF version of the python notebook** used for the model development including critical analysis of your approach and ultimate judgement. This is in PDF format. See canvas for instructions on converting the notebook to PDF.
2. **A set of predictions** from your ultimate judgement. This is in CSV format. If your model predicts the patient will be discharged from the hospital within 3 days, the associated "LengthOfStay" value in CSV should be 0 (1 otherwise).
3. **Your code** (Jupyter notebooks) used to perform your analysis. This is a ZIP file containing all the support files. We strongly recommend you to attach a README file with instructions on how to run your application.

The submission portal on Canvas consists of two sub-pages.

Sub-page 1 (link here) for PDF-Notebook submission

- Please name the report by following this convention: *COSC2753_A1_YourStudentID*

Sub-page 2 (link here) for code and other file submission

- Include only the source code and the set of predictions in a *zip file which follows the above naming convention.*
- Please note that your code will be checked for plagiarism by our specialised software and not by Turnitin.

Rubric

You can review the rubric for this assignment before making a submission to Turnitin.

COSC2753_2021_A1

Criteria	Ratings						Pts
Approach 1) Data exploration leading to well informed approach. 2) Identifying an adequate evaluation framework that is tailored to the problem. 3) Well justified method/algorithms selection. 4) Hyper parameters selection strategy. 5) Through analysis of the models. 6) Approach satisfy all the requirements and restrictions.	**50 to >40.0 pts** **HD+** Outstanding across the course.	**40 to >35.0 pts** **HD** The approach is an excellent and extremely thorough investigation of the chosen ML problem. All elements adequately analysed. Goes beyond what is done in class.	**35 to >30.0 pts** **DI** The approach is a good and reasonably thorough investigation of the chosen ML problem. There are small gaps between in the investigation in what could have been explored. Goes beyond what is done in class.	**30 to >25.0 pts** **CR** The approach is sufficient, but not a thorough investigation of the chosen ML problem. There are gaps in the investigation and alternative algorithms or techniques are better than the ones in the approach. The approach has a limited consideration of the unique aspects of the chosen ML problem.	**25 to >10.0 pts** **PA** The approach is a minimally sufficient investigation of the chosen ML problem. It only examines the bare minimum requirements of suitable techniques and algorithms. There are many gaps in the investigation and there are algorithms or techniques are clearly more suited to the chosen ML problem.	**10 to >0 pts** **NN** Poor, superficial, or incomplete approach that does not meet the minimum requirements for PA.	50 pts
Ultimate Judgement & Analysis 1) Analysis of the model and the outputs using suitable methods. 2) Make a clear ultimate Judgment. 3) Rational behind the ultimate model is clear and considers all the aspects. 4) Limitations of the model identified.	**20 to >16.0 pts** **HD+** Outstanding across the course.	**16 to >14.0 pts** **HD** Ultimate Judgement is established and exceptionally justified. Evaluation of the Ultimate Judgement is exceptional and clearly demonstrated the viability of the trained model in real-world practice and limitations.	**14 to >12.0 pts** **DI** Ultimate Judgement is established and suitably justified. Evaluation of the Ultimate Judgement is sound and suitably explained, however, the reader may not be fully convinced and have minor questions.	**12 to >10.0 pts** **CR** Ultimate Judgement is established, but there are unexplained choices, or the justification is hard to follow. An sufficient attempt at evaluating the Ultimate Judgement is made.	**10 to >4.0 pts** **PA** An Ultimate Judgement is made by not justified.	**4 to >0 pts** **NN** An Ultimate Judgement is not made.	20 pts
Test Performance Performance on the unseen test set	**10 to >8.0 pts** **HD+** Outstanding across the course (Top 5%).	**8 to >7.0 pts** **HD** In Top 5-20%.	**7 to >6.0 pts** **DI** In Top 20-25%.	**6 to >5.0 pts** **CR** In Top 25-50%.	**5 to >2.0 pts** **PA** Better than random model but not in top 50%.	**2 to >0 pts** **NN** Less accurate than a random model.	10 pts
Implementation 1) Code is well documented and easy to understand.2) Code does not contain errors.3) Code contain evidence of all investigations mentioned in report.4) Code is optimal and shows good programming practices.	**20 to >16.0 pts** **HD+** Outstanding across the course.	**16 to >14.0 pts** **HD** Code is exceptional and satisfy all the elements.	**14 to >12.0 pts** **DI** Code is styled and organised reasonably. Commenting could be improved.	**12 to >10.0 pts** **CR** Code is styled and organised reasonably. Commenting could be improved. Few minor errors.	**10 to >4.0 pts** **PA** Code is styled and organised poorly, not following general good programming practices. Commenting is rare. Implementation has minor issues but works.	**4 to >0 pts** **NN** Code is styled and organisedpoorly, not following general good programming practices. Contain major errors.	20 pts

Total Points: 100