RMIT UNIVERSITY

Machine Learning & Computational Machine Learning
COSC 2673 & COSC 2793
Assignment 1

Assessment Type	Individual assignment. Submit online via Canvas → Assignments → Assignment 1. Marks awarded for meeting requirements as closely as possible. Clarifications/updates may be made via announcements/relevant discussion forums.
Due Date	Week 6, Friday 16th April 2021, 11:59pm
Marks	30%

1 Overview

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 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. The notebook should include markdown text explaining the rational, 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.

AWS educate classroom (name: RMIT_Machine_learning_2021S1_Assign1_group$\langle 1, 2, 3 \rangle$) *should be used* for this assignment.

2 Learning Outcomes

This assessment relates to the following course learning outcomes (CLOs):

- **CLO 1:** Understand the fundamental concepts and algorithms of machine learning and applications.
- **CLO 3:** Set up a machine learning configuration, including processing data and performing feature engineering, for a range of applications.
- **CLO 4:** Apply machine learning software and tool-kits for diverse applications.

3 Assessment details

3.1 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 if a patient will be discharged from a hospital early or, will stay in hospital for an extended period (see task below for exact definition), based on several attributes (features) related to: patient characteristics, diagnoses, treatments, services, hospital charges and patients socio-economic background.

The machine learning **task** we are interested in is: "*Predict if a given patient (i.e. newborn child) will be discharged from the hospital within 3 days (class 0) or will stay in hospital beyond that - 4 days or more (class 1)*".

The data set to develop your models is given to you on canvas. Note that you need to transform the target column ("LengthOfStay") to match the two classes mentioned in the above task. Class 0 if LengthOfStay < 4 and class 1 otherwise.

- You need to come up with an ***approach*** (that follows the restrictions in 3.2), 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 (only use techniques taught in class up to week 5 - inclusive) 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.
- 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.

3.2 Restrictions

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 (attributes) as input (except the "ID" and "Health Service Area" fields which are not attributes).

You are only allowed to **use techniques taught in class up to week 5 (inclusive)** for this assignment. That is, you are NOT allowed to use ML techniques such as: Neural networks or SVM for this task.

3.3 Dataset

The data set for this assignment is available on Canvas. There are the following files:

- "`README.md`": Description of dataset.
- "`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 in filename to your student ID.

The description of the data fields are given in Table 1.

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 dataset 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.

4 Submission

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. Should be in PDF format. See canvas for instructions on converting the notebook to PDF.
2. A **set of predictions** from your ultimate judgement. Should be 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. Should be a ZIP file containing all the support files. will be used for plagiarism checking - notebook should match PDF.

The submission portal on canvas consists of three sub-pages. First page for PDF-Notebook submission, the second page for code submission and the third page for submitting predictions on test set (CSV file). More information is provided on canvas. Include only source code in a zip file containing your name. We strongly recommend you to attach a README file with instructions on how to run your application. Make sure that your assignment can run only with the code included in your zip file!

After the due date, you will have 5 days to submit your assignment as a late submission. Late submissions will incur a penalty of 10% per day. After these five days, Canvas will be closed and you will lose ALL the assignment marks.

Assessment declaration:
When you submit work electronically, you agree to the assessment declaration - `https://www.rmit.edu.au/students/student-essentials/assessment-and-exams/assessment/assessment-declaration`

5 Teams

Not relevant. This is an individual assignment.

6 Academic integrity and plagiarism (standard warning)

Academic integrity is about honest presentation of your academic work. It means acknowledging the work of others while developing your own insights, knowledge and ideas. You should take extreme care that you have:

- Acknowledged words, data, diagrams, models, frameworks and/or ideas of others you have quoted (i.e. directly copied), summarised, paraphrased, discussed or mentioned in your assessment through the appropriate referencing methods
- Provided a reference list of the publication details so your reader can locate the source if necessary. This includes material taken from Internet sites. If you do not acknowledge the sources of your material, you may be accused of plagiarism because you have passed off the work and ideas of another person without appropriate referencing, as if they were your own.

RMIT University treats plagiarism as a very serious offence constituting misconduct. Plagiarism covers a variety of inappropriate behaviours, including:

- Failure to properly document a source
- Copyright material from the internet or databases
- Collusion between students

For further information on our policies and procedures, please refer to the following: `https://www.rmit.edu.au/students/student-essentials/rights-and-responsibilities/academic-integrity`.

7 Marking guidelines

A detailed rubric is attached on canvas. In summary:

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

Approach: You are required to use a suitable approach to find a predictive model. You may use any ML technique taught in class during week 2-5, 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.

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 `test_data.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.

Table 1: Data-set Description

Column Name	**Attribute/Target**	**Description**
ID	N/A	Unique number to represent patient ID
HealthServiceArea	N/A	A description of the Health Service Area (HSA) in which the hospital is located. Capital/Adirondack, Central NY, Finger Lakes, Hudson Valley, Long Island, New York City, Southern Tier, Western NY.
Gender	Attribute 1	Patient gender: (M) Male, (F) Female, (U) Unknown.
Race	Attribute 2	Patient race. Black/African American, Multi, Other Race, Unknown, White. Other Race includes Native Americans and Asian/Pacific Islander.
TypeOfAdmission	Attribute 3	A description of the manner in which the patient was admitted to the health care facility: Elective, Emergency, Newborn, Not Available, Trauma, Urgent.
CCSProcedureCode	Attribute 4	AHRQ Clinical Classification Software (CCS) ICD-9 Procedure Category Code
APRSeverityOfIllnessCode	Attribute 5	All Patient Refined Severity of Illness (APR SOI) Description: Minor (1), Moderate (2), Major (3), Extreme (4)
PaymentTypology	Attribute 6	A description of the type of payment for this occurrence.
BirthWeight	Attribute 7	The neonate birth weight in grams; rounded to nearest 100g.
EmergencyDepartmentIndicator	Attribute 8	Emergency Department Indicator is set based on the submitted revenue codes. If the record contained an Emergency Department revenue code of 045X, the indicator is set to "Y", otherwise it will be "N".
AverageCostInCounty	Attribute 9	Average hospitalization Cost In County of the patient
AverageChargesInCounty	Attribute 10	Average medical Charges In County of the patient
AverageCostInFacility	Attribute 11	Average Cost In Facility
AverageChargesInFacility	Attribute 12	Average Charges In Facility
AverageIncomeInZipCode	Attribute 13	Average Income In Zip Code
LengthOfStay	target	The total number of patient days at an acute level and/or other than acute care level. Need to be transformed to match the task class 0 id LengthOfStay <4 and class 1 otherwise