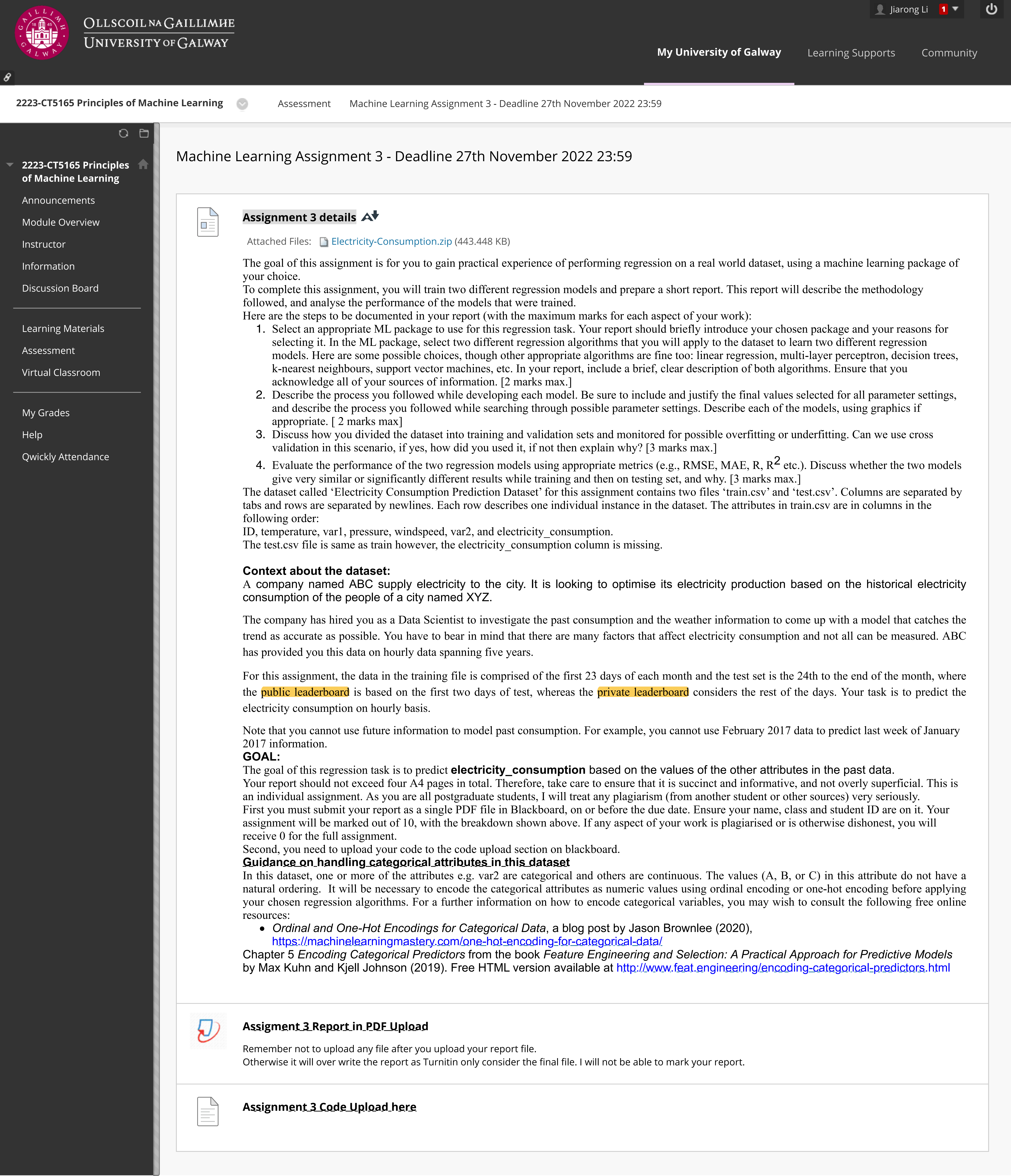

Machine Learning Assignment 3 - Deadline 27th November 2022 23:59

Assignment 3 details

Attached Files: Electricity-Consumption.zip (443.448 KB)

The goal of this assignment is for you to gain practical experience of performing regression on a real world dataset, using a machine learning package of your choice.

To complete this assignment, you will train two different regression models and prepare a short report. This report will describe the methodology followed, and analyse the performance of the models that were trained.

Here are the steps to be documented in your report (with the maximum marks for each aspect of your work):

1. Select an appropriate ML package to use for this regression task. Your report should briefly introduce your chosen package and your reasons for selecting it. In the ML package, select two different regression algorithms that you will apply to the dataset to learn two different regression models. Here are some possible choices, though other appropriate algorithms are fine too: linear regression, multi-layer perceptron, decision trees, k-nearest neighbours, support vector machines, etc. In your report, include a brief, clear description of both algorithms. Ensure that you acknowledge all of your sources of information. [2 marks max.]
2. Describe the process you followed while developing each model. Be sure to include and justify the final values selected for all parameter settings, and describe the process you followed while searching through possible parameter settings. Describe each of the models, using graphics if appropriate. [2 marks max]
3. Discuss how you divided the dataset into training and validation sets and monitored for possible overfitting or underfitting. Can we use cross validation in this scenario, if yes, how did you used it, if not then explain why? [3 marks max.]
4. Evaluate the performance of the two regression models using appropriate metrics (e.g., RMSE, MAE, R, R^2 etc.). Discuss whether the two models give very similar or significantly different results while training and then on testing set, and why. [3 marks max.]

The dataset called 'Electricity Consumption Prediction Dataset' for this assignment contains two files 'train.csv' and 'test.csv'. Columns are separated by tabs and rows are separated by newlines. Each row describes one individual instance in the dataset. The attributes in train.csv are in columns in the following order:

ID, temperature, var1, pressure, windspeed, var2, and electricity_consumption.

The test.csv file is same as train however, the electricity_consumption column is missing.

Context about the dataset:

A company named ABC supply electricity to the city. It is looking to optimise its electricity production based on the historical electricity consumption of the people of a city named XYZ.

The company has hired you as a Data Scientist to investigate the past consumption and the weather information to come up with a model that catches the trend as accurate as possible. You have to bear in mind that there are many factors that affect electricity consumption and not all can be measured. ABC has provided you this data on hourly data spanning five years.

For this assignment, the data in the training file is comprised of the first 23 days of each month and the test set is the 24th to the end of the month, where the public leaderboard is based on the first two days of test, whereas the private leaderboard considers the rest of the days. Your task is to predict the electricity consumption on hourly basis.

Note that you cannot use future information to model past consumption. For example, you cannot use February 2017 data to predict last week of January 2017 information.

GOAL:

The goal of this regression task is to predict **electricity_consumption** based on the values of the other attributes in the past data.

Your report should not exceed four A4 pages in total. Therefore, take care to ensure that it is succinct and informative, and not overly superficial. This is an individual assignment. As you are all postgraduate students, I will treat any plagiarism (from another student or other sources) very seriously.

First you must submit your report as a single PDF file in Blackboard, on or before the due date. Ensure your name, class and student ID are on it. Your assignment will be marked out of 10, with the breakdown shown above. If any aspect of your work is plagiarised or is otherwise dishonest, you will receive 0 for the full assignment.

Second, you need to upload your code to the code upload section on blackboard.

Guidance on handling categorical attributes in this dataset

In this dataset, one or more of the attributes e.g. var2 are categorical and others are continuous. The values (A, B, or C) in this attribute do not have a natural ordering. It will be necessary to encode the categorical attributes as numeric values using ordinal encoding or one-hot encoding before applying your chosen regression algorithms. For a further information on how to encode categorical variables, you may wish to consult the following free online resources:

- *Ordinal and One-Hot Encodings for Categorical Data*, a blog post by Jason Brownlee (2020), https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/

Chapter 5 *Encoding Categorical Predictors* from the book *Feature Engineering and Selection: A Practical Approach for Predictive Models* by Max Kuhn and Kjell Johnson (2019). Free HTML version available at http://www.feat.engineering/encoding-categorical-predictors.html

Assigment 3 Report in PDF Upload

Remember not to upload any file after you upload your report file.

Otherwise it will over write the report as Turnitin only consider the final file. I will not be able to mark your report.

Assignment 3 Code Upload here