Data Mining 1 Assignment 2

Regression and Classification using a bike-sharing dataset

Dr Bernard Butler, SETU

27 October 2023

Issue Date: 27 October 2023

Submission Date: Sunday 19th November 2023 at 21:00

1.1 Aim

The purpose of this assessment is to give students the opportunity to show what they have learned in the module to date, especially relating to Exploratory Data Analysis (EDA), Regression and Classification.

1.2 Format of submissions

Students are asked to submit one (or more) Jupyter notebooks, together with any supporting resources (other than the data files) needed to run the analysis, so that they can be "run" by the lecturer. If a student uses any additional python packages, please indicate where they can be found too. There is no need to add the files to a zip archive before uploading.

Students are reminded to use the markdown cells to explain their reasoning, discuss outputs and report on the progress of their data investigation. Students are also advised to use both tabular and plotted output to present their work. Each notebook should read like a technical report, with all findings supported by evidence.

Please use python comments only to explain how the code works, and use the markdown cells to explain steps in the data analysis. Professional data scientists generally arrange their notebook cells in the following manner

1. Explain what you are about to do and why (markdown)
2. Perform the analysis (python code and comments)
3. Output of the analysis (generated by the code cell above)
4. Interpret and discuss the results (markdown)

repeating this pattern as necessary. Students are strongly encouraged to do the same. Please note that markdown cells can have headings, lists and other formatting, so please use this to make your notebooks and reasoning easier to follow.

Another feature of professional notebook design is that frequently used code is placed in python functions and then reused (with suitable changes to their calling arguments) as necessary, rather than being copied and modified from cell to cell. Marks will be allocated for good practice in this regard because it greatly eases the task of machine learning engineers, who take the notebooks as living specifications and build machine learning pipelines that implement these “specifications” in production systems.

1.3 Submission mode and Deadline

Assignment attempts should be uploaded via the relevant moodle submission page. The deadline for submission is **Sunday 19 November 2023 at 21:00**.

id	Criterion	Marks
1	Read the *hourly* data and split into training and test data	5
2	For the training data only, use exploratory data analysis to learn about the data and to indicate how to build a model	15
3	Using a forward selection approach, build a regression model that offers the best performance	30
4	Which of the 3 target columns is easiest to predict accurately?	5
5	Using this preferred target, derive a new target whose values are the grouped label	5
6	Use *two* classification procedures to predict these demand quartiles	35
7	Does regression or classification provide the highest classification accuracy on the test set? Why?	5

This assessment is worth 30% of your overall marks.

The key to success in this assessment is good data preparation and a structured approach to building models, keeping track of various options and paying close attention to multiple performance measures, most of which are subject to trade-offs, i.e., you cannot improve performance measure *A* indefinitely without harming at least one of the other performance measures.

1.4 Specification

A detailed specification has been prepared in the format of a python notebook. All submissions should follow its format.

For more details, please refer to the attached CA2spec.ipynb notebook.

While the requirements should be clear, there are many ways to meet them, and students are invited to be creative when answering the questions, i.e., marks will be assigned for “flair” and going beyond the treatment given in the notes and labs.

1.5 Other

If you use any resources (such as code, data, analysis) from others, *you must cite the source*. This is easy to do in practice: markdown has support for citing/referencing both online resources and books. Otherwise you are passing off someone else’s work as your own - this is known as *plagiarism* and is not acceptable in SETU.

As with CA1, you are advised to upload your work frequently, but please remember to *submit* the version you wish to be graded.