FIT5149 S2 2020 Assessment 1
Predict Bike-Sharing need in Metropolitan Area

Aug-2020

Marks	15% of all marks for the unit
Due Date	Friday, 18 September 2020, 5:00 PM
Extension	An extension could be granted for circumstances. A special consideration application form must be submitted. Please refer to the university webpage on special consideration.
Lateness	For all assessment items handed in after the official due date, and without an agreed extension, a 10% penalty applies to the student's mark for each day after the due date (including weekends, and public holidays) for up to 5 days. **Assessment items handed in after 5 days will not be considered/marked.**
Authorship	This assignment is **an individual assignment** and the final submission must be identifiable your own work. Breaches of this requirement will result in an assignment not being accepted for assessment and may result in disciplinary action.
Submission	You are required to submit two files, a Jupyter notebook and the PDF file generated by the Jupyter notebook. The two files **must** be submitted via Moodle. Students are required to accept the terms and conditions in the Moodle submission page. **A draft submission won't be marked.** Please carefully read the assessment description. All the members in the teaching team will not accept submissions via email.
Programming language	R in Jupyter Notebook

Introduction

For the metropolitan area, one of the solution to traffic/environmental issues is to enhance the current renting services for bicycles. The benefit of bike sharing is that the sharing system allows people to rent a bike from one location and return it to a different place on an as-needed basis. Sharing program is easier to get access which is normally at a low cost or free of charges, the program is often integrated with mobile devices making it easier to use.

In recent years, bike sharing has been well received all around the world, and leads to significant impacts on establishing a larger cycling community, increasing the use of transportation, minimizing greenhouse gas emissions, enhancing public health and also traffic troubles. It is important to make sure that the right amount of rental bikes are available and accessible at the right place and at the right time, providing a city with a stable and sustainable supply of rental bikes. In order to do so, one important task is to predict the hourly rental bike demand as accurately as possible.

In this task, we are interested in exploring machine learning approaches to predict the demands for bike sharing based on relevant data such as weather, season, holiday, etc, which are known to influence the demands for bike renting. The aim is to build statistical learning models that can predict how many bikes are required in a particular area at a particular time, and identify the major factors that affect the bike demand. Specifically, the problem you are going to solve is:

- Can you accurately **predict** the bike sharing demands given the collected data?
- Can you well **explain** your prediction and the associated findings? For example, identify the key factors that are strongly associated with the response variable, i.e., the demand.

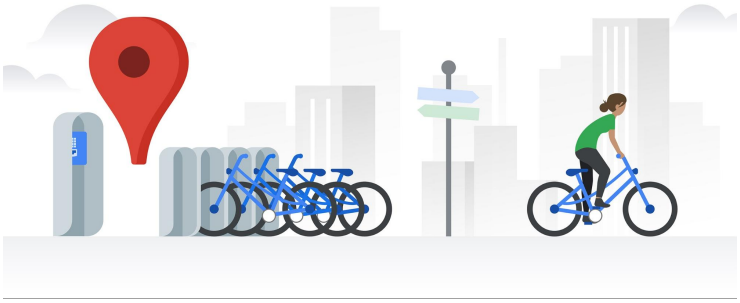

Figure 1: "Bikesharing is booming as this two-wheeled technology transforms how people get from A to B in cities around the planet". The picture is from here

.

Dataset

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information. There are 8760 instances in total with 14 attributes containing 0 missing values. The rented bike count is the target you are required to make your prediction The dataset has been split into training set and testing set to facilitate your analysis, which can be **downloaded from the Moodle site**.

Task description

In this assessment, you will finish the following two tasks.

Prediction task

For the prediction task, the underlying problem is to predict the demands for bikes using the collected attributes. The provided dataset is well organised and cleaned. It is important that you understand each attribute.

To measure the performance of your model(s), you should fit the model to the training dataset, perform the prediction on the test dataset and finally compute some performance metrics of your choice.

In this task, you are required to develop models that can accurately predict the number of bikes required. To finish the task, you should

1. develop and compare 2 types of models;
2. describe and justify the choice of your models;
3. analyze and interpret your results.

Inference task

The purpose of the inference task is to identify the key factors that have strong effects on the bike demands. In other words, which attribute contributes the most to your model's performance? The inference task can be based on variable correlation analysis, regression equations, or any other statistical analysis. The descriptions and the accompanying interpretation must be comprehensible, useful and with statistic support whenever it is possible. To finish this task, you should use proper data analysis techniques to

1. identify a subset of attributes that have a significant impact on the prediction of the bike demands;
2. report your identification with statistical evidence (e.g. correlations, p-values) and interpret the identified attribute subset (e.g. as to why certain attributes have certain impacts on the prediction).

Some Hints:

- Avoid just plainly showing the results without meaningful interpretation/discussion. For example, if you use any plot, you will need to clearly discuss the information delivered by the plots in the context of the task.
- Choose the appropriate plots or statistics to show the right information.
- While developing the model, make clear, for example, how the optimal parameters are chosen if there is any, etc.
- Be precise in the use of various tools and the corresponding discussion. Avoid submitting extremely long Jupyter notebook, which could result in a lot of redundant information, easily losing the focus of your work.

Files to be submitted

There are two files required to be submitted, which are

- The **R** implementation of the tasks in one file.
  - The file **must be** a **Jupyter notebook**. Besides the R code, all the discussions must also be included in the file.
  - The name of the file **must be** in one of the following formats: **XXXXXXXX_FIT5149_Ass1.ipynb** You should replace "XXXXXXXX" with your student ID. **Please keep your running results in the submitted notebook file**.
- A PDF file generated by the Jupyter notebook or R Markdown. The name of the PDF file must be in the following format
  - XXXXXXXX_FIT5149_Ass1.pdf

  Before you generate the PDF file, **please clear all the outputs**. Please note that the PDF file will be used by Turnitin for the purpose of plagiarism check. It is your full responsibility to make sure that all the outputs are cleared before the PDF file is generated, as the outputs can contribute significantly to the Turnitin scores.

Please refer to the Assessment 1's Moodle page for how to submit the two files and note that **If you do not follow the above way to name your submission, your submission will not be marked and will receive 0 mark directly**.

Academic integrity

Please be aware of University's policy on academic integrity. **Monash University takes academic misconduct very seriously. You can learn from the above materials and understand the principle of how the analysis was done. However, you must finish this assessment with your own work.**