QBUS6810 Statistical Learning and Data Mining
Group Assignment

Contents

1 Key information

Kaggle competition ends: 11:59pm 3rd November 2023
Report and notebook due date: 11:59pm 5th November 2023
Weight: 30% of your final grade

Simple extensions: These can only be obtained for the report and notebooks submission. Only one person in the group will need to apply for a simple extension. Please apply by emailing qbus6810.admin@sydney.edu.au.

1.1 Groups

The assignment is to be completed in groups of up to 5 students. Groups can be formed across different tutorials. Please make sure that you have registered your group on Canvas: those groups will be used for identification and assessment purposes.

You are ultimately responsible for forming your own groups. If you would like to be randomly allocated to a group, please contact qbus6810.admin@sydney.edu.au as early as possible. Additionally if you are a small group and would like new members to be randomly allocated to your group, please contact qbus6810.admin@sydney.edu.au. Groups are expected to be finalised by **Wednesday the 18th of October**. Any student not allocated to a group by Wednesday will be randomly allocated to a group on Thursday the 12th of October.

1.2 Group expectations

You are expected to complete the group expectations form provided on Canvas assignment page by Wednesday the 18th of October. You will also need to submit this to Canvas. This is worth 1 mark in your overall assignment.

2 Problem description

Airbnb is a global platform that runs an online marketplace for renting and leasing short-term lodging. It is interested in developing a pricing service for its users that will compute a recommended price based on the features of a listing. As a consultant working for a data analytics company, you are approached by Airbnb to develop a model for predicting nightly prices of Airbnb listings based on state-of-art techniques from statistical learning. The goal of your analytics team is to predict the price per night of listings for properties along the east coast of Australia. Such information can be used to estimate the prices of new listings or to guide new hosts in advertising their properties. Airbnb can also use the information to identify which of their listings produce the most profit.

You are provided with a training dataset containing detailed information on a number of existing Airbnb listings along the east coast of Australia. As part of the contract, you are asked to write a report according to the instructions given in Section 3.2.

3 Datasets

You have been provided with the following datsets, which can be downloaded from Canvas.

- `train.csv`: for training and validating your models.
- `test.csv`: for making predictions.
- `sample_submission.csv`: your predictions must be in the same format as this file. This particular file was generated used the code in the provided Jupyter notebook scaffold.

3.1 Data dictionary

The data correspond to Airbnb listings in Australia with each row corresponding to a single listing.

id	Identifier for each listing to comply with the Kaggle format.
price	The price per night for each listing in Australian Dollars (AUD).
description	Detailed description of the listing.
neighborhood_overview	Host's description of the neighbourhood.
host_acceptance_rate	The rate at which the host accepts booking requests.
neighbourhood	The neighbourhood of the listing.
latitude	The geographic latitude location of the property.
longitude	The geographic longitude location of the property.
property_type	The property type as selected by the host.
room_type	All homes are grouped into the following four rooms: Entire place, private room, shared room and hotel room.*
accommodates	The maximum capacity of the listing.
bedrooms	The number of bedrooms of the listing.
beds	The number of beds in the listing.
amenities	Information on the available amenities at the listing.
minimum_nights	The minimum number of nights you can book the listing for.
maximum_nights	The maximum number of nights you can book the listing for.
number_of_reviews	The number of reviews the listing has.
review_scores_cleanliness	The listing review rating for cleanliness.
review_scores_communication	The listing review rating for communication.
review_scores_location	The listing review rating for location.

* **Entire places** are best if you're seeking a home away from home. With an entire place, you'll have the whole space to yourself. This usually includes a bedroom, a bathroom, a kitchen, and a separate, dedicated entrance. Hosts should note in the description if they'll be on the property or not (ex:"Host occupies first floor of the home"), and provide further details on the listing.
Private rooms are great for when you prefer a little privacy, and still value a local connection. When you book a private room, you'll have your own private room for sleeping and may share some spaces with others. You might need to walk through indoor spaces that another host or guest may occupy to get to your room.
Shared rooms are for when you don't mind sharing a space with others. When you book a shared room, you'll be sleeping in a space that is shared with others and share the entire space with other people. Shared rooms are popular among flexible travelers looking for new friends and budget-friendly stays.
Hotel rooms provide a level of service and hospitality associated with traditional hotels. The rooms are available in boutique or lifestyle hotels, hostels, bed and breakfasts, or similar properties. They typically include vibrant common areas and rooms with unique touches.
Reference: https://www.airbnb.com.au/help/article/5.

3.2 Written report

The purpose of the report is to describe, explain, and justify your solution to the client. You can assume that the client is trained in business analytics, however, is not an expert in statistical learning.

Your report should be a maximum of 15 pages (single spaced, 11pt font). Note that the cover page, a table of context, reference list and appendix do not count towards the page limit.

3.3 Suggested outline of the report

1. Introduction
2. Data processing
3. Exploratory data analysis
4. Feature engineering
5. Methodology
6. Validation and comparisons
7. Conclusion

More detailed information is provided in the report scaffold, which you can download from Canvas. Additionally, a guide for the page length is provided in the marking rubric.

3.4 Requirements

1. Your report must provide the validation scores (those from the Public Leaderboard on Kaggle) for three different sets of predictions, including your final model. These should generally be your best performing models within the model requirements specified below. You will need to make a submission on Kaggle (see Section 5 for instructions) to get each validation score.
2. The three sets of predictions must come from different statistical learning methods. At least one of the models should to be an interpretable linear model (OLS, Lasso, etc); at least one should be an interpretable model specified by a single regression tree; and the remaining model should be more advanced (e.g. using bagging, random forests, boosting or a stacked model).
3. In the methodology section you will discuss your three models in detail (including both the description of the methods/algorithms and the interpretation of the estimated models).
4. You must pay special attention to, and report on, the relationship between the location and the price, both during the exploratory data analysis and during the model interpretation. You must comment on the patterns in pricing around the east coast of Australia. As part of feature engineering, you must create (and describe in the report) at least one new location-related variable by using the existing variables and, if you wish, external information.
5. You are expected to hold at least three group meetings during the course of the assignment (not including the meeting to complete your group expectations). You will need to take meeting minutes as outlined in the appendix of the assignment template.

4 Jupyter notebook

You must provide a Jupyter notebook containing all of the relevant code used to produce the results in your report. The notebook should be well formatted and easy to understand. A notebook scaffold has been provided for you on Canvas.

Once you are ready to submit your notebook, you can used Ed to check that your notebook runs without error.

5 Kaggle competition

You will participate in the Kaggle competition that will be run on www.kaggle.com. This competition will allow you to incorporate feedback into your model building process and compare your performance with that of other groups. Participation in the competition is part of the assessment, so please make sure that your final submission is correct. Your ranking in the competition will affect your mark.

You will need to create a Kaggle account, identifiable by your name, to access the competition and make submissions. Please note that you can significantly simplify your registration with Kaggle by using social logins (Facebook, Yahoo, Google) to sign in. Those options are available on the Kaggle sign-in page. After you have created an account and logged into Kaggle, you should be able to access the competition here (you need to be logged in to get to the competition page via the link). For convenience, this link has also on the Canvas Assignment page.

On this page you will click on the 'Join Competition' link, located in a dark box near the top right corner of the page. After you accept the competition rules, you will have joined the Kaggle competition for the group project. Each group will need to create a team on Kaggle. The group leader can create a team by joining the competition and then going into the 'Team' tab, which will appear near the top of the competition page. The leader can then invite other group members using their Kaggle names (they need to first join the competition before they are able to be invited). Kaggle team composition must be identical to that of the groups you formed on Canvas, and the team number must match the group number. Each student in the group is required to sign up and be identifiable as a member of a Kaggle team.

Kaggle randomly splits (just once) the listings in the test.csv file into validation (50%) and test (50%) cases, but you will not know which ones are which. When you make a submission during the Kaggle competition, you get a score equal to the RMSE computed on the validation listings. These scores are displayed on the 'Public Leaderboard' and provide an ongoing ranking of teams. You can use the scores of your submissions to help you select the best predictive model.

You will need to manually select one of your Kaggle submissions to be used as your final model at the end of the competition. Once the competition is over, Kaggle will rank teams' final submissions based on the test cases only, and those will be displayed on the 'Private Leaderboard'. Your goal is to do as well as possible on the Private Leaderboard at the end of the competition, so please be careful not to overfit the validation cases in an attempt to improve your public ranking. Please note that the competition ends at 11:59pm on the 3rd of November, which is exactly 2 days before the due time for the assignment report.

Rank	Mark
1	10
25	8
50	5
75	3
105	0

Table 1: Examples of rankings in the Kaggle competition and their corresponding awarded mark (out of 10). This table assumes that there are a total of 105 groups, which may change as the group registration is finalised.

5.1 Kaggle marks

Your ranking in the Kaggle competition in the **private leaderboard** will count towards 10% of this assignment. We will look at the rank of your group and your mark will be:

$$\text{mark} = 10 \times \frac{\text{number of groups} - \text{your rank} + 1}{\text{number of groups}} \tag{1}$$

rounded to the nearest half mark. Examples of mark calculations from rank are given in Table 5.1.

5.2 Real world relevance

The ability to perform in a Kaggle competition is highly valued by employers. Some employers go as far as to set up a Kaggle competition just for recruitment.

6 Submission details

6.1 Required submissions

- Written report (one **.pdf** file per group)
- Jupyter notebook (one **.ipynb** notebook per group)

Your report and notebook files should be named:

- QBUS6810_GroupXXX_report.pdf
- QBUS6810_GroupXXX_notebook.ipynb

where XXX is your group number. For example, if you were group 32, this would be *Group032*.

Your assignment should be submitted on Canvas. To find the submission page go to Modules > Group Assignment. You may submit multiple times but only your last submission will be marked.

6.2 Late submissions

In accordance with University policy, these penalties apply when written work is submitted after 11:59pm on the due date:

- Deduction of **5%** of the maximum mark for each calendar day after the due date.
- After ten calendar days late, a mark of zero will be awarded.

7 Academic Integrity

We take academic integrity issues seriously in QBUS6810. If you are suspected of dishonest behaviour you will be referred to the Academic Integrity Office who will process your case. This may result in delayed results, mark reduction, failure of the unit or expulsion.

Please refer to University policy for more details.

8 Expected Contribution to Group Work

It is expected that each member contributes equally to the project.

In the event a student has not sufficiently contributed to the assignment, their mark will be adjusted. This will be judged on a case-by-case basis. If a student does not contribute to the group project at all, they will be given a mark of 0 for the assignment.

Please keep a detailed record of meeting minutes as evidence of each member's contribution. We will also look at the statement of contribution to identify how much each member has contributed to the project. Additionally, please keep other evidence of other group communications (e.g. emails), which we may be provided as proof of student contribution if necessarily.

9 Generative AI

You may use generative AI tools (such at chatGPT) as much as you would like but you must ensure that your group have all agreed on how you want to use it and you must acknowledge its use in your statement of contribution.

Note that you are 100% responsible for your assessment submission. So if you are using generative AI you must ensure that the assignment you submit is of the quality you intend.