QBUS6810
Statistical Learning and Data Mining
Semester 1, 2019

Group Project: Airbnb Pricing Predictions

1. Key information

Required submissions: 1) Written report (submitted as one pdf file per group via Assignment submission on Canvas); 2) Predictions for the test data (via Kaggle); 3) Python code (via email, the address to be provided). Further instructions will be posted on Canvas.

Deadline: Friday May 31st at 5PM.

Weight: 30% of your final grade.

Groups: Complete the assignment in groups of four or five students. Make sure to sign into your group on Canvas; Canvas groups will be used for identification and assessment purposes.

Length: Your written report should have a maximum of **15** pages (single spaced, 11pt, cover page not included).

Marking and key rules:

- A separately posted rubric indicates the marking criteria for the report.
- Carefully read the requirements for each part of the assignment.
- Please follow any further instructions announced on Canvas, particularly for submissions.
- You must use Python for this assignment. It is fine to use Excel for data manipulation (however, this approach is generally not recommended due to its inefficiency).
- The predictions for the test data on Kaggle *must* come from your own analysis in Python. An examination of the code will be conducted for verification purposes.
- Please note that it is your responsibility to be informed of and to follow the University of Sydney and Business School rules and guidelines.

2. Getting the data

The data is posted on the Kaggle competition page. To be able to join the competition, you will need to access the competition page via the following link:

https://www.kaggle.com/t/28cb00294fe94927ac801164794b75dd

You will need to create a Kaggle account, identifiable by your name, to access the competition, download the data and make submissions. After you have created an account and logged into Kaggle, use the above link to get to the competition page (you need to be logged in to get to the competition page via the link). On this page you will need to click on the "Join Competition" link, located in a light blue 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 should 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 teams must be identical to 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.

3. Problem description

Airbnb (www.airbnb.com) is a hospitality company 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 focus of your analytics team is on the properties in London, UK.

You are provided with a dataset containing detailed information on a number of existing Airbnb listings in London. As part of the contract, you are asked to write a report according to the instructions given below. The client will use a test set to evaluate your work.

4. Understanding the data

A training dataset and a test dataset are posted on Kaggle. The latter omits the price values. Furthermore, Kaggle randomly splits the observations in the test set into validation (30%) and test (70%) cases, but you will not know which ones are which.

When you make a submission during the competition, you get a score equal to the RMSE computed on the validation cases. 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 select one of your submissions to be used as final at the end of the competition. Once the competition is over, Kaggle will rank the 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.

Data Description:

- Each row corresponds to a separate Airbnb listing in London, UK. As a consequence of using real data, a detailed description of all the variables is not available. However, the names of the variables are self-explanatory. The first column in the data provides an identifier for each listing and is included to comply with the Kaggle format. It should not be used as a predictor in the analysis. The response variable, price, is the second column in the training dataset. It gives the British pound sterling (GBP) price per night for each listing. Variables security_deposit, cleaning_fee and extra_people are also measured in GPB and correspond to surcharges. Variables latitude and longitude specify the geographic location of each property. Several variables are Boolean, with the word true recorded as "t" and false recorded as "f". Some of the listings have missing values under some of the variables. Note that, in many cases, a missing value means that the corresponding characteristic does not apply to that particular Airbnb listing. This is information, rather than lack of information, and you could make use of this information in your analysis.

5. 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.

Requirements:

Your report must provide the validation (i.e. Public Leaderboard) scores for at least **five** different sets of predictions, including your final model. You need to make a submission on Kaggle to get each validation score. The five sets of predictions should all come from different statistical learning methods.

In the methodology section you will discuss **two** of the five models in detail (the other three do not need to be discussed). One of these two models will be your final model. Also, one of these two models should be an interpretable model (e.g. OLS, subset selection, Lasso, Ridge, Elastic net, a single regression tree), and the second one should be a more advanced model (bagging, random forests, boosting, or a model that contains one of these three as a part).

You will 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. As part of feature engineering, you should create one new location-related variable by using the existing variables and, if you wish, external information.

Suggested outline of the report:

1. Introduction: write a few paragraphs stating the business problem and summarising your final solution and results. Use plain English and avoid technical language as much as possible in this section (it should be for a wide audience).
2. Data processing and exploratory data analysis: provide key information about the data, discuss potential issues, and highlight interesting facts that are useful for the rest of your analysis.
3. Describe and justify your process of feature engineering.
4. Methodology: here you will focus on the two models as outlined above (your rationale for choosing the models and why they make sense for the data, description of how these models are fitted, interpretations of the models in the context of the business problem at hand). This part is allowed to be more technical than the rest of the report.
5. Validation set results from Kaggle and comparison of the methods.
6. Final remarks (non-technical).

6. Kaggle Competition

The purpose of the Kaggle competition is to incorporate feedback by allowing you to compare your performance with that of other groups. Participation in the competition is part of the assessment, and you must make sure that your final submission is correct. Your ranking in the competition will typically not directly affect your marks (apart from thenbonus marks and the Benchmark requirement, as explained below), however, we will assess whether your participation represents a genuine effort to make good predictions and improve them (in particular, you should make sure to beat the "Benchmark" score on the Public Leaderboard).

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.

Bonus marks:

The five teams with the best performance on the Private Leaderboard will receive bonus marks for the assignment (with the total Group Project score capped at 100). The best performing team will receive 10 bonus marks, the second team will get 8 marks, the third will get 6 marks, the fourth and fifth will each get 3 marks (however, the maximum score will remain at or below 100). Please note that your choice of the final model has to be well justified in the report, and the Kaggle predictions must come from your own analysis in Python. An examination of the code will be conducted for verification purposes. Your code is required to reproduce the Kaggle predictions included in the report.