QBUS6810
Statistical Learning and Data Mining
Semester 2, 2019

Project: Airbnb Pricing Analytics

1. Overview

In this project your team will analyse data from Airbnb rentals in Sydney to provide market advice to hosts, real estate investors, and other stakeholders. Your team will have two tasks. The first will be to build a predictive model for vacation rental prices. The second will be to find interesting facts the data that can help your clients make better decisions.

2. Required Submissions

Main Report
Due: Friday November 8th at 5pm
Marks: 30% of final mark
Limit: 15 pages
How: Canvas

Team Expectations Agreement
Due: TBA
Marks: unmarked but required
How: hand in to your tutor

Kaggle Competition
Due: Friday November 8th at 7pm
Marks: part the project
How: Kaggle

Self and Peer Assessment
Due: Friday November 8th at 7pm
Marks: may lead to a mark adjustment
How: SPARKPLUS

Python Code
Due: Friday November 8th at 7pm
How: e-mail (address TBA)

3. Key Rules and Details

Marking: a separately posted rubric indicates the marking criteria for the report.

Originality: the analysis of the dataset must be entirely your own original work. If you borrow material from anywhere based the same or similar dataset (Airbnb rentals), it will be disregarded by the marking even with appropriate referencing (and you know what happens without referencing). This type of dataset (real problem, realistic complexity) provides the best

possible learning experience for you. However, these are hard to come by since companies are understandably not keen to share their data. Therefore, we need strict rules and your cooperation in order to not have to rely on less interesting made up data.

Groups: a separate document will provide the instructions and rules for teamwork (including the team expectations agreement).

Length: Your written report should have a maximum of 15 pages (single spaced, 11pt, cover page, references and appendix not included). Be objective. Find ways to say more with less. Every sentence, table, and figure has to count. When in doubt, delete or move to the appendix.

Python: you must use Python for this assignment.

Kaggle competition: your work should be strictly based only on the training, validation and test data files provided. Violating this rule would be considered a serious breach of academic integrity. The predictions for the test data on Kaggle must come from your own analysis in Python and be consistent with the description in the report. An examination of the code will be conducted for verification purposes.

Announcements: please follow any further instructions announced on Canvas, particularly for submissions.

University rules: please note that it is your responsibility to be informed of and to follow the University of Sydney rules and guidelines, in particular those related to academic integrity.

5. Problem description

Airbnb (www.airbnb.com) is a global platform that runs an online marketplace for short term travel rentals.

As a team of data scientists and business analysts working at a market intelligence and consulting company targeting the Airbnb market, you are tasked with developing an advice service for hosts, property managers, and real estate investors.[1]

To achieve your project's goals, you are provided with a dataset containing detailed information on a number of existing Airbnb listings in Sydney. Your team has two tasks:[2]

1. To develop a predictive model for the daily prices of Airbnb rentals based on state-of-the-art techniques from statistical learning. This model will and allow the company to advise hosts on pricing and to help owners and investors to predict the potential revenue of Airbnb rental (which also depends on the occupancy rate).

2. To obtain at least three insights that can help hosts to make better decisions. What are the best hosts doing?

[1] A real example is Airdna. Airbnb itself has a large data science and analytics team.
[2] This is similar to Airdna: https://www.airdna.co/airbnb-hosts.

We will refer to these tasks as supervised learning and data mining respectively.

As part of the contract, you are asked to write a report according to the instructions given below. The company will use test data to evaluate your work.

6. Understanding the data

6.1 Training, validation, and test sets

The data are split into two files, a training dataset and a second dataset for validation and evaluation. The latter omits the price values.

We will run a Kaggle competition as part of the assignment. Kaggle randomly splits the observations in the second file into validation (50%) and test (50%) 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 model 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.

6.2 Data description

Each row corresponds to a separate Airbnb listing in Sydney. As a consequence of using real data scraped from Airbnb, 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.

The response variable, *price*, is the last column in the training dataset. It gives the price per night for each listing in Australian Dollars. Variables *security_deposit*, *cleaning_fee* and *extra_people* are provided as percentages on the nighly rate. 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".

As with any real dataset, you will encounter many practical issues. The tutorials cannot possibly cover every problem that occurs in practice, so finding solutions is part of the assignment.

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 use it in your analysis.

7. Supervised 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.
- At least one of your models should be an advanced nonparametric model (bagging, random forests, boosting, or a model that contains one of these three as a part).
- You should create text-based features as part of feature engineering.
- Don't waste your time with k-NN (why?).

8. Data Mining

Key question: What are the best hosts doing?

Requirements:

- Extract at least three useful quantitative insights from the data that address the key question.
- It's better to keep the focus on price/revenue.
- Your insights should refer to estimates from a model.
- At least one of your insights should refer to the text fields of the listings (summary, description, etc).
- Remember that association is not causation. Do not oversell your insights.

9. Written report

The purpose of the report is to describe, explain, and justify your solution to the clients. You can assume that the clients have training in business analytics. However, they are not experts in statistical learning and data mining in particular.

Requirement

In the methodology section you will discuss two models in detail (the others do not need to be discussed, just mentioned). One of these two models will be your final model for the Kaggle competition, and the other is your data mining model (typically more interpretable). If your supervised learning and data mining models are the same, then your second model should be an interpretable benchmark.

Suggested outline:

1. Introduction: write a few paragraphs stating the business problem, summarising your final solution, and highlighting your key insights. 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. Due to lack your space, you may want to refer to the appendix for most EDA plots.
3. 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. Predictive model validation.
6. What are the best hosts doing?

10. Kaggle Competition

The link to join the competition will be posted on Canvas.

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.

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 the bonus 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.