QBUS3820
Machine Learning and Data Mining in Business
Semester 1, 2021

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 and the second will be to uncover interesting facts from the data that can help your clients make better decisions.

Please read all the instructions carefully.

2. Required Submissions

Team expectations agreement
Due: Friday April 16th at 11:59pm
Marks: unmarked
How: Canvas assignment

First Kaggle prediction
Due: Friday May 7th at 11:59pm
Marks: unmarked
How: Kaggle

Main Report
Due: Friday June 4th at 11:59pm
Marks: 30% of final mark
Limit: 20 pages (excluding references and Appendix)
How: Canvas assignment

Kaggle Competition
Due: Friday June 4th at 11:59pm
Marks: part of the project
How: Kaggle

Python Code
Due: Friday June 4th at 11:59pm
How: Canvas assignment

Self and Peer Assessment
Due: Friday June 4th at 11:59pm
Marks: can lead to a mark adjustment
How: Sparkplus, link on Canvas

Note: the 11:59pm deadline is based on the University policy determining what would constitute a late submission (see unit outline). An earlier due time would be meaningless under the university rules. I'm not expecting or suggesting that you work until late on the due date.

3. Key Rules and Details

Marking: the rubric (posted separately) 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. 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 protective of their commercial data. Therefore, we need strict rules and your cooperation in order to not have to rely on less interesting artificial datasets in future assignments.

Groups: the groups are self-selected. The assignment must be done in groups of up to three students (minimum of two). There are not exceptions to this rule: if you are more than three then you will need to split the group. Students who do not have a group by the census date will be randomly allocated one, which can be an existing group that is not full or a new one. A separate document will provide further instructions and rules for teamwork (including the team expectations agreement and self and peer assessment).

Length: Your written report should have a maximum of 20 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 material or move to the appendix. That said, there will be no penalties for exceeding the page limit, within reason.

Technology: 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. 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.

Announcements: please follow any further instructions announced on Canvas.

University rules: you agree to follow the University of Sydney rules and guidelines on assignments. The links are on Canvas.

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]

[1] A real example is Airdna. Airbnb itself has a large data science and analytics team.

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 machine learning techniques. 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 rentals (which also depend 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?

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.

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 seconds dataset omits the rental prices.

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**. 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. Because the assignment is based on real data scraped from Airbnb, a detailed description of all the variables is not available. However, the names of the variables should be 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*

[2] This is similar to Airdna: https://www.airdna.co/airbnb-hosts.

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 several practical issues such as uninformative columns. The tutorials cannot possibly cover every problem that occurs in practice, so finding solutions to these problems is part of the assignment and practical training for a real work in data science. Ask us for help if needed.

Some of the listings have missing values for 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.

7. Supervised Learning (Task 1)

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 must be based on models that are substantively different from each other.
- At least one of your models should be a linear model.
- At least one of your models should be an advanced nonparametric model (bagging, random forests, boosting, etc).
- At least one of your models should be a model average or model stack.
- Identify one of your five models as the benchmark.

Suggested:

- Try to build at least some features based on text data.

8. Data Mining (Task 2)

Key question: What are the best hosts doing?

Requirements:

- Extract at least three interesting insights from the data that address the key question.
- The meaning of "best hosts" is for the group to decide upon based on the context of the project. Your clients are hosts and real estate investors, so they are probably interested in maximising their rental income. Therefore, you want to consider outcomes that relate to that such as price and revenue.

Notes:

- This task is open-ended as is the nature of data mining applications. Here you should think creatively and explore the data in a way that you find interesting. The ability to explore open-ended problems is important for industry work in data science.
- Insights that refer to estimates from models (including but definitely not limited to coefficient estimates) tend to be more compelling than insights that are only justified by EDA.
- 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 machine learning and data mining specifically.

Preparing the report will involve careful consideration of what you should include in it. Focus on the highlights of your analysis. Note that there is no page limit for the appendix. It is OK to put extra material in the Appendix and refer to it in the main part of the report.

Requirement:

- In the methodology section you will discuss three models in detail (the others do not need to be discussed, just mentioned). One model is your best linear model, the other your best nonparametric model, and the third is the model stack (or average), all according to your Kaggle validation scores (Public Leaderboard).

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 (you should pitch it to a general 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 possible lack space, you may want to refer to the appendix for most EDA plots.
3. Feature engineering.
4. Methodology: here you will focus on the three 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. Model validation (the Kaggle validation scores should go here).
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 using your student e-mail address to access the competition 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). The name of the Kaggle team must be identical to the group name on Canvas, i.e. 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.

Requirements:

- Every member of the group must sign up to the Kaggle competition and join a team.
- You must be identifiable on Kaggle by your real name or student ID.
- The name of the Kaggle team must be identical to the group name on Canvas, i.e. the team number must match the group number.
- The Kaggle team must be set up and have a valid submission to the Kaggle competition by the date specified in Section 2 (required submissions).

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, explained below) if we judge that your participation represents a genuine effort to make good predictions and compete.

Real world relevance: The ability to participate in a Kaggle competition is highly valued by employers. Some employers in Australia go as far as to set up a Kaggle competition just for recruitment.

Bonus marks: The team with the best performance on the Private Leaderboard will receive 3 bonus marks for the unit. In order to qualify for the bonus, the choice of final model needs to be well justified in the report and your Python code must reproduce the winning predictions.

Attention! You have to manually select which submission Kaggle will use to compute the test (Private Leaderboard) results. It will not necessarily pick the best submission for you (if it did, this wouldn't satisfy the definition of prediction).