QBUS3820
Machine Learning and Data Mining in Business
Semester 1, 2022

Regression 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.

2. 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 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 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?

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

As part of the contract, you are asked to write a report according to the instructions given below.

[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.

3. Understanding the data

Each row corresponds to a separate Airbnb listing in Sydney. Because the dataset was scraped from Airbnb, a detailed description of the variables is not available. However, you can identify their meanings from the context.

The response variable, *price*, is the last column in the training dataset. It gives the price per night for each listing in Australian Dollars. The *latitude* and *longitude* variables specify the geographic location of each property. Some variables are binary, with the word "true" recorded as "t" and "false" recorded as "f".

Since this is a real dataset, you will encounter several practical issues, such as redundant columns and missing values. Overcoming these practical problems is part of the assessment.

4. Machine Learning (Task 1)

Requirements:

- Your report must show results for at least five different sets of predictions.
- At least one of your models should be a linear model.
- At least one of your models should be a tree-based model.
- At least one of your models should be a model average or model stack.
- Identify one of your five models as a benchmark.
- Your report must compare your models in terms of cross-validation or validation metrics.
- Your report must show model evaluation results.

Note that these are only minimum requirements. Refer to the rubric for the details on the marking criteria.

Suggested:

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

5. Data Mining (Task 2)

Business question: What are the best hosts doing?

Requirements:

- Extract at least three quantitative insights from the data that address the business question.
- The meaning of "best hosts" is for the group to decide based on the context of the project. Your clients are hosts and real estate investors, so they'd probably be interested in maximising their property 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 is interesting for you. The ability to explore open-ended problems is important for industry work in data science.
- Remember that association is not causation. Do not oversell your insights.

6. 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, please do not assume that they are experts on the methods used in your project.

Preparing the report will involve careful consideration of what should go in the main text (20 pages). The main text should focus on the highlights of the project. Note that there is no page limit for the appendix. It's ok to put extra material (such as additional figures and tables) in the appendix and refer to it in the main text.

Requirements:

- The report should discuss problem formulation, exploratory data analysis, feature engineering, methodology, and results.
- Write about the data mining task in a separate section.
- In the problem formulation section, discuss the business problem from the perspective of decision theory. In particular, is Airbnb pricing a prediction problem? In what ways can machine learning meaningfully help hosts to increase revenue or reduce costs?
- Discuss three models in detail in the methodology section. One model should be your best linear model, the other your best nonlinear model, and the third is the model stack (or average).
- When you submit the report on Canvas, include the Python code that generates all the results that appear on the report as an additional attachment.

Suggested outline:

1. Introduction: write a few paragraphs introducing the project and overview the methodology and main results. Use plain English and avoid technical language as much as possible in this section (write it for a broad audience).
2. Problem formulation and objectives: state the problem to be solved and the goals of the project.
3. Exploratory data analysis: provide essential information about the data, discuss potential issues and highlight the most interesting findings. Due to a possible lack of space, you may want to refer to the appendix for most EDA plots.
4. Feature engineering.
5. Methodology: focus on the three models specified above. Explain the rationale for using these learning algorithms and explain the choices that you've made regarding configuration, training and hyperparameter optimisation. This part is allowed to be more technical than the rest of the report.
6. Results.
7. What are the best hosts doing?