

MIS772 2020 T1

A1 Predictive Models for Business

AirbnbAI approached you to develop a RapidMiner process of determining if the New York City rental accommodation is viable economically, i.e. when getting more than one review per month:

- The number of reviews per month (n)
- Economic viability $\equiv n > 1$ (extra marks – propose own derived measure)

AirbnbAI provided you with a sample of 49,000 listings of rentals, 20% of which have not received any reviews as yet. The listings include the following information:

- Property id, name, type, price
- Id and name of the property host
- Property geo-location and its neighbourhood classification
- How many nights a dwelling is rented per year
- Minimum nights stay
- How many rooms are being rented in a building
- The number of occupants allowed in a rental
- Whether the listing is licensed or not

AirbnbAI would like you to use RapidMiner to generate some insights into the rental listings and these questions are of interests:

- A) What neighbourhoods have the most *attractive* AirBnB rentals? (define)
- B) What kind of rentals attract the *majority* of reviews? (define)
- C) Which of the rentals, which lack reviews, are still *likely economically viable*?

AirbnbAI wants you to use RapidMiner to cleanup and explore the provided data, then develop and evaluate a classifier to predict the rentals' long-term viability, and to minimise mis-classifications.

This assignment aims for students to learn how to ...

- Articulate problems and solutions in business terms
- Gain insights from data
- Prepare data for different models
- Develop classification models
- Assess and report model performance
- Become curious about the world through data and analytics

The following mini-case study will be used in assignment A1.

Data (in UTF-8): <http://www.deakin.edu.au/~jlcybuls/pred/data/NYC-AirBnB.zip>

Source: <https://www.kaggle.com/dgomonov/new-york-city-airbnb-open-data>

Use file: AB_NYC_2019_Clean_4.csv
and follow in-lab instructions

Individual Tasks and Deliverables

Partial Submission (Questions A, B - marked with the final submission)

Exec Problem: Define your problem in business terms, in doing so answer questions A and B, cross-reference with other report sections for support.

Data Exploration: Visualise the selected attribute characteristics. Use the visualisations to support answering questions A and B.

Final Submission (Question C)

Exec Solution: Describe your solution in business terms, in doing so answer question C, cross-reference with other report sections for support.

Data Preparation: Deal with duplicates, bad and missing values. Transform the selected attributes or create the new ones as needed. Use appropriate analysis and data visualisation to investigate relationships between attributes. Interpret results.

Model: Create and explain one or two classification models, i.e. k-NN and Decision Tree, to address question C. Explain and justify your models' properties. Investigate and deal with the class imbalance.

Evaluation: Use hold-out or cross-validation of the model. Include an honest testing. Compare the performance of different models and select the best. Qualify how much we can trust the answer to question C.

See CloudDeakin for more info about this assignment, especially the assignment template and the assessment rubric. Some students will also be asked to present their work to an academic panel, failure to do so will result in zero marks. Late penalty of 5% per day will apply, up to 5 days, at 11:59pm the submission is already 1 day late.

