Predicting HDB Rental Rates: Finding the Right Flat
CS5228 Final Project – Task Description (v1.0)

AY 2023/2024 Semester 1

1. Overview

1.1. Purpose of the Project

The purpose of the final project is for you to show how you perform data mining in a practical setting. Given a dataset and a defined task, you need to select appropriate techniques to solve this task, justify design and implementation issues, as well as interpret your results and assess any limitations of your approach. We tried to design the project to make the task both interesting and relevant. As often emphasized in the lectures, there is rarely one single best way to solve a data mining task, and many steps will therefore benefit or even require your own creativity to come up with appropriate solutions. While this project will require a certain amount of effort, we also hope that you will have fun completing it, and that it will be a valuable learning experience.

1.2. Project Scenario

In this project, we look into the rental market in Singapore, more specifically the rental market for HDB flats. Accommodation is one of the biggest financial burdens in Singapore. Particularly in the most recent years, both the resale and rental market saw a significant rise in prices and rental rates. Figure 1 shows an example listing for an HDB flat on RentInSingapore.vom.sg. For non-homeowners (e.g., singles, young couples waiting for a BTO flat, or foreign workers with or without family) who do not have the financial means to go with the resale market – or are not eligible – renting is often the only practical solution (even if only temporarily). Thus, potential renters therefore want to know what they can get for their money, where they can best save money, and simply spot bargains and/or rip-offs. On the other hand, landlords and real estate agents aiming to maximize rental rates want to know how to best present and advertise their flats. Also, with the limited amount of available land, affordable housing is a major issue of the Singaporean government, which may deploy ”cooling measures” to influence the housing market. In short, there are many stakeholders that rely and benefit from a deeper understanding of the Singaporean housing market.

More specifically, the main goals of this project is the prediction of rental rates of HDB flats, given the information about an HDB flat (e.g., size, #rooms, location), your task is to predict the expected rent based on historical data. This regression task is implemented as a Kaggle InClass Competition. Note that the focus is less on your final results but on the understanding of the data, task and results, as well as on the motivation and justification of your approach, incl. discussions about potential alternatives and limitations.

In the following, after giving a brief overview to the core dataset as well as to the auxiliary data we have collected, we detail on both subtasks. If you have any questions, please do not hesitate to post your question on Canvas, or send me an email (chris@comp.nus.edu.sg). All the best, good luck, and have fun!

2. The Dataset

2.1. Core Dataset

The core dataset of rental rates for HDB flats has been collected from data.gov.sg. In Singapore, HDB owners need to apply for approval to rent our their flat or single rooms. This information about approved applications is made available on data.gov.sg; of course without any personal information. Each entry in this dataset comes with the

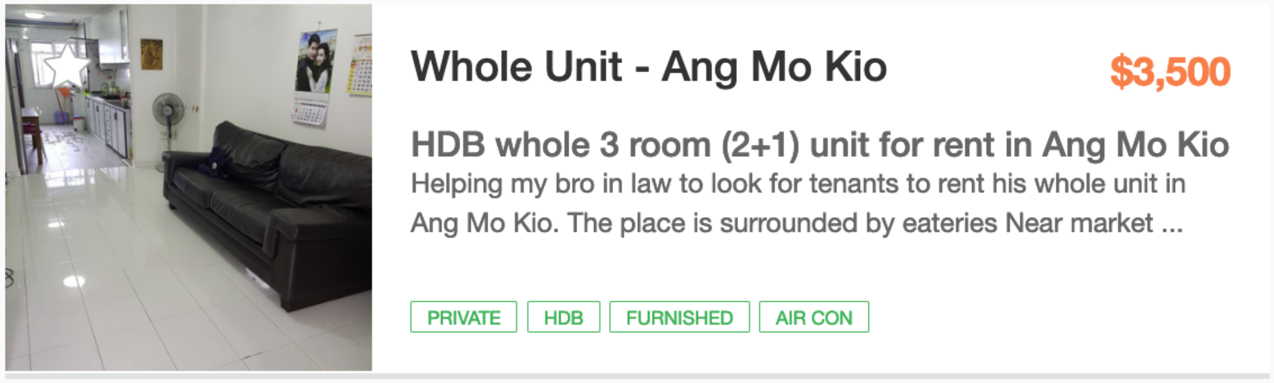
- Rent approval date (e.g., 2022-05, 2023-10)
- Town (e.g., Punggol, Bedok, Jurong West)
- Block (e.g., 650, 321A, 95B)
- Street name (e.g., Punggol Walk, Boon Lay Ave)
- Monthly rent in SGD

The dataset contains all the approved applications from Jan 2021 to Xxx 2023.

Using additional data sources, we have extended each data entry by several additional attributes such as the flat type and size, the lease commence date, as well as further location information (incl. latitude and longitude). Overall, this core dataset consists of around 15 attributes. While the meanings of all attributes should be rather self-explanatory, we do provide a brief description of each attribute on the Kaggle page for our InClass competition. If you still have questions, you can ask your question on the Canvas or via email.

2.2. Auxiliary Data

While a flat’s most basic features (e.g., size and age) certainly affect the rent, other factors are also very important



Whole Unit - Ang Mo Kio

$3,500

HDB whole 3 room (2+1) unit for rent in Ang Mo Kio

Helping my bro in law to look for tenants to rent his whole unit in Ang Mo Kio. The place is surrounded by eateries Near market ...

PRIVATE HDB FURNISHED AIR CON

Figure 1. Example of a listing for an HDB flat for sale on RentInSingapore.vom.sg.

when it comes to assessing the rental rate for a given flat. We therefore provide some additional data to enrich the core dataset described above:

- Location is often touted to be one of the most important aspects of accommodation. Apart from extending the core dataset with exact coordinates, we also collected the location of important ”landmarks” such as MRT stations (both existing and planned stations), shopping malls, and primary schools.
- Rental rates are not decided on in isolation but also depend on the overall ”climate” like the general economic situation. As some indicators reflecting the economy, we collected the stock prices for the largest Singapore-based companies by market capitalization, as well as the Certificate of Entitlement (COE) prices for the same time interval as the core dataset.

2.3. Additional Comments

You are of course not required to use all the data or all the features provided. In fact, it is unlikely that you have enough time to try out all possibilities throughout all tasks. Do not worry, this is on purpose! It is up to you to decide which data and features you deem useful for solving the different tasks. On the other hand, you are also very welcome to collect any additional data that is not provided but you think can help you in the project.

3. Task: Prediction of Rental Rates

The goal of the project is to predict the rental rates for HDB flats in Singapore, as well as to get deeper insights into the local housing market. It is therefore first and foremost a **regression task**. The different information you can extract from the core dataset and the auxiliary data allow you to come up with features for training a regressor. It is part of the project for you to justify, derive and evaluate different features. Besides the prediction outcome in terms of a dollar value, other useful results include the importance of different attributes, the evaluation and comparison of different regression techniques, an error analysis and discussion about limitations and potential extensions, etc.

This task will be implemented as **Kaggle InClass competition**. On the competition page on Kaggle, you can download various files. `train.csv` and `test.csv` split the dataset into the training and test set. Naturally, `training.csv` will contain the numerical attribute `price` for each flat; this column is missing in `test.csv`. The predictions you submit should be via a `csv` file with a single column that contains the predicted rent for each row in the test dataset; we provide the file `example-submission.csv` to show you an example of a submission. To prevent overfitting to the leaderboard, we will limit the number of submissions per day. We also use a 30/70 split for the public and private leaderboard.

The `train.csv` file contains the features of the core datasets. Additionally, you can download the file `auxiliary-data.zip` which contains all `csv` files with the auxiliary data (e.g., the locations of MRT stations, shopping malls, and schools, as well as the stock and COE prices). It is up to you if and how you want to consider and integrate the auxiliary data into the dataset for training your regression model(s).

4. Deliverables

4.1. Progress Report

The progress report will be a simple slide deck as a PDF document of 10 slides max. – to be uploaded to Canvas. The purpose of the progress report is two-fold: (a) to give us a chance to check if your project goes into the right direction, and (b) to provide you with a little incentive to start early. There is no official layout or structure. As the name suggests, it should outline your progress with your project work (e.g., goals and questions, EDA results, first design decisions or results, but also with issues/challenges/obstacles that you are facing). The last 1-2 slides should outline the next steps until the end of the project.

- **Deadline: End of Week xx, Xxx xx (11:59 pm)**

Note: You are welcome to submit your progress report earlier. Ideally, this will in turn give you earlier feedback, but also allow us to better balance the workload. The progress report will not be explicitly graded but not submitting any report will negatively affect your final grade.

4.2. Final Report

The final report will be a PDF document in the format of a scientific paper of at most **8 pages** including tables, plots and figures, but excluding references and the appendix. The appendix may contain supplementary content but should be used sparingly. As a rule of thumb, the report should be readable and completely comprehensible without the appendix. The appendix typically may include plots or tables that elaborate on the results of your EDA or your evaluation. For the layout and presentation in the report, we will provide a Word and LaTeX template.

4.3. Structure & Content

Your final report should include the name and student IDs of all team members as well as your team name. Please also include a breakdown of your workload, i.e., some overview what team member was (mainly) responsible for each part of the project. This can be a table, Gantt chart, etc. to be added to the appendix.

While the overall structure of the report is up to you, it should cover the following aspects:

- **Motivation.** Motivate and outline the goals and questions you address. Note that this is also relevant for Task 1 as different teams may focus on different aspects for those tasks. For example, simply aiming for a top rank on Kaggle for Task 1 is not is not a sufficient motivation :).
- **Exploratory Data Analysis & Preprocessing.** Explain and justify your approach to understand the data, and how it informed your data preprocessing steps (e.g., data reduction, data transformation, outlier removal, feature generation).
- **Data Mining Methods.** Describe how you chose and applied appropriate data mining techniques (e.g., clustering, regression models, recommendation methods – all depending on your approaches for Task 1 & 2). This description should include which techniques you used, how you chose their hyperparameters, etc. Note that you do not need to explain the techniques themselves. However, in case of more advanced methods or models, you should add relevant references.
- **Evaluation & Interpretation.** Evaluate and compare the performance of different methods. Discuss which method(s) performed best and why. Understand in what cases your methods perform bad, and discuss principle limitations and potential future steps for improvement.

The structure of your report should, of course, reflect the 2 subtasks you need to address in this project. While EDA & Preprocessing might be only one section in your report, Data Mining Methods and Evaluation & Interpretation might very likely require their own instances for each subtask. The exact structure will depend on your choice of Task 1, particularly if and how it relates to Task 1.

4.4. Submission

The final submission contains both the report as a PDF document as well as your source code, uploaded to Canvas in a zipped folder. Instead of uploading source code, you can also add a link to a GitHub repository. Note that the reproducibility of your approach is part of the grading (cf. Section 5) which includes the organization, documentation, and readability of your code.

- **Deadline: End of Week xx, Xxx xx (11:59 pm)**

Side note: We generally do not run all your code, simply due to time constraints. However, we do try to run code in cases where something is unclear from reading your report.

5. Grading

In a nutshell, a good grade requires that your approach and all design decisions are well motivated and methodologically sound, and that the outcome – mainly the report but also your source code – is of a high quality. In more detail, we weigh the core criteria for the grading as follows:

Methodological Quality (60%). While the exact distribution may depend on your exact approach, methodological quality generally covers the following aspects:

- **Preprocessing:** appropriate preprocessing methods are chosen (informed by the results of the EDA) and correctly implemented; missing values, categorical attributes, etc. are handled correctly.
- **Visualization:** appropriate plots, figures and tables are used to visualize results, architectures and work flows.
- **Methods:** applied methods are well motivated and correctly implemented; alternatives are discussed and design decisions are justified.
- **Evaluation:** different methods are compared or evaluated using appropriate metrics and experimental setups (e.g., cross-validation); common errors and principle limitations are evaluated and discussed.

Quality of Report (30%). The report describes your methodology and explains your results in a clear, concise and comprehensible manner. Related work should be appropriately referenced; the limit of 10 pages should not be exceeded (excluding references and appendix!).

Reproducibility (10%). The code you submit is complete, well-organized, documented, and readable. Simply put, it should be easy for an outsider to use and understand your code to retrace your steps and reproduce results. This does not only mean to run your code, but also to follow your thoughts when reading the code.

There is no explicit separation between Task 1 & 2 in the grading scheme, as this typically highly depends on the choice for Task 3 – and both tasks often rely on common methods (e.g., preprocessing steps). However, in general, there is more weight on Task 1 as it is more well-defined and specified to allow for a fairer comparison. To give you some rough idea, you can consider a weight 70% for Task 1 and 30% for Task as some guideline.

Important: For the Kaggle InClass competition, your position on the public and private leaderboard will only be used as part of the bigger picture, primarily as part of the methodological quality. Getting a good grade does not require a top position on the leaderboards as long as the overall approach is sound and of high quality. This also means, in turn, that a top position does not automatically guarantee a top grade. Of course, a sound approach and good results typically go hand in hand, and results (significantly) below the average are likely to indicate problems with the methodology. The main purpose of implementing Task 1 as a Kaggle InClass competition is to provide you with incentives for solving this task, to give you a way to compare your solutions with the ones of other teams, and to hand out bragging rights to the top competitors.