”I Need a Place to Settle Down!”
CS5228 Final Project – Task Description (v1.0)

2022/2023 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, but also to provide you with enough flexibility for your approach(es). 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 Singapore housing market. Home ownership in Singapore is rather expensive, making buying a home the most significant financial decision for most people in their lives. Buyers therefore want to know what they can get for their money, where they can best save money, and simply spot bargains rip-offs. On the other hand, sellers and real estate agents aiming to maximize prices want to know how to best present and advertise their properties. 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. To cover various such as aspects and to provide you with some flexibility, the final project is split into 3 subtasks:

- Task 1: Prediction of Property Resale Prices – given the information about a property (e.g., size, #rooms, location), your task is to predict its price based. This regression task is implemented as a Kaggle InClass Competition.
- Task 2: Property Recommendations – given the information about a property a user is interested in, your task is to find ”meaningful” alternatives that can be shown to the users in the form of recommendations.
- Task 3: Open Task – given the provided property dataset, you explore your own ideas to gain interesting insights into the data. There are no specific rules, apart from your insights being non-trivial and useful (cf. our definition of data mining in Lecture 1).

In the following, after giving a brief overview to the core dataset as well as to the auxiliary data we have collected, we detail these three 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 property prices has been collected from 99.co. Figure 1 shows an example listing for a 5-room HDB flat for sale on 99.co, giving an overview over the different features for a listing (e.g., the asking price, property type, address, number of bedrooms, and so on); the dataset includes HDB units, condos, as well as landed properties. The core dataset consists of around 20 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

When buying a property – apart from its basic features – its location is argued to be a very important factor as well. We therefore provide you with additional data to enrich the core property dataset:

- We extended the original attribute of the core dataset by each property’s location in terms of its geo coordinate (i.e., lat/lng), as well as the subzone and planning area the property is located in.
- We collected the location of important ”landmarks” such as MRT/LRT stations, shopping malls, primary and secondary schools.

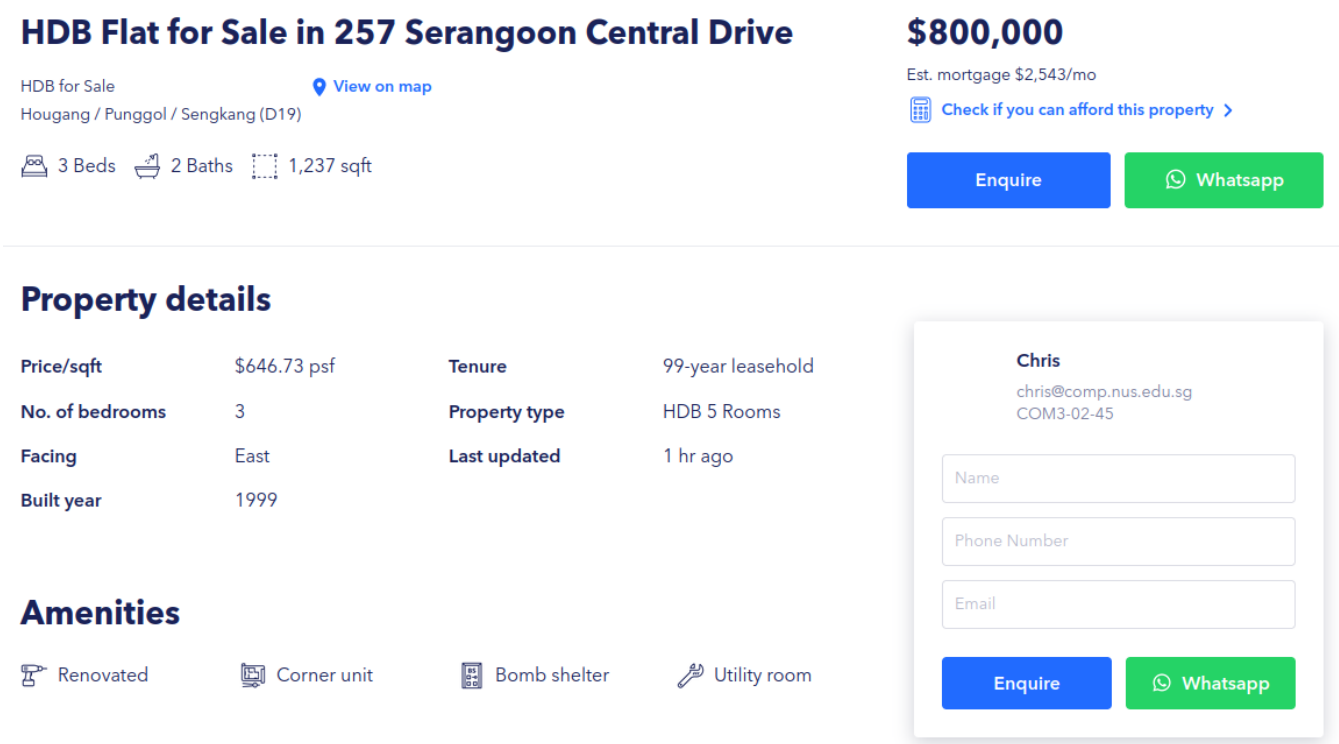

Figure 1. Example of a listing for a HDB unit for sale on 99.co. The dataset will be available on the Kaggle website for the InClass Competition with more details about the different attributes.

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. The 3 Tasks

The following subsections give an overview to the three subtasks and outline the requirements for the final submission. As a general note, when tackling the different subtasks, you are not limited by the methods or algorithms covered in the lectures. You are also free to use any packages (NumPy, pandas, matplotlib, scikit-learn, NetworkX, and beyond) to implement your solutions. In short: All gloves are off, anything goes!

3.1. Task 1: Prediction of Property Prices

The goal of this task is to predict the (asking) prices in Singapore. 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 property; 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 sales price 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 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). 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).

3.2. Task 2: Property Recommendation

When browsing a property listing on 99.co, the site provides alternatives in the "Similar listings" section. However, it is not clear how these recommendations are made nor has the user a way to influence them. The goal of this task is to design and implement your own basic recommendation engine. Similar to Task 1, there are different approaches to solve this. It is therefore important – apart from the implementation of your solution itself – to discuss alternative approaches and justify your design decisions. If you make any important assumptions (e.g., users can specify their preferences, or you utilize a user's browsing history of previous property listings), you should also make those explicit and justify why those are realistic and meaningful assumptions.

Since "good" recommendations are very subjective, Task 2 will require you to submit a self-contained Jupyter notebook incl. any code or data used in this notebook (note that your source code will be part of your submission anyway). The notebook should motivate your approach and important design decisions (like in the final report) and should provide a method `get_top_recommendations()` that computes your, say, top-10 recommendations given a row from the dataset (i.e., given the property listing a user is currently browsing on). Of course, `get_top_recommendations()` may use all kinds of input parameters or any additional information you want to utilize beyond the provided dataset. We will provide more details regarding the submission of your solution for Task 2 towards the end of the project.

3.3. Task 3: Open Task

Tasks 1 & 2 describe two concrete goals to utilize and gain insights into the dataset. Now, in Task 3, we expect you to explore your own goals to extract further non-trivial and useful information from the data. You are welcome to utilize your results from Tasks 1 & 2. There are really limitations other than that the result should be non-trivial and useful. If available and possible to collect, your analysis may require the collection of additional data for the analysis. It is up to you to design your task around available data. Your report should contain a clear motivation for your task of choice and why the results are of practical interest for any stakeholder. If your task computes individual results for a given input (like in Task 2), we also strongly encourage you to organize your solution as a self-contained Jupyter notebook.

We recommend that you address this task last. Firstly, after tackling Task 1 & 2, you will have a very good understanding of the dataset and might even utilize your results for Task 3. And secondly, Task 1 & 2 have a higher weight in the overall grading. Of course, you can always collect ideas for Task 3 when solving the other two tasks.

4. Deliverables

4.1. Progress Report

The progress report will be a simple slide deck as a PDF document of approx. 10-15 slides. 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 8, Oct 09 (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 **10 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 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 & 2 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., regression and classification models, recommendation methods). 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 3 different 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.

4.4. Submission

The final submission contains both the report as PDF document as well as your source code, uploaded to Canvas in a zipped folder. Instead of the 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: In of Week 12, Nov 06 (11:59 pm)**

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.

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.