CS5228 Final Project

2020/2021 Semester 2

1. Overview

The purpose of the final project is for you to show how you perform data mining tasks in a practical setting. Given a dataset and task, you need to select appropriate techniques to solve the task, justify design and implementation issues, as well as interpret your results and assess any limitations of your approach. To provide you with some flexibility for the final project, you can choose between two options:

- **Option A: Kaggle InClass Competition** – you will apply your data mining skills on a given dataset and task (prediction of HDB resale prices).
- **Option B: Open Project** – you explore and area of interest and propose your own task to solve using data mining.

In the following, we detail on both options. If you have any questions, please do not hesitate to post your question on the LumiNUS forum, or send me an email (chris@comp.nus.edu.sg)

2. Option A: Kaggle InClass Competition

2.1. The Task

The resale market of HDB flats is big business in Singapore. To find a good prices as either a buyer or a seller, it is important to have good understanding of what affects the market value of a HDB flat. Most people would accept that attributes such as the size and type of flat, its floor, but also its location to nearby amenities (e.g., MRT stations, parks, malls, commercial centers) influence the resale price of the flat; Figure 1 illustrates this using an online resale ad. However, it is not obvious which attributes are indeed most important in a quantified sense.

The goal of this project is to predict the resale price of a HDB flat based on its properties (e.g., size, #rooms, type, model, location). It is therefore first and foremost a regression task. Besides to 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.

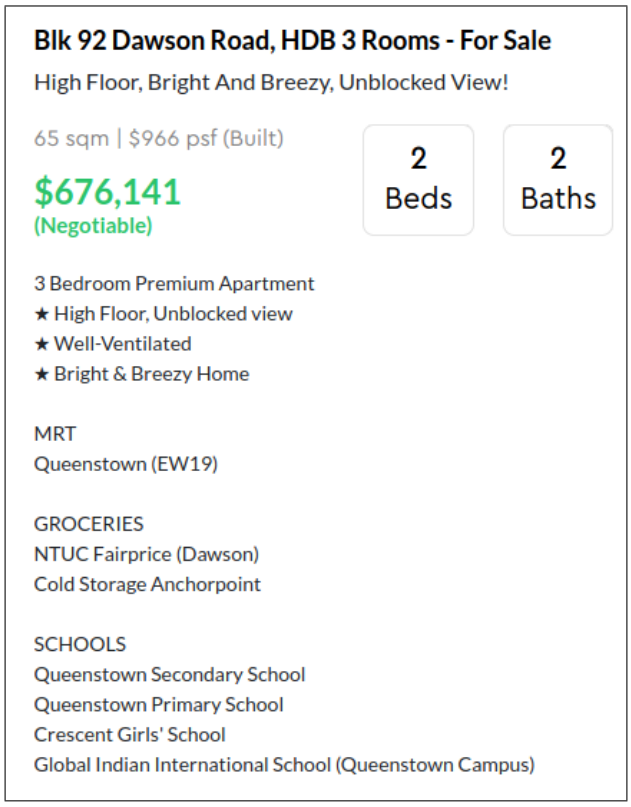

Figure 1. Example of HDB flat resale ad taken from STProperty.sg.

2.2. The Data

The core dataset of past HDB resale transaction is publicly available on Data.gov.sg.[1] However, we customized the dataset for you to allow for a richer analysis:

- The core dataset file already combines all resale transactions – on Data.gov.sg the data is split across 5 files spanning different years. Figure 2 shows small sample of the core dataset
- The core dataset file contains additional attributes such as the subzone or geographic location in latitude and longitude
- The dataset features new supplementary files containing information such as the location of MRT stations, malls commercial centers, and schools.

1. https://data.gov.sg/dataset/resale-flat-prices

month	flat_type	block	street_name	floor_area_sqm	flat_model	lease_commence_date	latitude	longitude	subzone	planning_area	resale_price
2013-02	4 room	4B	boon tiong road	91.0	model a	2005	1.286589	103.831950	tiong bahru	bukit merah	663888.0
1997-08	3 room	11	holland drive	65.0	improved	1975	1.309054	103.794063	holland drive	queenstown	228000.0
2007-08	executive	558	jurong west street 42	157.0	maisonette	1985	1.354198	103.717344	hong kah	jurong west	305000.0
1996-02	4 room	354	hougang avenue 7	105.0	new generation	1986	1.372408	103.899433	kangkar	hougang	235000.0
2015-06	3 room	99	old airport road	56.0	standard	1969	1.308590	103.888245	aljunied	geylang	320000.0

Figure 2. Sample of core dataset containing the attributes and resale prices of HDB flats. The complete datasets will contain supplementary information such as the location of MRT stations, malls commercial centers, and schools that can be used to derive additional useful features.

We believe that the meanings of all attributes are rather self-explanatory. If you still have questions, you can check the description of original dataset on Data.gov.sg, or ask your question on the forum or in an email.

These different types of information allows you come up with new features for training a regressor. It is part of the project for you to justify, derive and evaluate different features. Lastly, you are also welcome to collect additional data that might improve the accuracy of your predictions.

2.3. Kaggle Submission

On the Kaggle page for the InClass competition, you can download two files, `train.csv` and `test.csv`, that split the dataset into the training and test set. Naturally, `training.csv` will contain the numerical attribute `resale_pice`; this column is missing in `test.csv`. The predictions you submit should by `csv` file with a single column that contains the predicted resale price for each row in the test dataset. To prevent overfitting to the leaderboard, you are allowed to make 3 submissions per day to the leaderboard.

3. Option B: Open Project

The open project allows you to explore and area of your interest in the form a data mining task. For example, if you are currently working in a particular industry area, or doing research within a particular academic area, it is highly advisable to pursue a topic in those areas. This could take the form of performing some data analysis on a dataset from your domain of interest, or proposing a new method relevant to data mining on a particular type of data.

3.1. Project Proposal

If you choose the open project, You will have to submit a 1-page proposal shortly motivating the task, describing data, outline important research questions expected results; we will provide a basic template for that. This proposal is not graded, and is purely to allow the course staff to help you ensure your project topic has reasonable scope and feasibility, and possibly offer some helpful advice. If needed, you can still change your project topic after the deadline, but please let the course staff know first (so they can help to ensure that the new topic and new time frame is reasonable). Also feel free to approach the course staff if you would like feedback on your project topic before the deadline.

3.2. Things to Consider

If you are looking into Option B, here are a couple of points you should consider before making your final decision.

- **Scope.** The project should be in line with the course aims of CS5228. This includes the emphasis on understanding the data, the pros and cons of different data mining methods for a given task, and the meaningful interpretation of the results. This is generally very difficult in case of unstructured data such as images and videos. Just having a good deep learning model for, say, image classification does not guarantee a good grade.
- **Dataset availability.** Make sure that the dataset for the task you have in mind is indeed available and clean enough for a meaningful analysis. This particularly includes the case where you first need to collect the dataset on your own (e.g., using public APIs[2]).
- **Computing power.** Although we try not encourage it, you task of choice will require the processing and analysis of large volumes of unstructured data such as images and videos. This typically requires large amounts of processing power. You should make sure that processing time does not become a bottleneck for your project.
- **Team size.** We encourage a team size of 3 students. This assumes that the whole team has to agree on the project topic. Teams of 4 students are possible if the scope and ambition of the project justifies it, as motivated in the project proposal. But please talk to the course staff first.

2. Note: We cannot consider data collection methods such as Web scraping as this is a legal gray area.

3.3. Useful Resources

Repositories of Public Datasets. Below you can find links to a selected set of repositories of publicly available dataset. There are many more such repositories, particularly for very domain-specific datasets, to can find with your Web search engine of choice.

- Google Dataset Search: https://datasetsearch.research.google.com/
- Kaggle public datasets: https://www.kaggle.com/datasets
- UCI Machine Learning Repository: http://archive.ics.uci.edu/ml/index.php
- Github page with links to public datasets: https://github.com/awesomedata/awesome-public-datasets
- Subreddit of answers: https://www.reddit.com/r/datasets/
- Singapore open datasets: https://data.gov.sg/

Data collection using public APIs. In principle, you can also collect your own dataset using public APIs. For example, in the context of Singapore. LTA and NEA provide public APIs to collect live information about bus arrival times, carpark availability, estimated travel times taxi availability etc. in case of LTA, and temperature, rainfall, UV index, etc in case of NEA. Here some resources where you can browse for public APIs:

- ProgrammableWeb: https://www.programmableweb.com/
- RapidApi: https://rapidapi.com/collection/list-of-free-apis
- APIs.guru: https://apis.guru/browse-apis/
- AnyAPI: https://any-api.com/
- GitHub page with links to public APIs: https://github.com/public-apis/public-apis

If you are considering to collect your own data, here are some things to consider:

- Data collection might take take, particularly if an API provides only live but not historical data, and you need a sufficient amount of data for a meaningful analysis (e.g., several weeks and months)
- Make sure that the API is indeed public and free to use; this is not always very obvious until you are indeed trying to access the API.
- Most open APIs have rate limits, i.e., you can only submit a limited number of requests within a time interval. Make sure that such rate limits will not cause a bottleneck for your data collection.

In short, relying on self-collected data will involve some risks when it comes to achieving your project goals in the allocated time frame. So please be aware of that.

4. Deliverables

4.1. Progress Report

The progress report will be a simple slide deck as PDF document of approx. 10-20 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 better start early than too late. 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 for the progress report: TBD

4.2. Final Report

The final report will be a PDF document 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 readable and completely comprehensible without the appendix. 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 Kaggle team name for Option A. Please also include a breakdown of your workload, i.e., some overview what team member was (mainly) responsible for each parts 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 the Kaggle Competition project as different teams may focus on different aspects of the task (beyond the regression task).
- **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). 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 model(s) perform bad, and discuss principle limitations and potential future steps for improvement.

4.4. Submission

The final submission contains both the report as PDF document as well as your source code, uploaded to LumiNUS in a zipped folder. Instead of the source, 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 and readability of your code.

- Deadline for the progress report: TBD

5. Grading

To keep it fair, the grading for the Kaggle competition and the open project will be based on the same criteria. In a nutshell, a good grade requires that your approach is 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 chosen project and 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 decision are justified.
- **Evaluation:** different methods are compared or evaluated in 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 ad readable. Simply put, it should be easy for an outsider to use and understand your code to retrace your steps and reproduce results.

In case of Kaggle InClass Competitions – i.e., for Option A or you chose to enter a public competition for Option B – your position on the public and private leaderboard will only be used as part the bigger, primarily as part of the methodological quality. Getting a good grade does not require a top position on the leaderboard as long as the overall approach is sound of and high quality. Of course, a sound approach and good results typically go hand in hand, and results (significantly) below the average indicate problems with the methodology.

6. FAQ

Please use this Google Doc to submit any immediate questions and comments about the project you have: https://docs.google.com/document/d/1IfYAuZPMc1Wpv-w2n5XbuZrvzkjF-AoVz2-OoJ4BRFU/edit?usp=sharing

Based on your feedback there, we will extend this FAQ section and publish an updated version of this project description.