DSE3101 Project

AY23/24 Semester 2

Contents

1 Project guidelines

The project of this class is meant to provide a context for practicing the following skills:

- Integration of knowledge and datasets to solve the problem at hand.
- Self-learning of new topics as needed.
- Project and timeline management.
- Communication and collaboration within a team and between teams.
- Solving a problem for the user, not merely to our satisfaction.

1.1 Project group

Each group consists of up to 4 members. The members will be divided into two sub-groups of two. One sub-group will handle the modeling, while the other team focuses on the front-end.

- Please indicate your project group on Canvas by **Wednesday of Week 8 (March 13).**
- A group can either be *open* (anyone can join) or *by invitation* (specific classmates only). You can follow the Canvas instructions here: https://wiki.nus.edu.sg/pages/viewpage.action?pageId=381919310

1.2 Deliverables and submissions

Each group will submit:

- **A project proposal (5%).** Choose a topic and write up your proposal. Note that during your project work, you can make changes to the original proposal as you encounter challenges. This is also part of real-life project work. If this happens, you should describe it in the project journal: What difficulties did you encountered that prevent you from realizing the original proposal, and how you worked around them in the end.
- **Two 10-minute videos (20%).** Each sub-group will be responsible for one video. For instance, the modeling team will present the model performance, while the frond-end team will present details on the interaction. The videos should be directed at the target audience.
- **A project journal (10%).** An overview of the development of the project. The purpose is for us to understand, from your perspective, how the problem was approached, the division of labor, and the difficulties that were overcome as a group and as an individual.
- **A technical documentation (10%):** Each group will compile a technical manual of the work done. Both sub-group should include a brief description of the source code on GitHub, and then dive into the code portions relevant to their sub-group. The report can include code snippets.

Additionally, each student will submit:

- **Peer review (5%):** Each individual student will review videos from other teams and submit a short survey to rate them. You may also post suggestions/admiration on each video.

Key activities and due dates:

	Activities	Due dates
Recess & Week 7	Read group project instructions	
Week 8	Submit group info on Canvas	March 13
	Submit project proposal on Canvas	March 15
Week 12	Submit short videos on Canvas	April 12
Week 13	In-class presentation	
	Peer review	
	Submit everything else on Canvas	April 19

2 Project descriptions

2.1 Public transportation in Singapore

The main data for this project can be obtained from LTA DataMall: https://datamall.lta.gov.sg/content/datamall/en.html

- Bus and train ridership at different location across Singapore at particular time/day of the week.
- Location of bus stops and trains stations.

Other possible data can be sourced, including the planning region boundaries in Singapore (https://beta.data.gov.sg/collections/1749/view) as well as the population size and residents' mode of transport (https://www.singstat.gov.sg/).

2.1.1 Problem statement

The aim of this project is to identify commuter hubs within Singapore's public transportation network and the accessibility of public transport across regions in Singapore. This can include examining the density and distribution of transport services in relation to population centers.

Focus area for analysis can include:

- Public transport availability: Analyze data to uncover the current availability and accessibility of public transport in Singapore. Suppose the LTA wants a ranking of locations in Singapore by accessibility by public transport. Propose a measure that could form a basis for such an accessibility score, e.g., higher score for being close to an MRT station vs. a bus stop etc. Another focus in this area could be on measuring location accessibility relative to private transport. One concern of the LTA is that the time disparity relative to car travel gets very large for long distances.
- Data visualization: Create an interactive tool to provide uses with critical insights. These can include
  - The flow of passenger by region and specific bus/train routes.
  - The connectivity index of a region based on its transport links (e.g., the number of direct bus routes, frequency of buses per hour).

2.1.2 Possible use case

1. Alec is a daily commuter in Singapore who relies on public transport for daily work commute. She is exploring her options for buying/renting a new home. With insights from this project, she can check the current passenger flow and bus frequencies – this can help her decide the best time to leave home and avoid crowded conditions. Additionally, this tool enables her to understand public transport connectivity in different areas – this can help her gauge how her daily commute would be affected if she were to move to certain neighborhoods.

2. A government agency is planning to implement decentralization strategies to enhance connectivity between workers and their workplaces. They are looking to understand accessibility between residential areas and industrial and commercial clusters in order to refine planning efforts.

2.2 Understanding the HDB resale market in Singapore

The potential sources of data for this project could be:

- https://beta.data.gov.sg/ – search for "resale flat prices", several datasets would cover different years. There are other potentially relevant series in the "HDB" dataset category.
- https://www.onemap.gov.sg/ - potential source of information on amenities and distances.
- https://www.singstat.gov.sg/find-data/search-by-theme?type=publications – General Household Survey and Census of Population, available for 2010, 2015, 2020, could have useful information.
- Planning region boundaries in Singapore: https://beta.data.gov.sg/collections/1749/view

It may be helpful to consult some studies dealing with HDB resale prices or private condominium prices in Singapore for ideas on modeling methods and relevant variables.

2.2.1 Problem statement

The aim of this project is to visualize and find the best approaches to predicting HDB resale prices using relevant information. This can include descriptive visualizations that relate key market statistics such as prices and number of units transacted to environmental (e.g., amenities, transportation availability, distance from CBD) and demographic characteristics (e.g., population age makeup, income) as well as building predictive models to predict HDB prices as well as investigate the performance of these models over time.

Focus area for analysis can include:

- Data visualization: Create an interactive tool to provide users with critical insights. These can include
  - Town analysis by housing availability/affordability, geographic and socioeconomic characteristics. Availability of several years of data allows for illustrating dynamics of these characteristics.
  - Flat-level analysis to show which ones may have the highest chance of breaching some affordability red-line (for examples on definitions of affordable flats, see the speech by Minister of National Development (MND)).
  - Flat-level analysis to show the price difference between transactions of the same unit type to highlight areas of rapid price growth.
  - Town-level analysis of the different amount of wealth that can be unlocked by downsizing, lease buyback scheme, etc.

- Predictive price modeling:
  - Calculating and visualizing the performance of predictive models in for HDB resale prices across Singapore towns over time. Identifying dynamics in pricing performance and key predictor variables over time.

2.2.2 Possible use case

1. A traditional approach for investigating housing prices is hedonic modeling – relating the natural log of housing prices to structural characteristics of the property and environmental characteristics, such as nearby amenities, transport etc. Such models are typically cast in a linear regression form, called "hedonic regressions". The limitation of such models is that they are linear in parameters and, unless extensive feature engineering is used, focus mainly on linear relationship between the characteristics. Utilizing machine learning methods that can account for nonlinearities is, therefore, an interesting avenue of investigation for various stakeholders in understanding the key predictors of house prices. Additionally, the residual (prediction error) of such models can be thought of as land value according to urban economic theory – once the key characteristics of the apartments and nearby amenities are accounted for, the remaining unexplained price variation can be thought of as a proxy for the "value of a planning area". Obtaining good estimates of such land value and monitoring its evolution through time could yield insights into the dynamics of the perceived land attractiveness and expectations on land use.
2. A government agency is concerned whether today and future retirees have enough savings to tide them through their silver years. Housing is a substantial store of wealth for many, and there are schemes in place to allow retirees to 'cash out' to boost their savings, eg by downsizing or by the Lease Buyback Scheme. The agency is interested to know how much additional savings can be unlocked by different groups of retirees - minimally by age cohort and by estate.

2.3 Benchmarking macroeconomic forecasting models with real-time data

The main source of data for this project is: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research/real-time-data-set-for-macroeconomists

There are other potentially relevant data on the Philadelphia Fed real-time macroeconomic data webpage: https://www.philadelphiafed.org/surveys-and-data/real-time-data-research

Other relevant data may be sourced (e.g., financial data), but pay careful attention to vintages and availability in time, if applicable.

2.3.1 Problem statement

The aim of this project is to evaluate macroeconomic forecasting models in a realistic setting that properly accounts for data revisions. Frequently, academic and private forecasters run benchmarking (or back testing) exercises and demonstrate results that indicate superior forecasting

performance over older studies or professional forecasters' performance. However, such studies are usually done with the most recent version of the data. Macroeconomic data is by nature aggregated from multiple sources and thus is often subject to revisions. In fact, there could be multiple revisions and it could take years to arrive at the final numbers. Professional forecasters have to deal with the currently announced numbers, which suffer from a lot of uncertainty. The studies that follow frequently "beat" the results of professional forecasters all too easily, which could be merely the effect of having access to much cleaner data. You will choose a macroeconomic variable that is subject to frequent data revisions (e.g., US GDP growth rate), build time series models to forecast it at multiple horizons of practical interest, and evaluate its performance using real-time data vintages. Students are strongly recommended to go through at least the nontechnical background reading from the real-time dataset for macroeconomists titled "A Funny Thing Happened on the Way to the Data Bank" to get a better understanding of the practical problems posed by data revisions.

Focus areas for analysis can include:

- Building a set of time series forecasting models for one of the key macroeconomic indicators subject to frequent revisions and determining the best model at each forecast horizon via pseudo-out-of-sample performance.
- Given the above, building an interactive tool that shows, for user-specified training and predicted time periods (e.g., one may be interested how the model fares in booms vs. recessions or in special episodes such as the Great Financial Crisis or the Covid-19 pandemic), the forecasts generated by the chosen model, the underlying forecast uncertainty if possible (i.e., forecast intervals), and summary forecast performance measures (such as the MSE, turning points correctly predicted etc. depending on the variable used).
- Showcasing differences in forecast performance and ranking of models when using real-time vs. latest vintage data.

2.3.2 Possible use case

1. A central bank is evaluating different methodologies for macroeconomic forecasting to be put in production. As the forecasting model would be updating the forecasts as new data arrives, it is crucial to understand model performance in this environment rather than an "artificial" setting where the latest available vintage is used throughout the back testing exercise.
2. A researcher is peer-reviewing an article about macroeconomic forecasting, where the authors show much better forecasting performance than in the past studies using relatively simple models. The paper is using the most recent vintage, and the peer-reviewer would like to have an idea to what extent revisions affect the results.

2.4 Understanding the prospective 2024 DSE job market in Singapore

The main source of data for this project is:

Students would need to obtain (web-scrape) relevant job descriptions from LinkedIn and/or other suitable job sites. Some other sources could be utilized (e.g., salaries that are not published can be proxied from Glassdoor).

2.4.1 Problem statement

The aim of this project is to understand the state of the market for junior positions that are a good fit for DSE program graduates. Multiple stakeholders in the major have similar questions: what are the likely positions one can work in after graduating from the DSE program? What industry will one end up in? How much one is likely to earn? What are the key hard and soft skillsets required in these roles? There are no good answers to these questions to date. The challenge is that we cannot simply comb the "data scientist" ads, because such positions may pertain to different domains (e.g., a DSE graduate is unlikely to be employed by a company using data science in image recognition applications or robotics). At the same time, for many positions that require both data science and economic domain knowledge skills the title is quite different from "data scientist" or "data anything". For example, relevant roles could be: business consultant, financial analyst, statistical officer, research assistant, risk analyst, economist etc. In general, it is probably best to focus on roles that imply "an economist with data skills" rather than "a CS person with a little economics background". Also, the focus is on entry-level positions (something like 0-2 years of experience required) – obviously, nobody will become VP of Data Science fresh out of school.

Focus area for analysis can include:

- Building a dataset of job titles with relevant features extracted: organization or institution, skills or degrees required, expected salary, industry, public or private, additional requirements such as internships or years of experience and any other features you find relevant.
- Given the above, building an interactive tool that provides critical insights on the junior job market for DSE graduates using visualization techniques and potentially unsupervised learning methods.

2.4.2 Possible use case

1. The steering committee of the DSE program is wondering whether the curriculum of the program aligns well with industry requirements. They would like to investigate the nature of skills required for roles that are likely a good fit for DSE graduates.
2. Agnes is an 18-year-old female considering studying data science or economics at a local university. At the CHS open house, she is intrigued to hear about the new integrated Data Science and Economics program. However, while the career prospects and average salary ranges are more or less clear for Economics and Data Science and Analytics majors, she is keen to know more about employment prospects yielded specifically by DSE.