DSE1101 Project Guide

Aims and scope.

The main goal of this project is to take a chance to get a good feel for what it means to apply the techniques learned in this course to analyzing and building predictive models using real world data. You will be able to choose from four provided datasets for analysis, which range from "classical" economic data to more business/finance-oriented applications. You will be expected to give your best effort to provide as deep as possible analysis using tools acquired in class rather than push the boundaries of research (there will be plenty of time for that later in your studies/career).

Therefore, the tentative project workflow should go along the following lines:

1) Determine the dataset that interests you most: will you try to explain worldwide long-run economic growth, Singapore HDB prices, or dive into banking applications? – different datasets may require different approaches as, for example, the variable of interest could be binary or continuous.

2) Examine available data for the variable you are going to predict and any other variables you may use as potential predictors. At the initial stage, you may want to check that there are no missing values, obvious outliers (e.g., some variables may have impossible values) etc. Read carefully the data description files included with the datasets – they contain some crucial information that will help with your data processing. You may also need to recode some variables in a suitable format for analysis (e.g., text variables into factors). Split the data randomly into a training and test set (set seed for replicability!). Perform the rest of the analysis on the training set only – the test set will be invoked only in the final stage for evaluation.

3) Think about what could be the interesting descriptive statistics to report given the limited space – perhaps summaries of variables by certain categories, box plots, or correlation plots could be useful depending on the context. You may need to experiment with quite a few combinations here before you find something interesting but remember – report only the most interesting and relevant findings: your report

space is limited and in the real-life scenario the audience is only interested in graphs and tables that tell it something – do not include every single descriptive statistic output you have produced. Remember that there is no "model" answer here – different students may uncover different interesting facts about data through basic statistical analysis. Also, remember that the dataset may contain many variables, but not all may be equally useful – it is fine to focus more on variables that you believe are relevant based on your knowledge of economics, intuition, and obtained results.

4) Think about what predictive modeling methods learned in class are appropriate in your case and deploy all applicable methods to build a predictive model for the outcome variable. For the benchmark model, you should build either linear regression or a logit model with the most relevant predictors. You may find it easier to gradually work on this part week-by-week after the midterm exam, implementing methods one by one as they are covered in class. In your report, note briefly the details on model validation, the values of the key tuning parameters chosen, and any other details pertinent to model building.

5) Evaluate the performance of the models built in (4) on the test set and report nicely presented evaluations for comparison. Take care to utilize the relevant evaluation metrics based on the problem type (classification or regression). Discuss your results – do you think your model works well according to your results? Will the model likely be useful in a real-world scenario? How do more "advanced" machine learning methods perform compared to "basic" benchmarks: linear regression and logistic regression – is it worth it to go beyond baseline methods?

6) Discuss any potential limitations of your results: there are always limitations to any quantitative analysis and it is important to be able to recognize them. It could be insufficient data or the absence of data on a particular predictor, a violation of assumption(s) that prevents some method from working well, or the absence of a particular tool or technical skill that could have helped you in your view.

Producing a report. You can think of this report as providing a short research report or an executive summary to an academic or business supervisor. That means your report should have a particular structure (following closely the project workflow) and

not be contaminated by academically inappropriate language. View this as an opportunity to practice writing concise but informative research summaries. The structure of such a report could be as follows.

First, you have a short introduction, which should motivate what you are doing and explain why it is interesting to the reader. In the end, give a brief summary of your results and explain how the rest of the report is organized.

Afterward, it is customary to discuss the data. Mention the source, any transformations/cleaning you applied to the data, the train/test set split, and what are the respective sample sizes.

The data description can then be followed by the presentation of key descriptive statistics and visualizations just to give a taste of what the data is like and to potentially spot outliers/things that don't make sense, obvious correlations, potential trends, etc. Again, resist the temptation to report descriptive statistics in bulk and be more selective in reporting only the most interesting/relevant findings.

Finally, you may start on the main body of your work – reporting and discussing the results of your analysis. Try to connect your analysis to your research question and to the general economic knowledge that you have acquired so far. If there are things that seem out of place – focus attention on them and try to think why it happens. In short, try to make your discussion a bit deeper than "we can see that the line in Figure 1 is upward sloping". Label your figures and tables. Add legends and write notes if necessary so that the reader can understand what's going on in the figure/table without reading the text preceding it (it is customary to explain the variables, abbreviations, cases considered, etc. in the note to the figure/table). Compare and contrast performance across methods – e.g., if in the lecture we said one method is usually better performing than the other, do you observe this in practice? Why might this happen?

In the end, write some concluding remarks. Summarize what you have done and whether the results were as one might have expected. Highlight the interesting points you encountered along the way. Remember – a negative result is also a valuable result

that tells us that something doesn't work under certain circumstances, so don't despair if you find that the performance of some or all methods was not as expected. This is also a good place to comment on the limitations of your study (there always are) and possible enhancements for future work – it is an important indicator of how well you understand what you can and cannot do with the tools and data you have.

Project requirements.

The project report is not to exceed 4 pages in 11pt font with 1.5 line spacing, including any diagrams and tables. The report will be **due in Week 13 (tentatively Friday, but the exact deadline will be confirmed through Canvas based on student feedback).** Note that the 4-page limit is strict. If you find that you are breaching the limit, focus on the findings that most important for bringing your point across and omit the less pertinent material. The projects will be submitted via Canvas. The soft copy of the report is to be submitted together with all the data and code in a single zip file to ensure the replicability of the results. The code should be as clear and as reasonably detailed as possible (e.g., if you manipulate data inside R, all the variable transformations, etc. should be documented in the code), and should run given the raw data file provided (we recommend first import the raw data provided into R and save that as the R workspace, then your code can start from loading this workspace and doing all the data cleaning and subsequent analysis). Note that the submitted code should only include the operations that generate results reported in the project in the sequence they are reported – you may have generated more results in the course of your work that are not as interesting/relevant, but please keep them out of the replication code. Take care to set seeds for random number generators when splitting the sample randomly, performing cross-validation, etc. Verify that running your code produces exactly the same results as reported in your project text.

If you have any further queries, please do not hesitate to contact us regarding those.

Comments on project datasets.

You can choose to analyze one of the 4 suggested datasets for this project: 1) the economic growth dataset; 2) 2021 HDB resale prices; 3) credit card defaults; 4) bank telemarketing data. Below are some comments that should provide some context and help you in getting started.

1) **Economic growth.** This is one of the classical datasets in economic research, looking at explaining long run economic growth in a cross-section of countries. This particular version of the dataset comes from the Sala-i-Martin et al. (2004) paper published in American Economic Review. This dataset is compiled in an attempt to answer an important question: what are the key drivers of long-term economic growth? This is also a good example that data science does not always mean dealing with Big Data: since the number of countries in the world is more or less fixed, so is the sample size in such applications. The outcome variable here is the average GDP-per-capita growth rate between 1960-1996, so it is a regression problem.

   The paper is included together with the data mainly for you to be able to read the descriptions of variables (you should take a look at Section II Data and Table 1 there); you do not need to read about the methodology implemented there. Instead, you should build predictive models using methods that you learned in class. In this project, besides the usual evaluation of predictive performance, you may want to include some discussion about what variables appear to be important in driving long-run growth.

2) **HDB resale prices (a sample of transactions from 2021).** The property market in Singapore is a hot topic in conversations across ages and occupations. This dataset will allow you to delve a bit deeper into it and try to understand factors that may be important for pricing HDB flats on the secondary market. The data is taken from open sources (such as data.gov.sg) and most of the data cleaning has already been done – our thanks go to Nathanael Lam Zhao Dian who has graciously allowed us to use a subsample of his Honours Thesis dataset. The outcome variable is the HDB resale price. Since HDB prices are all large numbers, it may be smart to normalize them –

you may want to divide them by 1000 to have more manageable numbers to deal with when running regression models and reporting results. The dataset contains many predictors, but clearly, some are much more important (such as the area in square meters) and others maybe not be so much. You do not need to analyze all predictors necessarily – rather, you may opt to select the "core" few predictors and then try to test some conjectures and see whether certain perks such as closeness to MRT, amenities, and schools are as important as one might think in explaining HDB prices. A fun exercise that could be done at the end of the project – look up a couple of current listings for HDBs and predict their price using your best model: are the listed prices "fair"? If your prediction differs substantially from the listed price, discuss what may have contributed to the error in this case.

3) **Credit card default.** Identifying customers who are very likely to default on their credit card payments is a pertinent task for banks worldwide. This dataset from Yeh et al. (2009) contains data on credit card defaults in Taiwan in 2005. The outcome variable is whether a customer defaults on his/her credit card payment, so this is a binary classification problem. You will find out whether using only basic background information can help forecast credit card defaults. After discussing the customary evaluation metrics, you may want to discuss whether there are any surprising results regarding the variables that turned out to be important.

4) **Bank telemarketing.** When you graduate and start working, you will inevitably get calls from telemarketers trying to set you up with an insurance plan, a new mobile/fiber internet contract, or a financial product. How successful are those calls and what determines whether a customer "bites"? This is an important question for a business to answer before it engages in a direct marketing campaign. This dataset from Moro et al. (2014) collects data on a telemarketing campaign by a Portuguese bank offering term deposit subscriptions. The outcome variable is whether a customer subscribed to the offered financial product eventually, so it is again a binary classification problem.

Finally, let us reiterate: do not worry too much about “obtaining the right answer”, because there really is no ultimate answer in this case. The main objective is to utilize **the full potential** of the techniques you learn in this class in analyzing a dataset comprehensively, practicing the corresponding thought process, and organizing and reporting your findings. It is entirely possible for students working on the same dataset to obtain different results due to a multitude of factors but still get a good grade because of the correct logic behind the approach.