**STAT 425** **Statistical Modeling I** **Fall 2021**

**FINAL PROJECT**

**Due Date: December 13th 2021, Noon**

# 1 Introduction

This file describes the aims and structure of the final project for STAT425. You need to write a final report showing all your analysis and interpretation of results (maximum 15 pages, including front page, figures, tables and appendix (if any); fewer pages is better). Please note that this project is a group project. **Group projects may have up to three members maximum. Please include an additional title page with the names of the group project members as your project report cover, and a brief description of which parts of the project were developed by each member**.

# 2 Data Description

**Data set: Real estate valuation data**
This data set can be found at the UCI Machine Learning Repository[1]. It consists of market historical data set of real estate valuation collected from Sindian Dist., New Taipei City, Taiwan. More information about the data sources can be found in this link. The relevant publication for this data set is: Yeh, I. C., and Hsu, T. K. (2018). *Building real estate valuation models with comparative approach through case-based reasoning.* Applied Soft Computing, 65, 260-271. It is highly recommended that you read the corresponding publication to get familiar with the data and with the authors' analysis.

The inputs are as follows:

- $X1$=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.) Note that from variable $X1$ you can extract the transaction month.
- $X2$=the house age (unit: year)
- $X3$=the distance to the nearest MRT station (unit: meters)
- $X4$=the number of convenience stores in the living circle on foot (integer)
- $X5$=the geographic coordinate, latitude. (unit: degree)
- $X6$=the geographic coordinate, longitude. (unit: degree)
- $Y$= house price of unit area (10000 New Taiwan Dollar/Ping, where Ping is a local unit, 1 Ping = 3.3 squared meters)

You need to consider 7 predictors for your analysis: $X1$ to $X6$ and $X7$=the transaction month.

[1] https://archive.ics.uci.edu/ml/about.html

## 3 What do you need to include in your report?

- Section 1: Introduction: Provide a brief introduction of the goal of this final project. What is it all about? Where did you get the data from? What is the data background information?
- Section 2: Exploratory Data Analysis: Include some graphical displays of the data and numerical summaries of the data. Which variables and categorical? Which variables are quantitative? Also comment on any patterns/characteristics of the data which you find interesting like unusual/missing observations or anything relevant to your later analysis.
- Section 3: Methods: You are required to build at least two prediction models. For each method, include of a description of the methodology, and a description of the implementation if the implementation is not trivial. The implementation has to do with the programming approach you are using.
  - Section 3.1: Start with a simple model (a model that doesn't require much training), for example a multiple linear regression model. Include all necessary steps for variable selection, model diagnostics and model checking.
  - Section 3.2: Make predictions with your linear regression model using a testing set, and report your prediction errors.
  - Section 3.3: Predict with a different kind of regression model covered in this class, like a non-parametric regression or regression splines (just to mention a few). Compare the model fitted in 3.1 with this second model on a testing set, and compare both prediction errors.
  - Section 3.4 (Optional): You can use other methods studied in other classes for your analysis as a comparison. For example, a Random Forest approach (not covered in this class).
- Section 4: Discussion of results and Conclusions: In this part you should make a summary of your results and discuss the impact of your analysis. You should also write the main conclusions in three to four bullet points.

## 4 What do you need to submit in CANVAS?

Submit the following on the CANVAS final project link by **midnight, December 12th, 2021:**

- Final report (in .pdf), max. 15 pages including appendices and the cover page.
- R code (in R markdown)
- Summarize your numerical results using tables/figures. **PLEASE Do not include your R output in your report**. Assume that the reader does not have any idea of the R language!. Therefore, including $R$ code in the final report might be confusing.
- All the figures and tables (if there are any) must be labeled, and you should comment on the results displayed there in the main text. Do not include figures or tables that will not be necessary to support your statements.

- Add comment lines in your R script so its easy for us (me and the TA) to follow, e.g., "# Generate figure 1 in Sec 2", "# Model I: linear regression with the following variables ..... ".

## 5 General Rules and Academic Integrity

- You are NOT allowed to discuss the final project with any other groups/individuals outside your working group. If you have questions, please email me or post your question on the discussion board corresponding to the Final Project.
- You are allowed to use online resources. It will be a good idea to have an "Acknowledgment" Section at the end of your report where you acknowledge the author (or authors) of the online resources.
- You are NOT allowed to copy any sentences from other' work (paper, blog, or his/her post on the Forum) verbatim to your report. You have to either paraphrase or cite the source. Check some online websites on "how to avoid plagiarism".