STAT 542, Homework 6

April 22, 2018
Due date: May 11 (Fri), 5 pm to Compass

Requirements: This homework is mini-project for analyzing the "Melbourne housing market" data from Kaggle:

https://www.kaggle.com/anthonypino/melbourne-housing-market/data

Please read this carefully! You should submit your **report and `R` code** as separate files to Compass2g. The report should be in PDF/html/MS Word format, with no more than 20 pages, including plots. Your report should focus on describing the analysis strategy, demonstrating and interpreting the main result. No excessive `R` code should be included in the report. The `R` code file should include all corresponding code for generating the report. Make sure that you have **necessary comments** in your code such that our grader can easily replicate your results by re-running it. Font size should be 12pt and plots need to be clearly labeled. If you use R markdown and need a template, there is an example source file provided on our course website. This homework worth 100 points total. Late submission penalty is 5 points for each day (round up) of delay. **No submission will be accepted after May 13th, 5pm.**

Question 1: [30 points, 5-page limit] Data processing and descriptive analysis. The goal of this analysis is to predict the `Price` variable, hence any missing value of this variable can be discarded from the data (you can still use them if you want). Summarize each variable in the data set and provide simple descriptive statistics using tables/plots. You do not have to use the original scale of the variables. Any transformations, such as log, square root etc are allowed. The variables come in different forms (text, date, categorical, continuous, etc), you should provide a detailed discussion on how to transform them into a format that can be analyzed by regression models.

Question 2: [20 points, 4-page limit] The goal is to cluster the houses into categories. You can use any variable except the `Price` variable to define the clusters. The clusters you found should have intuitive interpretations such that it might help in the predictive model for `Price`. You can use any clustering algorithm. Clearly describe your approach, demonstrate the findings and interpret your results.

Question 3: [40 points, 9-page limit] Perform regression models to predict the `Price` variable. You could consider transformations of this variable, but the prediction needs to be done on the original scale. Your analysis should at least include the following models:

- Use two penalized linear regression models with variable selection property. For example, Lasso, Ridge, elastic net, AIC, BIC, etc.
- Use two nonparametric models. For example, random forests, splines, boosting, nearest neighbors, SIR, etc.

For each model, you need to 1) discuss and tuning the parameters (if any) and select the best tuning; 2) discuss the advantages and disadvantages compared with other models you considered; 3) properly demonstrate your findings, such as estimation errors, variable selection, etc.

Question 3: [10 points, 2-page limit] Based on all of the analyses above, can you suggest a new approach or a modification/integration of the approaches to improve the performance? Before implementing this new approach, you should particularly discuss

- Intuitively, why the model is likely to improve upon the existing models?
- Is there any computational challenges and if there is, how to address them?

Now, implement this method, and summarize your findings. Note that your new approach does not necessarily have to improve the accuracy. If there are computational difficulties that you cannot overcome, then consider a smaller subset of the data, such as one subregion of the city.