STAT 425 **Applied Regression and Design** **Spring 2020**

FINAL EXAM

Due Date: May 15th 2020, midnight

1 Description

This exam is a take-home project. Write your answers as a report (maximum 15 pages, including figures and tables; fewer pages is better).

Data set: Hotel booking demand

This data set contains booking information for a city hotel and a resort hotel, and includes information such as when the booking was made, length of stay, the number of adults, children, and/or babies, among other things. More information about the data can be found in Kaggle. I have modified the data set to analyze visitors whose origin is from the US only; and have excluded some variables from the original data. The modified file (stat425_fpdata.csv) has been uploaded in Compass and has 19 variables. There are two types of hotels (hotel variable): **Resort Hotel** and **City Hotel**. **Section 01 will analyze the data segment corresponding to *Resort Hotel*. Section 02 will analyze the data segment corresponding to *City Hotel*.**

2 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?
- Section 2: Exploratory Data Analysis: Include some graphical displays of the data. Also comment on any patterns/characteristics of the data which you find interesting or anything relevant to your later analysis. Provide a brief explanation/summary of variables you plan to include in your analysis.
  - Which variables are categorical and which are numerical?
  - Should we keep all time related variables in our analysis?
  - Should we keep week or month in our analysis?
  - For categorical variables, should we include any interactions?
  - For numerical variables, any evidence supporting nonlinear trends?
- Section 3: Method: 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.
  - Section 3.1: Start with a simple model, a model that doesn't require much training.
  - Section 3.2: Predict with linear regression models.
  - Section 3.3: Predict with Randomforest.

  - Section 3.4 (Optional): You can also try some other methods.

- Section 4: Discussion of results

3 What do you need to submit on Compass?

Submit the following on Compass (Assignment Dropbox) by midnight, May 15th, 2020:

- report (in pdf),
- R code (in .R or .txt, or .html or .pdf from R markdown)
- Summarize your numerical results using tables/figures instead of listing R output in your report.
- 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.
- Add comment lines in your R script so it's 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 ..... ". It maybe a good idea to prepare your R script using R markdown.

4 General Rules

- You are NOT allowed to discuss the exam with anyone else. If you have questions, please email me or post your question on the discussion board.
- You are allowed to use online resources. A good place to start would be the Forum section and Script section on Kaggle. 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".