# BUS212f: Analyzing Big Data II Fall 2018

## Case 2: Boston Airbnb Listings (b)-- Improving Your Models

### Introduction

For Case #1, you only knew how to create basic trees. Since that time, we have (or will have) studied bagging, boosting, random forests and k-nearest neighbor classifiers, and have example scripts to implement the techniques. We also held discussions in class about dealing with missing data and transforming available data. We also know a little more about evaluating model performance.

In this assignment, your team will revisit the first case and see if the alternative approaches can improve our classification accuracy. You will re-use the same “Main” and “New” csv files as in Case #1.

In addition, you now also know how to create an R project and to use GitHub for collaboration and version control. In this assignment, you will take advantage of this new knowledge. For the convenience of the teams, I will create separate private repos through GitHub Classroom as soon as everyone completes Comp2. You’ll do your work within these shared Team Repos.

One other minor change: for this report, imagine that you are preparing the analysis for Andrea J. Campbell, President of the Boston City Council President. This past spring, the City Council passed new regulations limiting short-term rentals through Airbnb. For our assignment, we’ll assume that the Council continues to monitor the Airbnb market closely, and is interested in developing a reliable model to identify Commercial hosts.

### Data Preparation:

You have the option of using the same data preparation steps that you used in Case 1, or decide to make additional changes. You may also merge in (JOIN) data from other sources if you see fit.

NOTE: As discussed in class and privately with some teams, do *not* create a variable counting the number of listings each host has.

As in the prior assignment, please split the data randomly into training and test sets (70% - 30%) before conducting the analysis. Use set.seed(*integer of your choice*) to permit replication.

### Your Challenge:

Using the observations from the main listings data, build a model that uses some or all the available measurement variables to classify listings as Commercial (Yes/No). You should initially train several models, one each using theses 5 approaches, and then choose ONE models as your best.

1. a “best-pruned tree” (with cross-validation), {this might be the exact one you used in Case 1, or you might introduce other variables)
2. a boosted tree model,
3. a bagged tree model,
4. a random forest model,
5. K-nearest neighbors model. For this model, use just the few best predictors as discovered from the tree-based methods.

Evaluate the models using the techniques covered in chapter 5, 7, and 9; your report should present the performance evaluations and explain why you have chosen the model you selected. In addition, use your analysis to identify the three or four predictors that provide the most predictive value.

### Deliverable:

Your team should prepare a Word document created with R Notebook (markdown) to report on your analysis and discussing your conclusions. In addition, prepare a PowerPoint presentation suitable for President Campbell and the City Council to explain your analysis and whether you think this model can improve upon the traditional human expert classification.

In all you will upload two files to LATTE: one from the knitted notebook, and one from PowerPoint. I will check your team repository to find the RMD file itself, so no need to upload that.

#### Your slides:

* should be clearly written, non-technical and specific. Remember that your audience is politicians, not data scientists or IBS students.
* you should not discuss all five models, but do explain the one model you have chosen and explain your reasoning.

#### The R Markdown document must include:

* Relevant R code that you used. NOTE: no doubt you explored models that you quickly rejected and/or wrote code that failed with error messages. BE SELECTIVE in reporting your code: show the code that was actually informative for you. Use comments and prose extensively to explain the code as well as the output.
  + Remember that you can use r markdown options like ```{r, echo = FALSE, results = “hide”} etc. to control what appears in the resulting Word document.
  + In RStudio, consult Help > Cheatsheets > R Markdown for details.
* Selected graphs, tables, or statistics summarizing the models you investigated and supporting your final choice. You should include confusion matrices and a comparison of misclassification rates.
* Some output that identifies which features (predictor) variables were most informative in the various tree models.
* A few sentences addressing this question: which of the five methods were helpful to you in deciding which variables to include in your model formulation, and why?