Final Project: CMPT 455

Project Due: 24 hours before your presentation (to be scheduled but should be last week of class)

Project Submission: Data, Slides, and code (either R or Jupiter Notebook in Python)

Group Size: Groups of two. If there is an odd number of students I will allow a single group of size 3.

Grade weight: 20% of final grade

Assignment (Major) Tasks:

1. Select a data set. It should have the following characteristics, but the most important characteristics are that you are interested in the data and think you can build an interesting project from it. If any of the following characteristics are not met, just clear it with me first.
   1. At least 10 variables
   2. At least 2 factor variables (create them with buckets if necessary)
   3. At least 3 quantitative variables.
   4. At least 500 records
2. Choose a dependent variable from the data set, that you will predict from models that take the other variables as input.
3. Give your project context by coming up with a backstory, in the form of a memo that hires you as a contractor. What is the business case for this modelling? Why is prediction of the dependent variable important? Should you prioritize predictability or explainability or both? The better backstory you give the better you will be able to justify subjective decisions you make so don’t skimp on this task. I won’t know the backstory until you present, so if you backstory evolves over the course of your project that is fine.
4. Univariate and Bivariate analysis: discuss the features of your variables as we have in class. My suggestion is to try and maximize the reader’s (mine) interest. That is, focus on variables that are important, and features that are unique (good or bad) and some transformations you may need to make.
5. Try and build at least one interaction: test a few (at least 3) and select at least one interaction that you can include in your modelling.
6. Transform variables and make / select features. This should be done to fit the needs of the problem you are solving. Don’t force this, but it is almost always necessary so be sure to discuss with me first if you think you don’t need this step, or risk losing marks.
7. Build models. Make sure to use Forward/Backward selection with AIC or BIC and justify you choices. You should also apply Regularization and justify ridge, Lasso, or elastinet.:
   1. At least 1 classification model (if your dependent variable is numerical you will need to bucket it for this task)
   2. At least 1 multi-linear regression model
   3. At least 1 GLM
   4. At least one unsupervised learning model (More on this when we cover it later.)
   5. At least 1 artificial neural network.
8. Compare the models using the techniques we discussed in class. Give context to the comparison. Don’t just report numeric results.
9. Select a model based on your comparison and relate it back to your backstory/business case.
10. Write a short memo to your employer that explains how the model works and how well you expect it to function. This task should be in NON-TECHICAL language that your employer can understand.