BUS 212a S1 – Analyzing Big Data II, Spring 2019

Project Assignment 3: Classification

The purpose of this assignment is to have you gain experience and skills with R, focusing on Classification. You will describe and discover structure in your project dataset, train and validate your models, and interpret the best model in a professional quality report. For each significant step, document your results, interpretations, assumptions, and process.

Select a target categorical variable to model with Logistic Regression, K-NN, and Classification Trees. Create and try to include at least one interaction or polynomial term, i.e., higher-order term, with all the methods. You may create indicator variables (dummy variables) or perform mathematical transformations, e.g., LN(x) or sqrt(x), to make your predictors more Gaussian.

For Logistic Regression, remember:

- Check for outliers.
- Each predictor should have a p-value statistical significance less than 0.05.
- Interpretation: For each predictor, look at the exp(B) column, i.e., the log odds. Compare the value against 1.0. Less than 1.0 means a decrease of 1.0-exp(B) percent. Greater than 1.0 means an increase of exp(B)-1.0 percent vs. the baseline category.

To show that each model is not overfitting, you will need to show that it works on your validation data (subset) and show that the sensitivity, specificity, and other metrics are still good.

Compare and contrast the confusion matrices of the best models from the different methods. Do they reinforce or contradict each other? For example, do the rules of your classification tree match your choice of K in K-NN and your interpretation of the Logistic Regression coefficients? Which model do you consider the best, and why?

Deliverable: zip file containing 1) Rmd file, 2) final report, 3) datasets or links to them on cloud storage if your datasets are too big for upload to LATTE.