Data Mining Principles

Assignment 4: Part 1

Use the Diabetes Readmissions data (on Canvas)

Logistic Regression

1. Generate training and Test samples for the Diabetes data set, of sizes 70% and 30%. Call them, say, Train and Test. **SAVE Train and Test. You will need these samples for comparison for other models you will build in later classes**.
2. Build a Logistic Regression model for Train for predicting the "Readmitted" variable, using the other variables. You should combine the < 30 days and > 30 days values to "Yes". Try to build the model with the best test accuracy that you can achieve.
3. Choose only the "main-effects". Main-effects are simply the original variables "as is" (of course, you need to ensure that categorical variables are represented with (k-1) dummy variables that have values 0 or 1. Try various independent variable combinations until you achieve the best test accuracy. [You can use GridSearchCV or RandomizedSearchCV in Python along with cross-validation to select the best solutions. RStudio users can use the stepAIC function from package MASS]. Choose this as the best model, and present its summary of the model. **You don't have to use any multiplicative interactions or non-linear transformations of the variables for this exercise. Save the best solution.**
4. Generate the confusion matrix (counts and proportions of actual versus predicted using the best model from 4.
5. Summarize the results.

Gradings: 1 Point per question. Total 5 points.