things to do

for each diagnostic: use PCA to check if there are specific patterns for people readmitted < 30, and people who were not readmitted or people readmitted > 30

undersampling: better precision + recall (best was random forest)

random oversampling: same as undersample in terms of accuracy and recall

oversampling: better accuracy using smote (tries to fit line and chooses points close to that line), svm

do bargraph of proportion of readmitted vs. non-readmitted for each top 10 diagnostic ID + other for diag1, 2, 3

Presentation structure

Introduction: names, title

Structure of presentation

Objective, goals, data set

Data cleaning (brief summary of what we've done)

General EDA (visualizations, 3)

Modelling

Talk about problem of imbalanced data (show table)

Brief summary of solutions – what is oversampling and undersampling

* Method 1:
* Method 2:

(out of 3 methods, find the two that are most similar and pick one from that, and the remaining to show to class)

For each method:

* What the method does
* Results (features with greatest effect)

Comparing the two methods: cross-validation, pros and cons

Conclusion: which to pick depends on the question at hand. Return back to objective to make conclusion about which is better (for us)

Optimization

Results from optimization

References