CSCI 452 (Data Mining)
Dr. Schwartz
Imbalanced Classes and Ensemble Methods
100 pts

This assignment emphasizes sampling methods for handling imbalanced classes but also allows you to experiment with ensemble methods. There is more research involved in this assignment since you will need to decide how to preprocess your data, what sampling methods you want to use (and why), etc. **You can work on this assignment in pairs!**

Useful libraries may include caret, ROSE, or DMwR (you might only use one of these or you might choose to use all three).

Note that throughout this lab, both the results and the analysis are graded. Be sure to provide your analysis, hypotheses and conclusions in your writeup. You will create an R Markdown file called sampling.Rmd and submit both the .Rmd and resulting .html (after knitting) files to autolab. Please be sure to follow the order of the questions and identify your answers/analysis clearly so that I can easily find the individual pieces.

You will be using the breast cancer data set from UCI (http://archive.ics.uci.edu/ml/datasets/breast+cancer) – be sure to have this dataset in the same directory as your .Rmd file.

Here is a paper that presents a similar analysis to what you will be doing in this lab: https://www.researchgate.net/publication/320719501_A_Comparison_of_Class_Imbalance_Techniques_for_Real-World_Landslide_Predictions

1) **(60 pts) Dealing with Class Imbalance** Use the dataset referenced above. This dataset has a class imbalance and this assignment involves investigating ways to address this. The overall approach will be for you to choose some machine learning techniques that you want to try and test them with the imbalanced data, then with the adjusted data. Choose at least two different models such as logistic regression, decision trees, naïve bayes, svms (no ensembles… yet).
    a) Examine the data (start from the original data) and do any preprocessing you deem necessary. Document your decisions. There are missing values in this dataset.
    b) Divide your data into training and test sets. This division will remain constant throughout the assignment.
    c) Before addressing the class imbalance, attain some baseline results with your selected classifiers. You should report *at least* accuracy and F-1.
    d) Address the class imbalance problem with sampling techniques. Attempt both undersampling and oversampling, then use an approach that includes generating synthetic examples such as ROSE or SMOTE. Present and evaluate your results for all of your selected classifiers.

2) **(40 pts) Ensemble Methods – Choose at least two ensemble methods for this part**
   a) Hypothesize as to whether your ensemble methods will gracefully handle the class imbalance in the original (imbalanced) data. Explain your reasoning.
   b) Test your hypothesis.
   c) Address the class imbalance in the same ways that you did in the first part of the assignment and present and evaluate your results.

Submission instructions: Present your answers to the questions above, analysis, commentary, R code and visualizations in an R Markdown file. Name your file sampling.Rmd and, when you knit, create an .html file. Zip these files together (along with your independent data set if it's not too large) and upload your submission to Autolab as the Class Imbalance lab. Be sure to just select and zip the files, don't zip a directory.