Disclaimer: This set of homework applies SMOTE to a seriously imbalanced dataset with a large number of features and data points. SMOTE is essentially a time consuming method. You need to start doing this homework early, so that you have enough time to run SMOTE on the full dataset.

1. **Tree-Based Methods**

   (a) Download the APS Failure data from: `https://archive.ics.uci.edu/ml/datasets/APS+Failure+at+Scania+Trucks` . The dataset contains a training set and a test set. The training set contains 60,000 rows, of which 1,000 belong to the positive class and 171 columns, of which one is the class column. All attributes are numeric.

   (b) Data Preparation

   This data set has *missing values*. When the number of data with missing values is significant, discarding them is not a good idea. [1]

   i. Research what types of techniques are usually used for dealing with data with missing values.[2] Pick at least one of them and apply it to this data in the next steps.[3]

   ii. For each of the 170 features, calculate the coefficient of variation $CV = \frac{s}{m}$, where s is sample standard deviation and m is sample mean.

   iii. Plot a correlation matrix for your features using pandas or any other tool.

   iv. Pick $\lfloor\sqrt{170}\rfloor$ features with highest CV, and make scatter plots and box plots for them, similar to those on p. 129 of ISLR. Can you draw conclusions about significance of those features, just by the scatter plots? This does not mean that you will only use those features in the following questions. We picked them only for visualization.

   v. Determine the number of positive and negative data. Is this data set imbalanced?

   (c) Train a random forest to classify the data set. Do NOT compensate for class imbalance in the data set. Calculate the confusion matrix, ROC, AUC, and misclassification for training and test sets and report them (You may use pROC package). Calculate Out of Bag error estimate for your random forset and compare it to the test error.

   (d) Research how class imbalance is addressed in random forests. Compensate for class imbalance in your random forest and repeat 1c. Compare the results with those of 1c.

   (e) XGBoost and Model Trees

   In the case of a univariate tree, only one input dimension is used at a tree split. In a multivariate tree, or model tree, at a decision node all input dimensions can

[1] In reality, when we have a model and we want to fill in missing values, we do not have access to training data, so we only use the statistics of test data to fill in the missing values.

[2] They are called data imputation techniques.

[3] You are welcome to test more than one method.

be used and thus it is more general. In univariate classification trees, majority polling is used at each node to determine the split of that node as the decision rule. In model trees, a (linear) model that relies on all of the variables is used to determine the split of that node (i.e. instead of using $X_j > s$ as the decision rule, one has $\sum_j \beta_j X_j > s$. as the decision rule). Alternatively, in a regression tree, instead of using average in the region associated with each node, a linear regression model is used to determine the value associated with that node.

One of the methods that can be used at each node is Logistic Regression. Because the number of variables is large in this problem, one can use $\mathscr{L}_1$-penalized logistic regression at each node. You can use XGBoost to fit the model tree. Determine α (the regularization term) using cross-validation. Train the model for the APS data set without compensation for class imbalance. Use one of 5 fold, 10 fold, and leave-one-out cross validation methods to estimate the error of your trained model and compare it with the test error. Report the Confusion Matrix, ROC, and AUC for training and test sets.

(f) Use SMOTE (Synthetic Minority Over-sampling Technique) to pre-process your data to compensate for class imbalance.[4] Train XGBosst with $\mathscr{L}_1$-penalized logistic regression at each node using the pre-processed data and repeat 1e. Do not forget that there is a right and a wrong way of cross validation here. Compare the uncompensated case with SMOTE case.

2. ISLR 6.6.3

3. ISLR, 6.6.5

4. ISLR 8.4.5

5. ISLR 9.7.3

6. Extra Practice: ISLR 5.4.2, 6.8.4, 8.4.4, 9.7.2

[4]If you did not start doing this homework on time, downsample the common class to 6,000 so that you have 12,000 data points after applying SMOTE. Remember that the purpose of this homework is to apply SMOTE to the whole training set, not the downsampled dataset.