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.

3. Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:
- i. Increase initially, and then eventually start decreasing in an inverted U shape.
- ii. Decrease initially, and then eventually start increasing in a U shape.
- iii. Steadily increase.
- iv. Steadily decrease.
- v. Remain constant.

) Repeat (a) for test RSS.
) Repeat (a) for variance.
) Repeat (a) for (squared) bias.
) Repeat (a) for the irreducible error.

5. It is well-known that ridge regression tends to give similar coefficient values to correlated variables, whereas the lasso may give quite different coefficient values to correlated variables. We will now explore this property in a very simple setting.

Suppose that $n = 2$, $p = 2$, $x_{11} = x_{12}$, $x_{21} = x_{22}$. Furthermore, suppose that $y_1 + y_2 = 0$ and $x_{11} + x_{21} = 0$ and $x_{12} + x_{22} = 0$, so that the estimate for the intercept in a least squares, ridge regression, or lasso model is zero: $\hat{\beta}_0 = 0$.

(a) Write out the ridge regression optimization problem in this setting.

(b) Argue that in this setting, the ridge coefficient estimates satisfy $\hat{\beta}_1 = \hat{\beta}_2$.

(c) Write out the lasso optimization problem in this setting.

lasso regression tries to find coefficient estimates $\hat{\beta}_j$ which minimize:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| \quad ①$$

plug in: $n=2$, $p=2$,

then $$① = \sum_{i=1}^{2}\left(y_i-\beta_0-\sum_{j=1}^{2}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{2}|\beta_j|$$

$$= (y_1-\beta_0-\beta_1 x_{11}-\beta_2 x_{12})^2+(y_2-\beta_0-\beta_1 x_{21}-\beta_2 x_{22})^2+\lambda(|\beta_1|+|\beta_2|) \quad ②$$

(d) Argue that in this setting, the lasso coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ are not unique—in other words, there are many possible solutions to the optimization problem in (c). Describe these solutions.

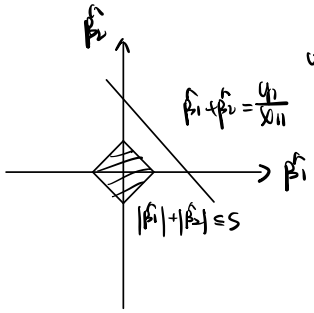
from (c), plug $x_{11}=x_{12}$, $x_{21}=x_{22}$, $y_1+y_2=0$, $x_{11}+x_{21}=0$, $x_{12}+x_{22}=0$ and $\beta_0=0$ into ②

and similar to (b), ② $= f(\hat{\beta}_1,\hat{\beta}_2) = 2(y_1-x_{11}(\hat{\beta}_1+\hat{\beta}_2))^2+\lambda(|\hat{\beta}_1|+|\hat{\beta}_2|)$

then change optimization to minimize: $2(y_1-x_{11}(\hat{\beta}_1+\hat{\beta}_2))^2$ subject to $|\hat{\beta}_1|+|\hat{\beta}_2| \le s$

so any $\hat{\beta}_1, \hat{\beta}_2$ satisfying $2(y_1-x_{11}(\hat{\beta}_1+\hat{\beta}_2))^2=0$

$\hat{\beta}_1+\hat{\beta}_2 = \frac{y_1}{x_{11}}$ makes RSS minimized



considering the constraint $|\hat{\beta}_1|+|\hat{\beta}_2| \le s$, the solutions of satisfied pairs $(\hat{\beta}_1, \hat{\beta}_2)$ are not unique since its values along the overlap of line $\hat{\beta}_1+\hat{\beta}_2=\frac{y_1}{x_{11}}$ and diamond $|\hat{\beta}_1|+|\hat{\beta}_2| \le s$, there are many points

5. Suppose we produce ten bootstrapped samples from a data set containing red and green classes. We then apply a classification tree to each bootstrapped sample and, for a specific value of X, produce 10 estimates of $P(\text{Class is Red}|X)$:

$$0.1, 0.15, 0.2, 0.2, 0.55, 0.6, 0.6, 0.65, 0.7, \text{ and } 0.75.$$

There are two common ways to combine these results together into a single class prediction. One is the majority vote approach discussed in this chapter. The second approach is to classify based on the average probability. In this example, what is the final classification under each of these two approaches?

3. Here we explore the maximal margin classifier on a toy data set.

(a) We are given $n = 7$ observations in $p = 2$ dimensions. For each observation, there is an associated class label.

Obs.	X_1	X_2	Y
1	3	4	Red
2	2	2	Red
3	4	4	Red
4	1	4	Red
5	2	1	Blue
6	4	3	Blue
7	4	1	Blue

Sketch the observations.

(b) Sketch the optimal separating hyperplane, and provide the equation for this hyperplane (of the form (9.1)).

(c) Describe the classification rule for the maximal margin classifier. It should be something along the lines of "Classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise." Provide the values for β_0, β_1, and β_2.

(d) On your sketch, indicate the margin for the maximal margin hyperplane.

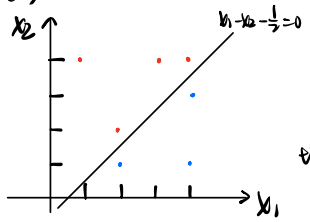
(e) Indicate the support vectors for the maximal margin classifier.

(f) Argue that a slight movement of the seventh observation would not affect the maximal margin hyperplane.

(g) Sketch a hyperplane that is *not* the optimal separating hyperplane, and provide the equation for this hyperplane.

(h) Draw an additional observation on the plot so that the two classes are no longer separable by a hyperplane.

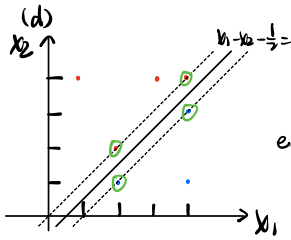
(a)



(b) equation: $X_1 - X_2 - \frac{1}{2} = 0$
this optimal separating hyperplane makes the gap between two class biggest

(c) classify to Red if $\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$, and classify to Blue otherwise, which $\beta_0 = \frac{1}{2}$, $\beta_1 = -1$, $\beta_2 = 1$

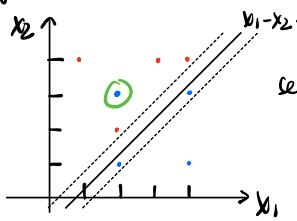
(d)



the margin for maximal margin hyperplane is two dashed lines
equations are $\begin{cases} X_1 - X_2 = 0 \\ X_1 - X_2 - 1 = 0 \end{cases}$

(e) indicate by green circle.

(g)



line $X_1 - X_2 - \frac{3}{4} = 0$ is not the optimal separating hyperplane
its margin $M = \frac{\sqrt{2}}{8}$
is smaller than that of $X_1 - X_2 - \frac{1}{2} = 0$ ($M = \frac{\sqrt{2}}{4}$)

(h) Show in green circle