Muhammed Halbutogullari

913938544

ECS 171 HW3

Preparation:

I used the file: ecs171.dataset.xlsx. When prepping the dataset, I noticed the final column had NaN values, so I discarded that column entirely. This left us with a dataset of size (194,4500).

Problem 1)

Ridge regression is said to be optimal for highly correlated feature variables, which seems to be the case with our massive dataset. Ridge regression is a form of linear regression, which is differentiated by the addition of a bias factor. The incorporation of a bias factor reduces variance and the rigidity of the response coefficients exhibit on features. Given these aspects, Ridge Regression seems like the optimal choice due to the high number of features in our data set.

With a lambda of 0, Ridge Regression does not have feature selection, so we need to implement regularization. Testing various alphas [0.01,0.1,1,2,3,5,10,25], our Ridge model found the optimal constrained parameter to be 0.1.

There were 4434 non-zero coefficients.

Our 5-fold cross validation generalization error for each fold: [0.42878123 0.12992284 0.33557091 0.03204945 0.17825269]

The average error of our 5-fold cross validation: 0.22091542557323918

Problem 2)

We begin by assuming we do not have any outliers and that our data is fully validated. We also assume that our input (to the function) bacteria is comparable to, or one of, the same bacteria than those in the dataset. We go through 200 iterations of bootstrapping. Each iteration samples roughly half the data set (randomly selects 100 bacteria), fits the optimal ridge model to this sample, and predicts the growth rate for the given bacteria (input for the function) using the newly fit model. We append the results of each iteration to a list and find the 2.5 and 97.5 percentile of that list. This gives us the 95% confidence interval cutoff values.

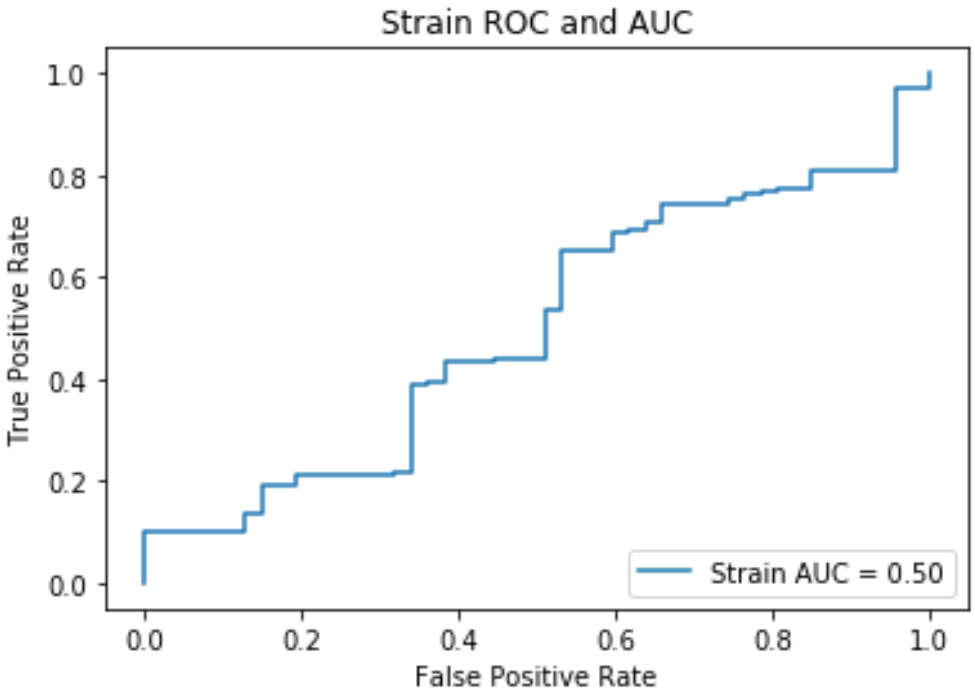
Problem 3)

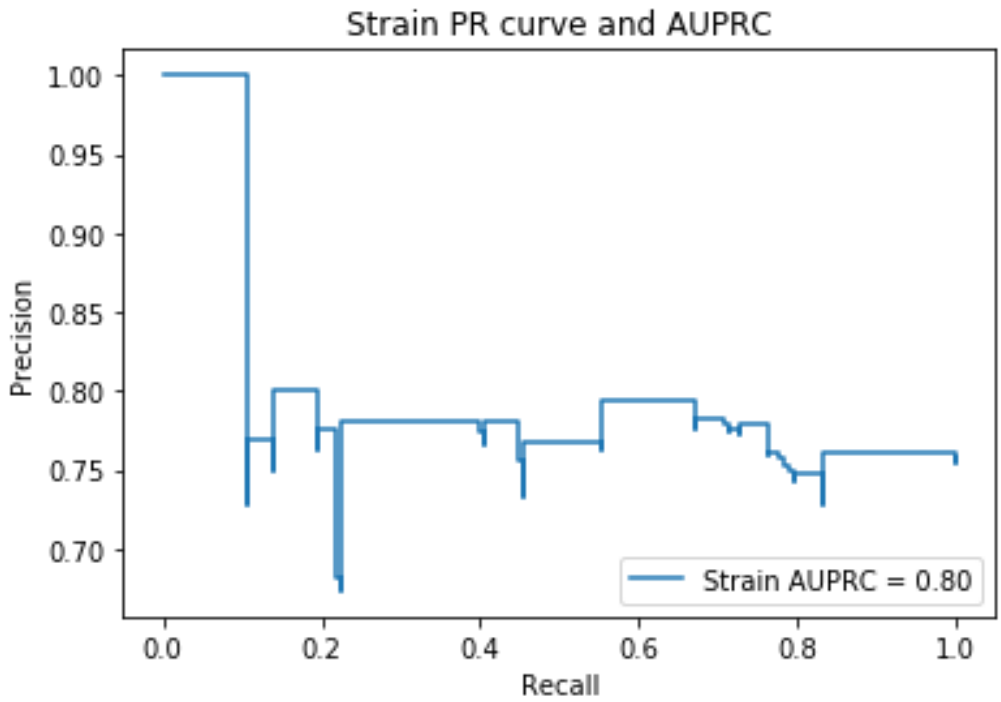
By taking the average of every column, we derived a hypothetical bacterium, comprised of genes expressed exactly at the mean expression value. Our ridge model predicted the growth rate for this bacterium to be 0.394. The 95% confidence interval fell between 0.382 and 0.405.

Problem 4)

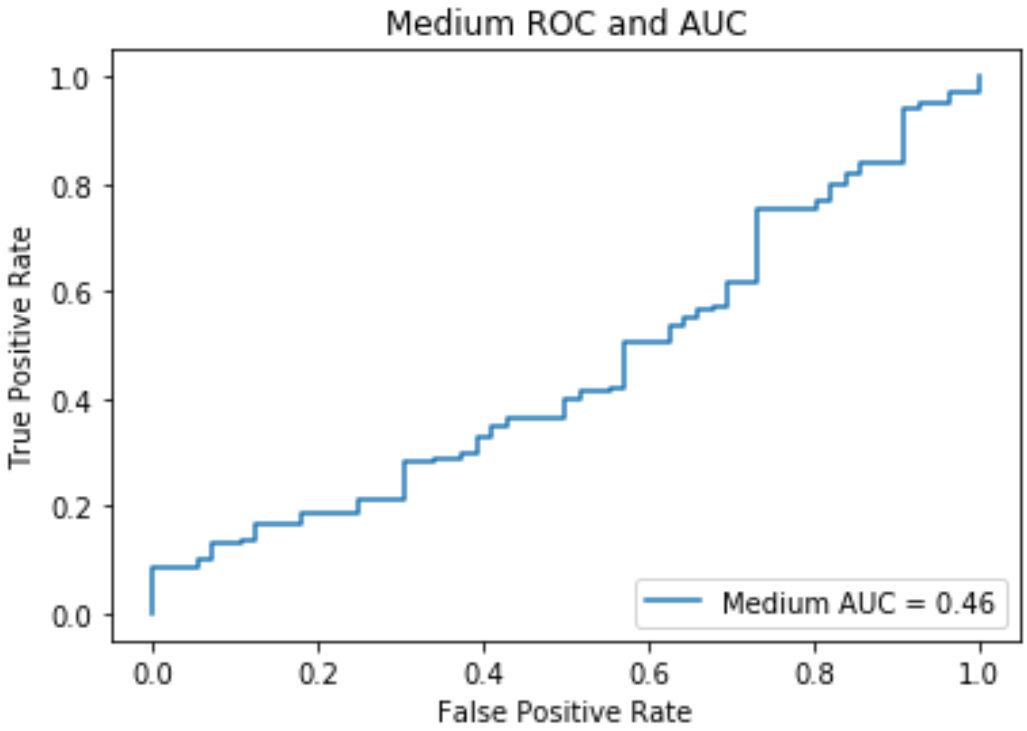
For this problem we will use the non-zero weighted features from problem 1. Therefore, we will be considering 4434 features for all our classifiers. AUC and AUPRC scores are reported within their respective graphs.

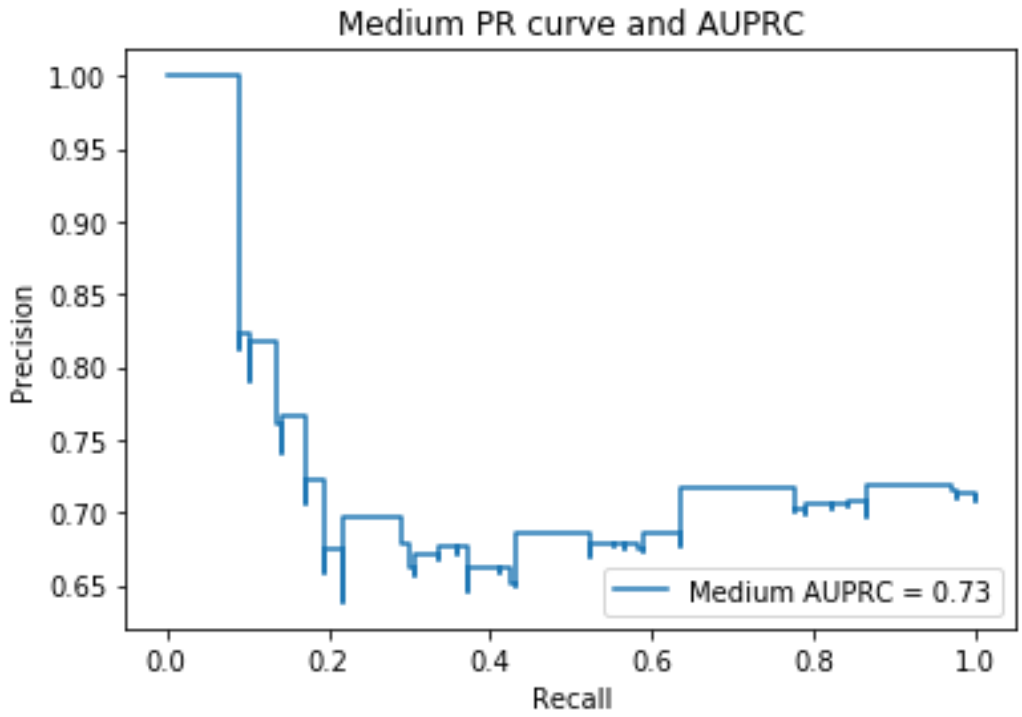
Strain:



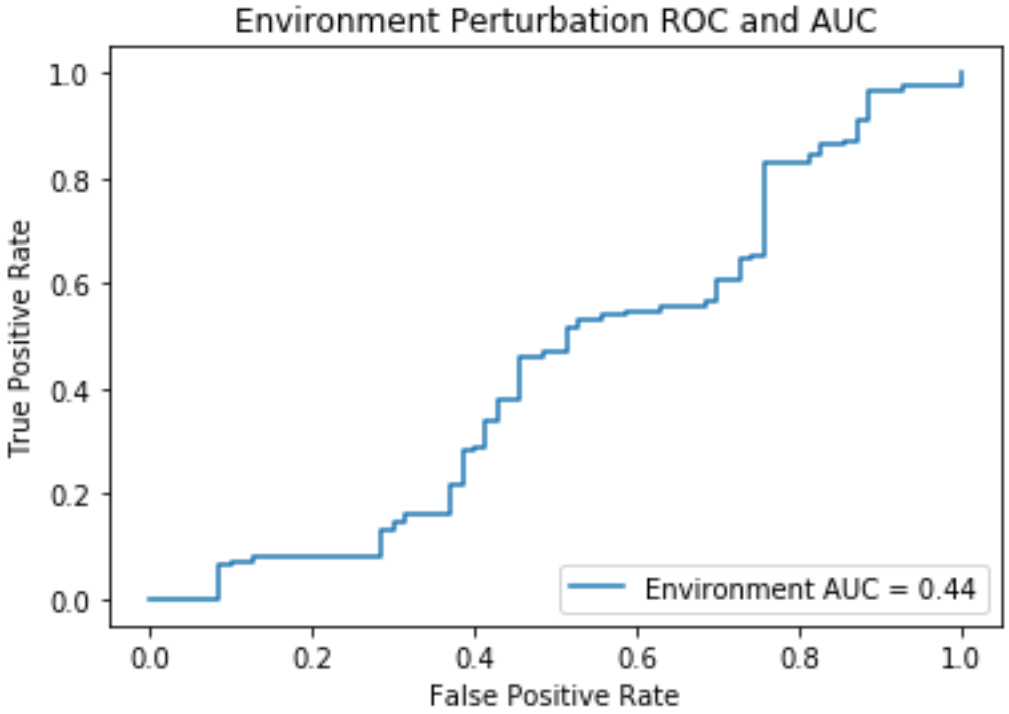


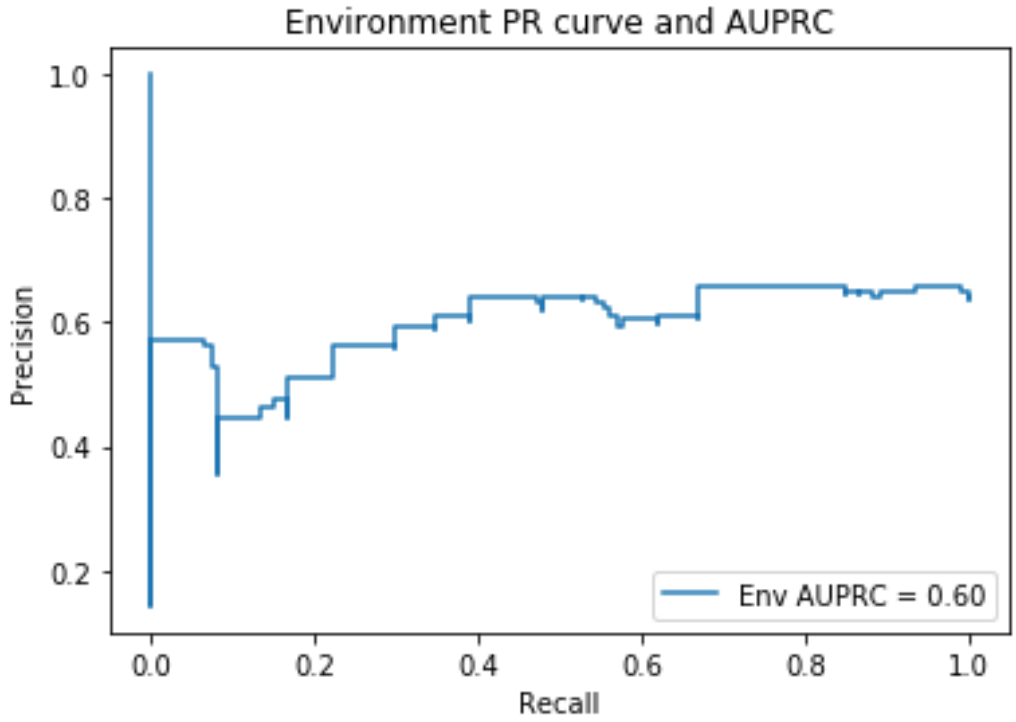
Medium:


Medium ROC and AUC
True Positive Rate
False Positive Rate
Medium AUC = 0.46
0.0
0.2
0.4
0.6
0.8
1.0

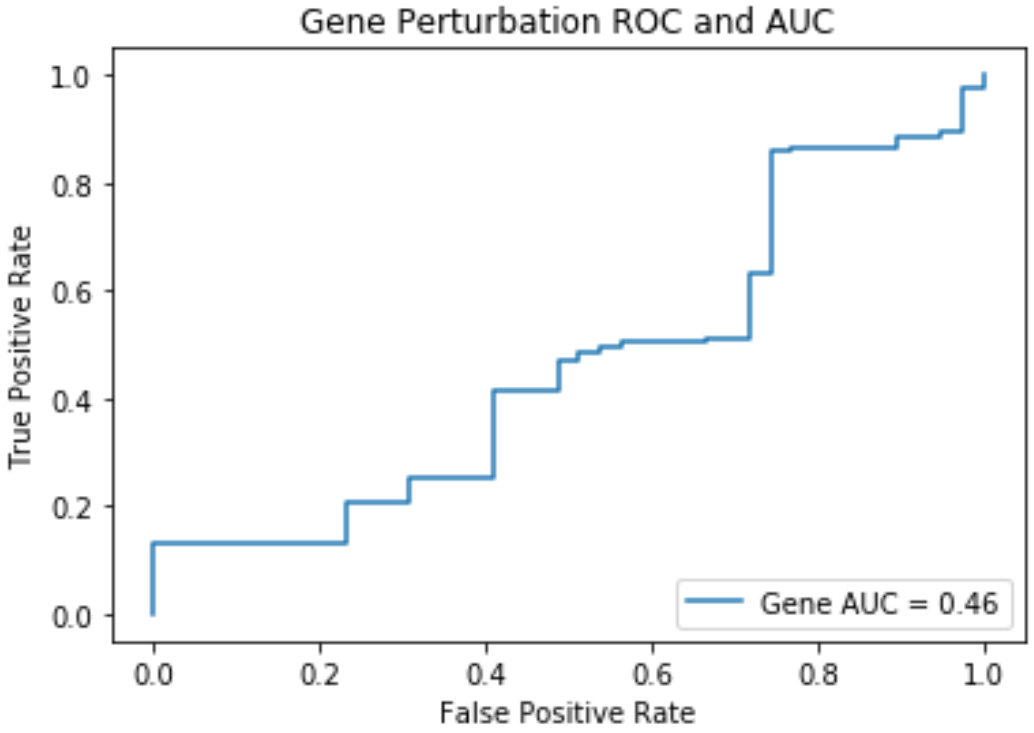

Medium PR curve and AUPRC
Precision
Recall
Medium AUPRC = 0.73
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.0
0.2
0.4
0.6
0.8
1.0

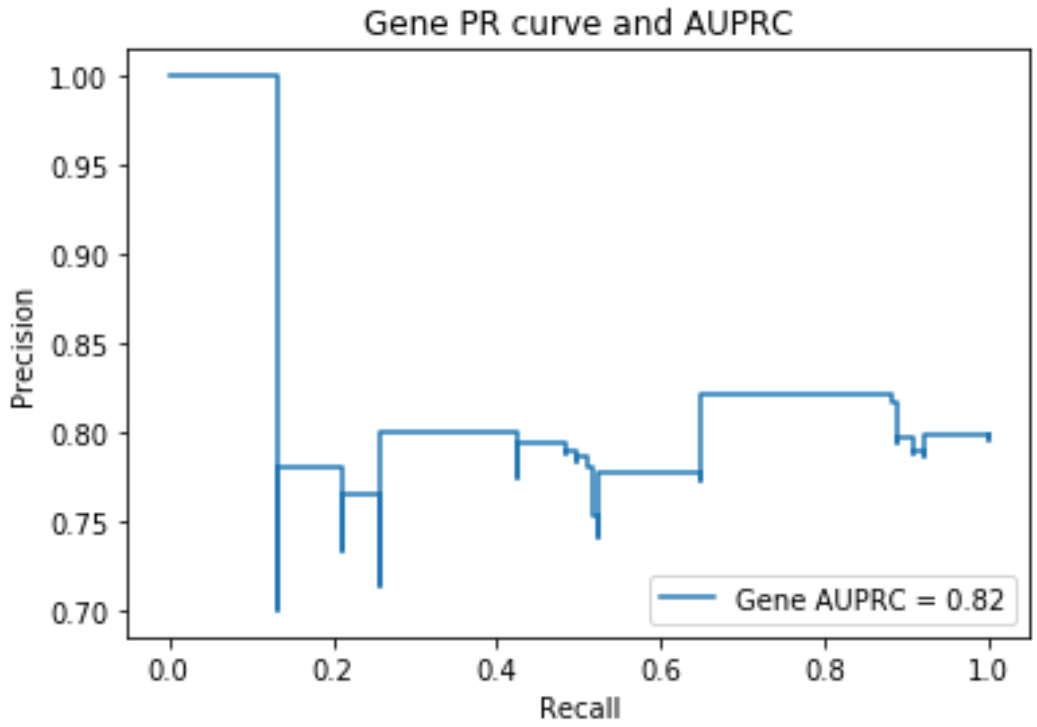
Environment perturbation:





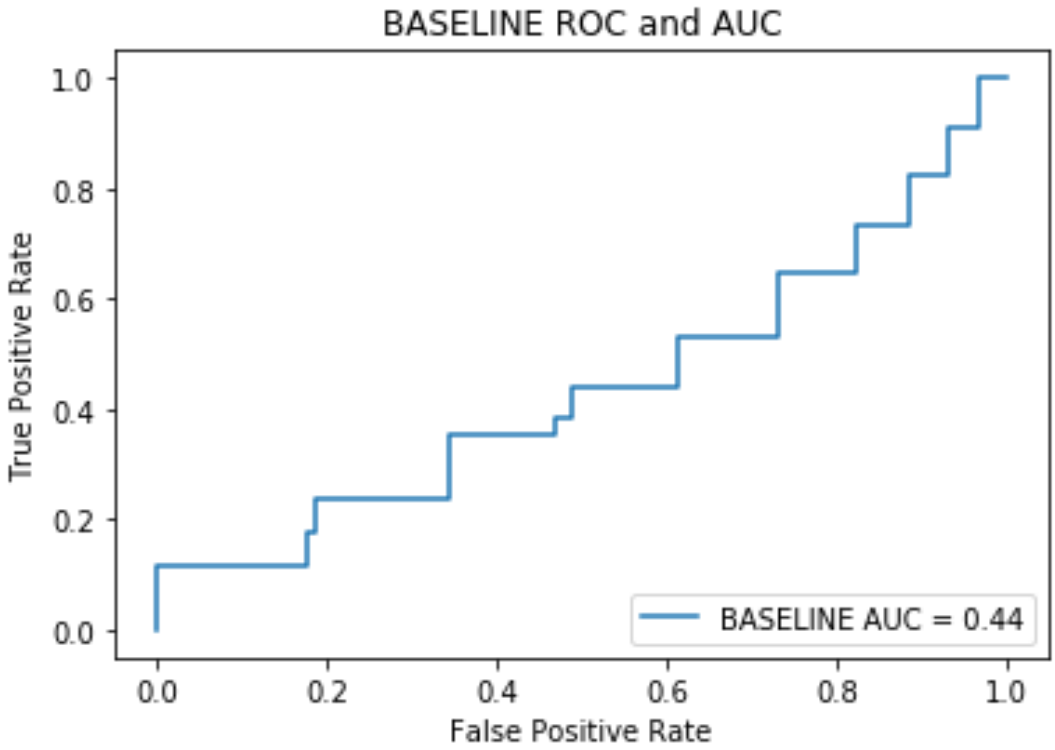
Gene perturbation:

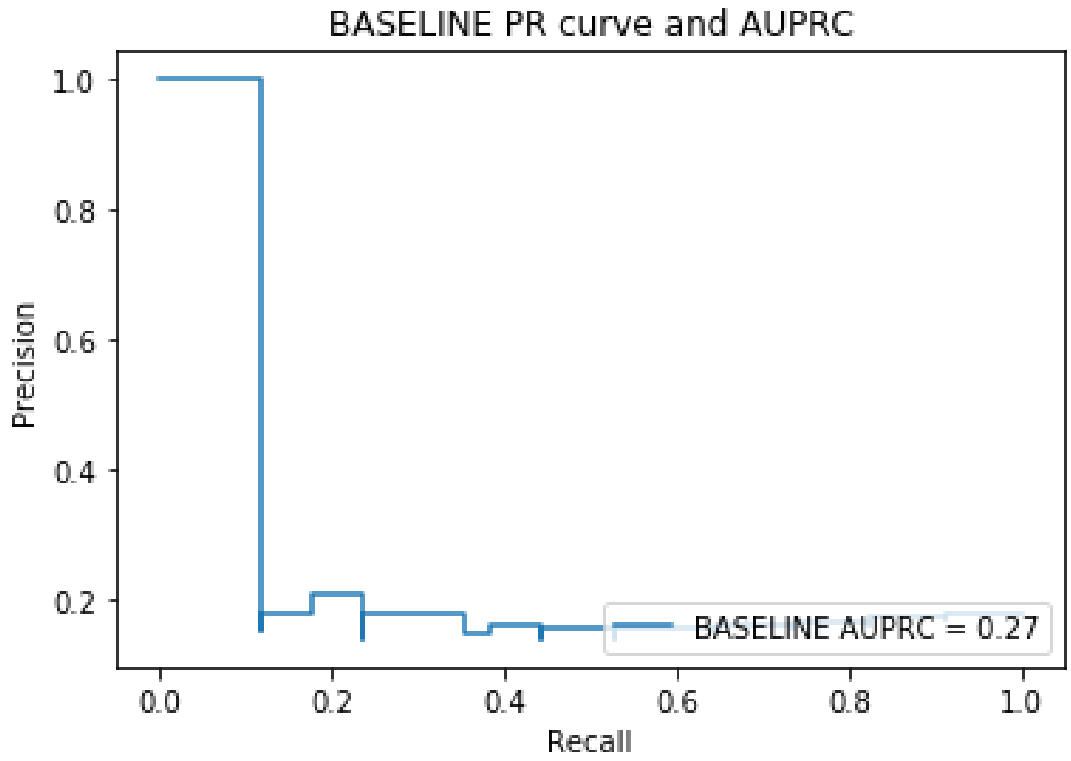




Problem 5)

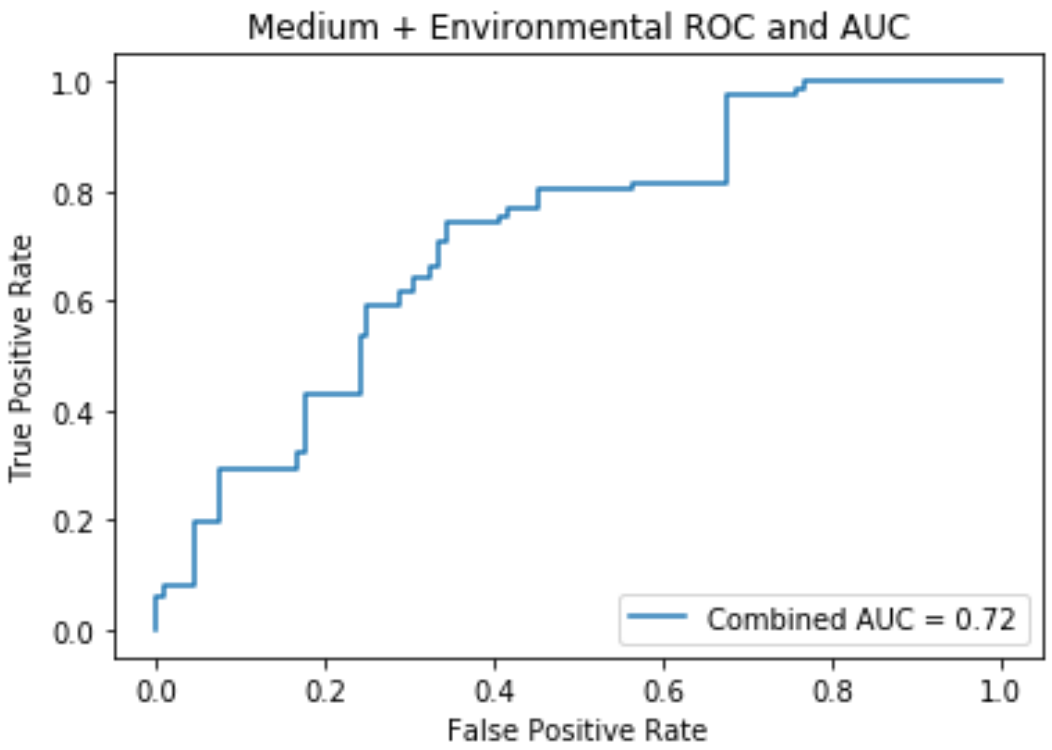
We will be using the non-zero features that we found in problem 1 (also previously used in problem 5). We are using 10-fold Cross Validation this time. Using the most common joint feature as the prediction (regardless of input), we constructed our baseline prediction ROC and PR.

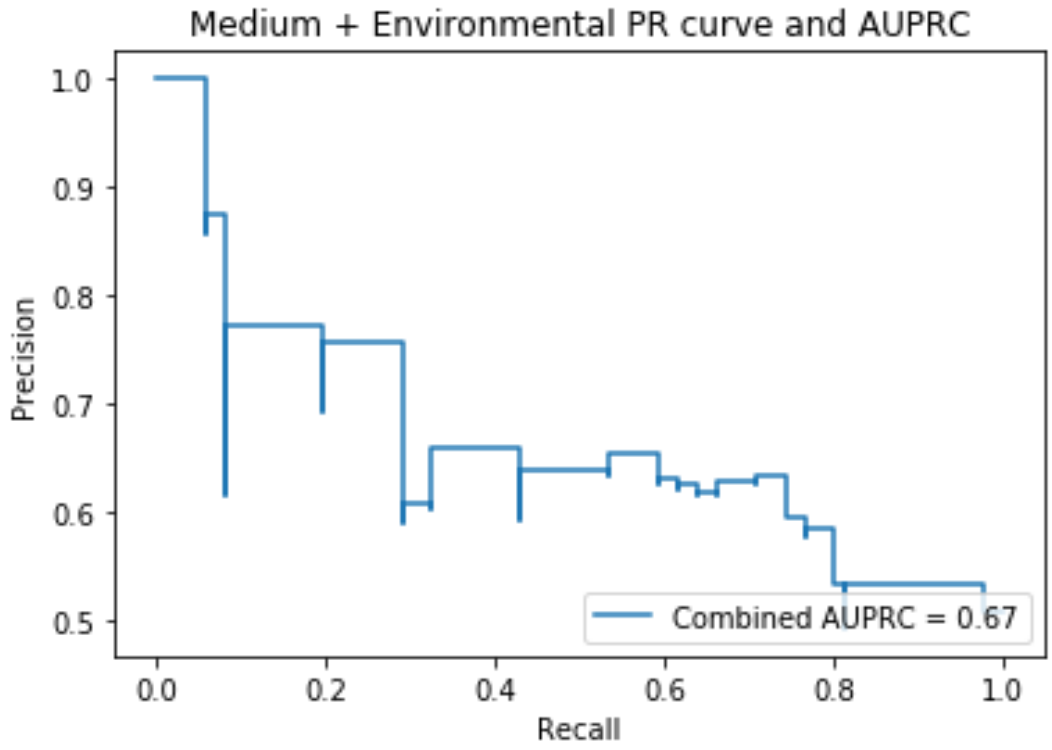




Continuing onto our combined classifier (graphs below): we see the AUC is much greater than the baseline, indicating that it is performing better than classification by simply choosing the mode of the expected output. The AUC of 0.72 is significantly greater than that of solely Medium (0.46), and that of Environment (0.44). This means that our combined classifier has a higher true positive rate than there was for either individual classifier, indicating we are better off using a combined classifier for regression, rather than use two separate classifiers for Medium and Environment.

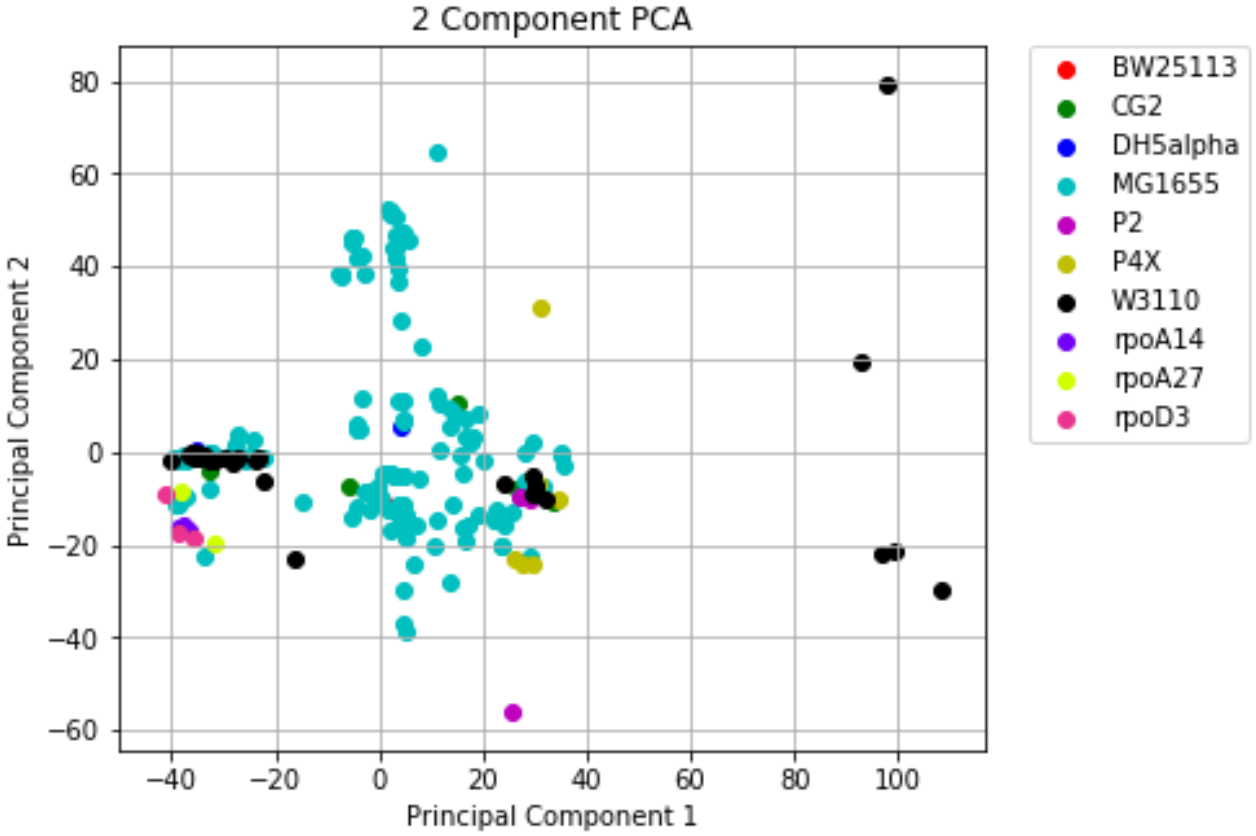
For the PR curves, we note that our combined classifier again has a much higher AUPRC (0.67) vs the baseline AUPRC (0.27). This means that our precision is much higher with the combined classification vs simply choosing the mode of the expected output Again, this indicates our combined classification is working, performing better than just a guess. The combined classifier had a higher AUPRC (0.67) than that of Environment (0.6), indicating the combined classifier had a had a higher precision rate than did the Environment classifier alone. The combined AUPRC was still below that of Medium (0.73), allowing us to conclude that the Medium classifier had a higher precision compared to the combined classifier.

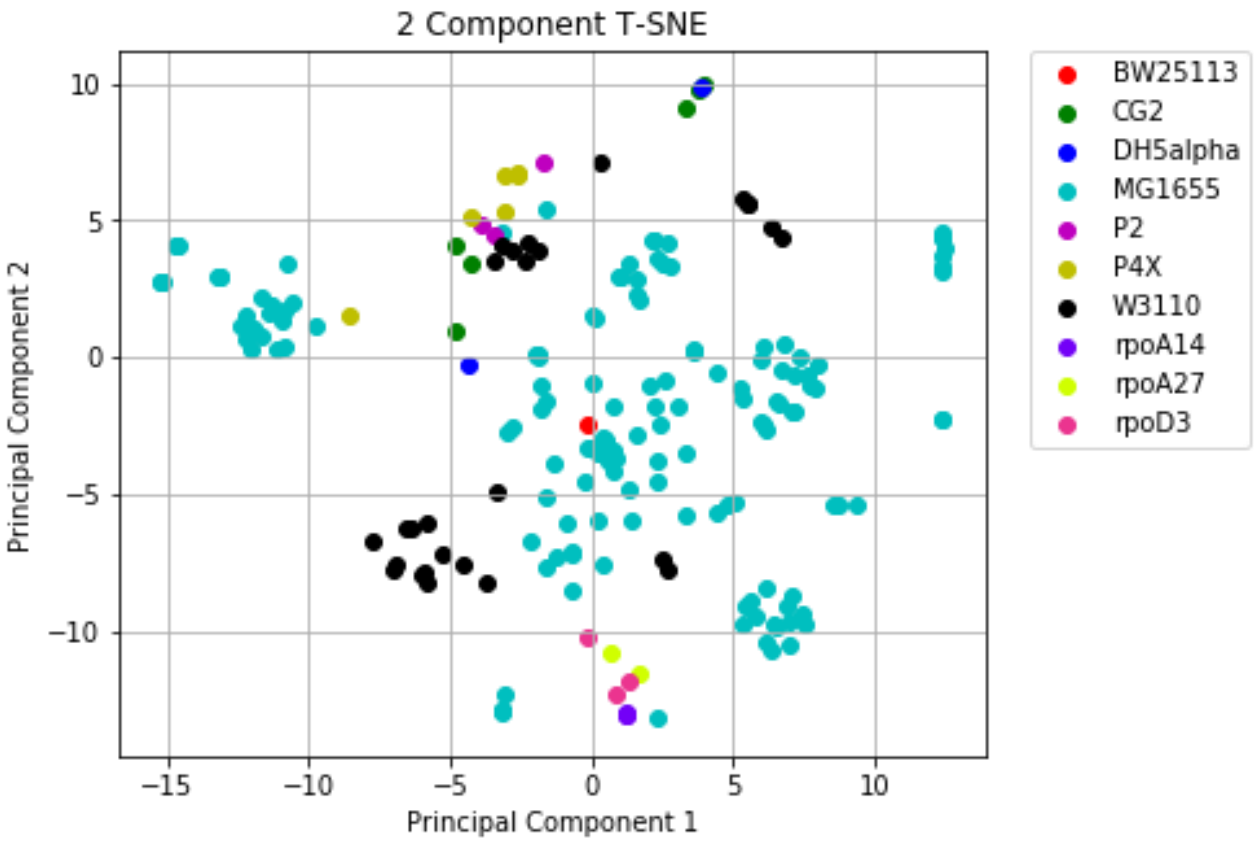

Medium + Environmental ROC and AUC
True Positive Rate
False Positive Rate
Combined AUC = 0.72
0.0
0.2
0.4
0.6
0.8
1.0


Medium + Environmental PR curve and AUPRC
Precision
Recall
Combined AUPRC = 0.67
0.5
0.6
0.7
0.8
0.9
1.0
0.0
0.2
0.4
0.6
0.8
1.0

Problem 6)

Before reducing the gene features, we normalize the dataset. This is crucial for both PCA and T-SNE to work properly. For both methods, we graphed the principal components along with the corresponding strain. As we see below, PCA seems to have reduced the dimensions to a small section of the graph, with some outliers (especially from the strain W3110). On the other hand, t-SNE does a better job: with more accurate grouping and virtually no outliers, it provides a significantly less dense graph, resulting in a better visualization which is easy to observe and understand.

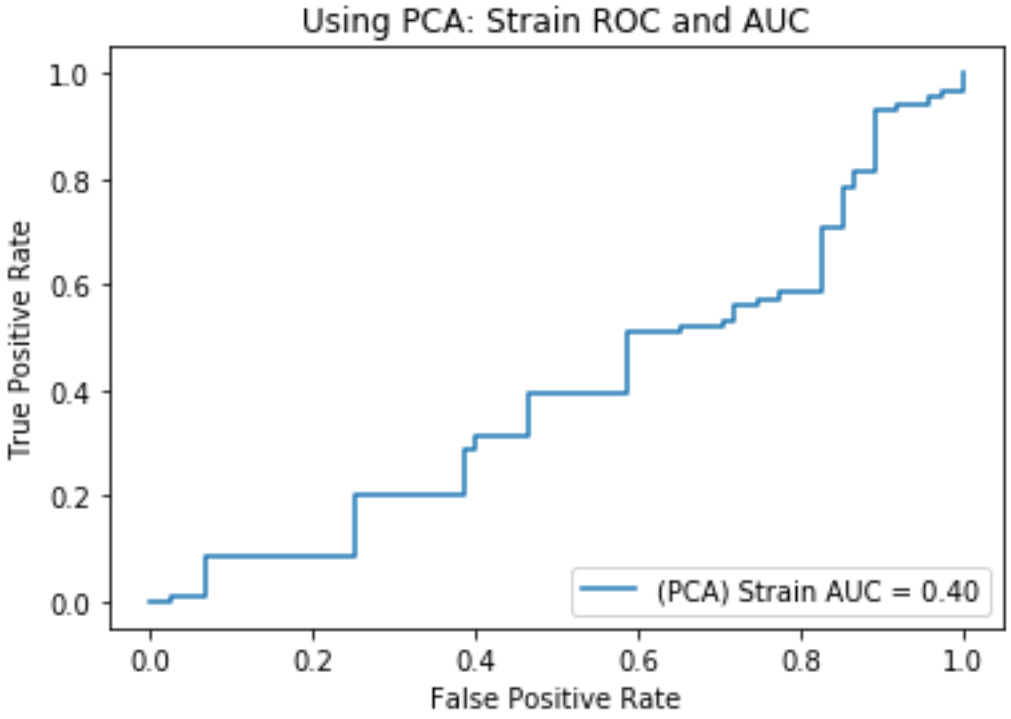


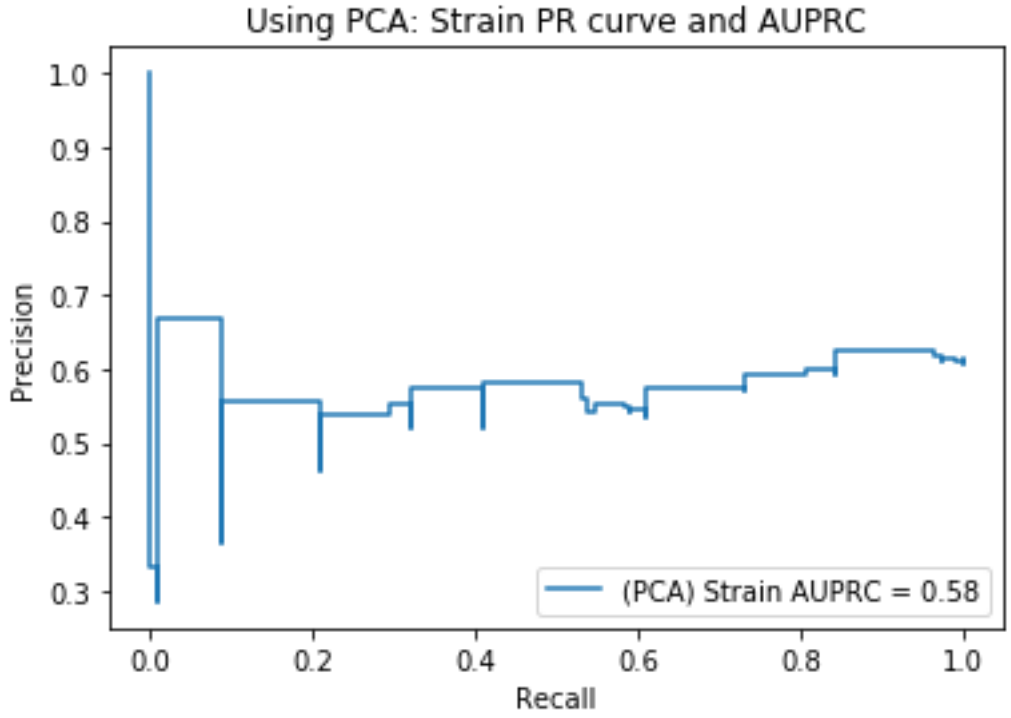


Problem 7)

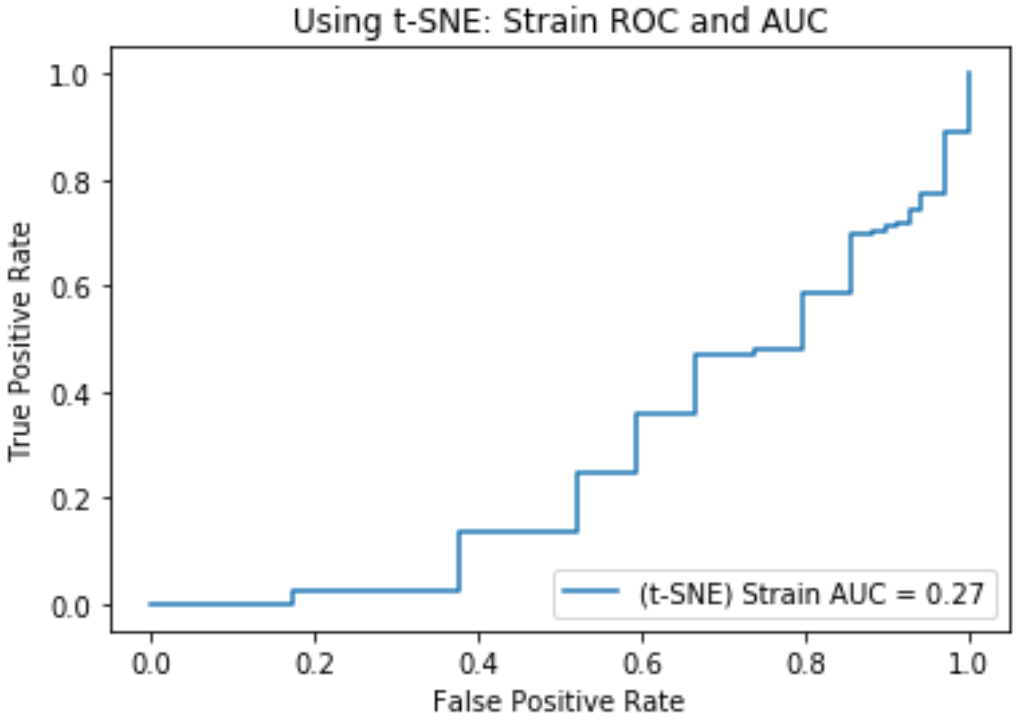
First, we recreate the reduced datasets (PCA, t-SNE) from step 6, both of which reduce the dimensionality of our dataset to two dimensions. We will rerun the entirety of problem 4 on each one of these datasets and determine the optimal approach for each classifier. Once again, the AUC and AUPRC values are located within their respective graphs.

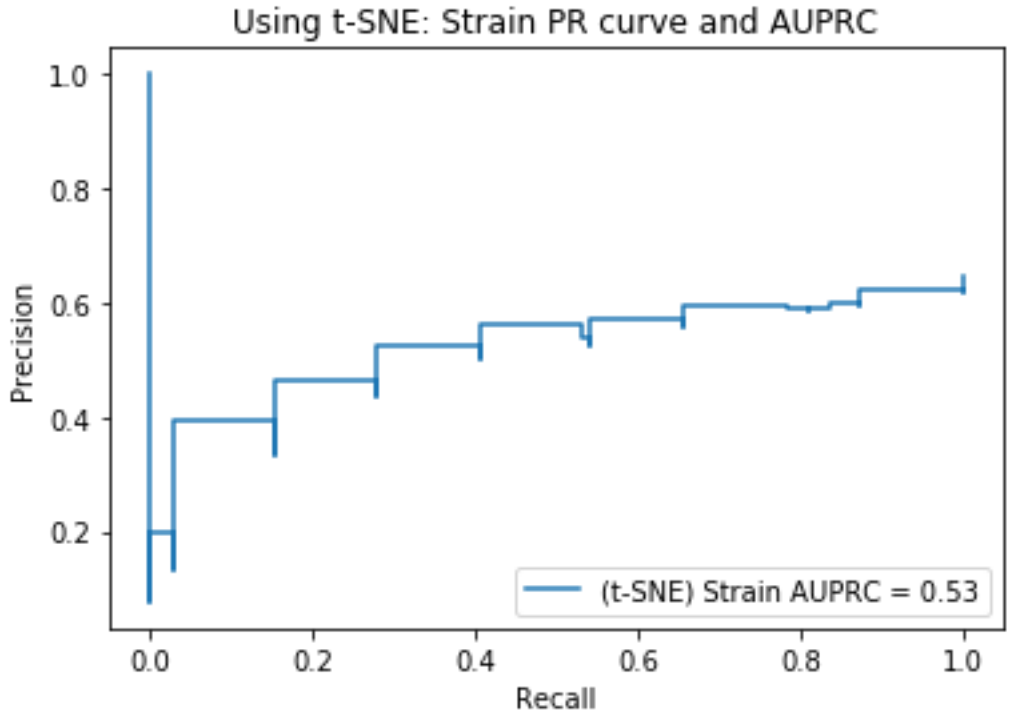
Strain using PCA:





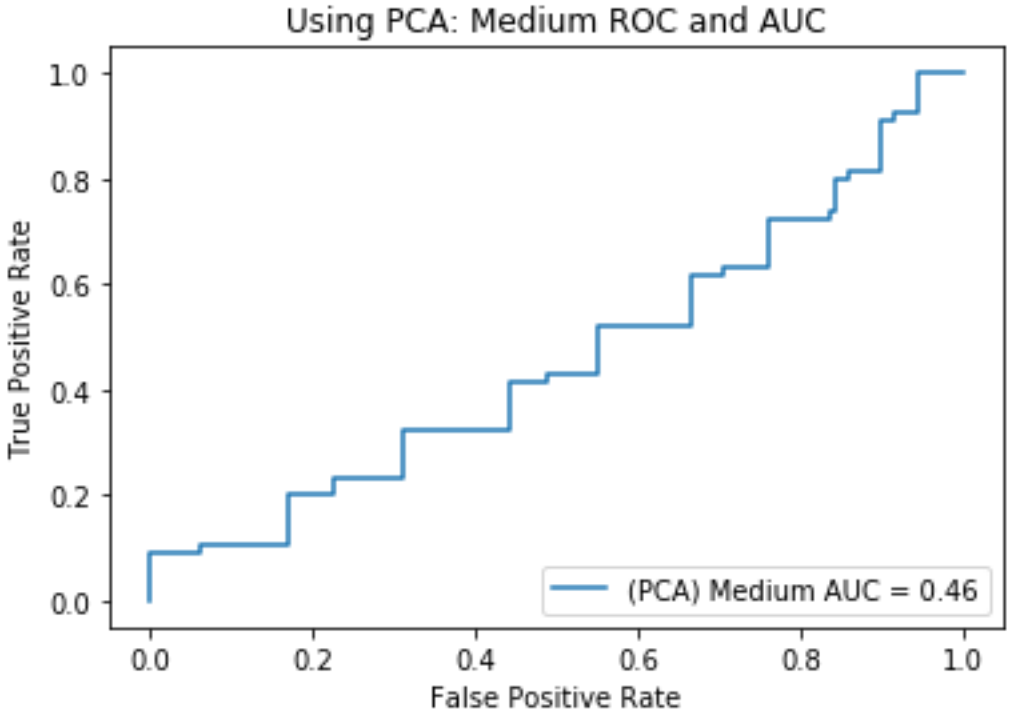
Strain using t-SNE:

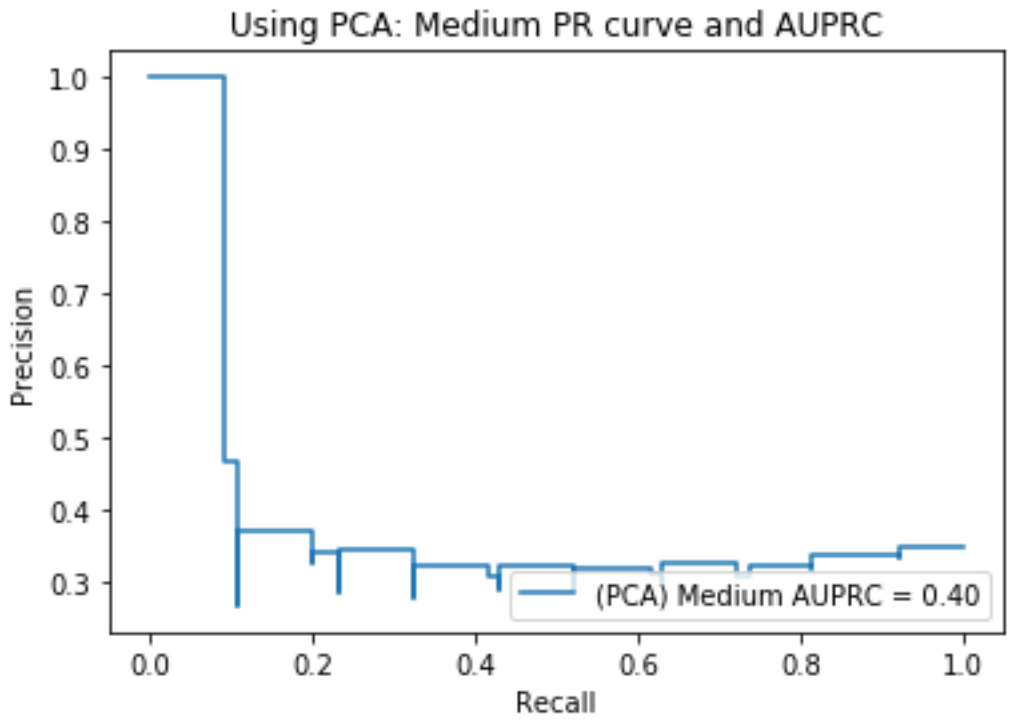




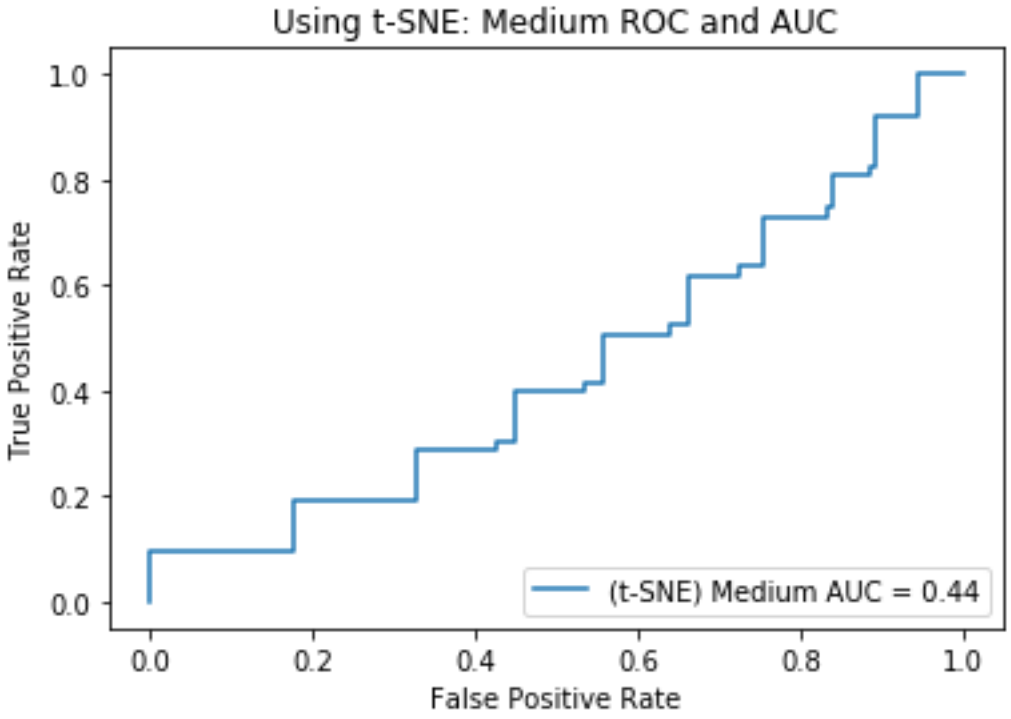
Comparing the results for Strain, we see that PCA had a higher AUC and AUPRC with respect to t-SNE. The original strain results found in problem 4 (where we used feature selection) resulted in a greater AUC and AUPRC, compared to either PCA or t-SNE. Therefore, we can conclude that the best pre-processing approach for our Strain ridge regression classifier is the feature selection we did in part 4.

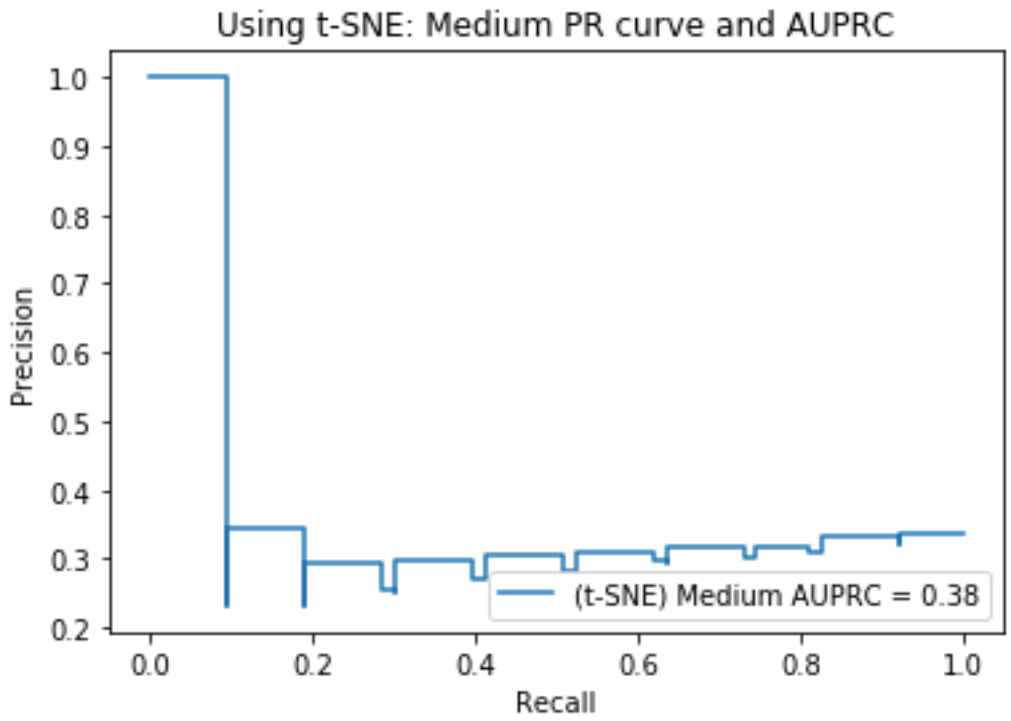
Medium using PCA:





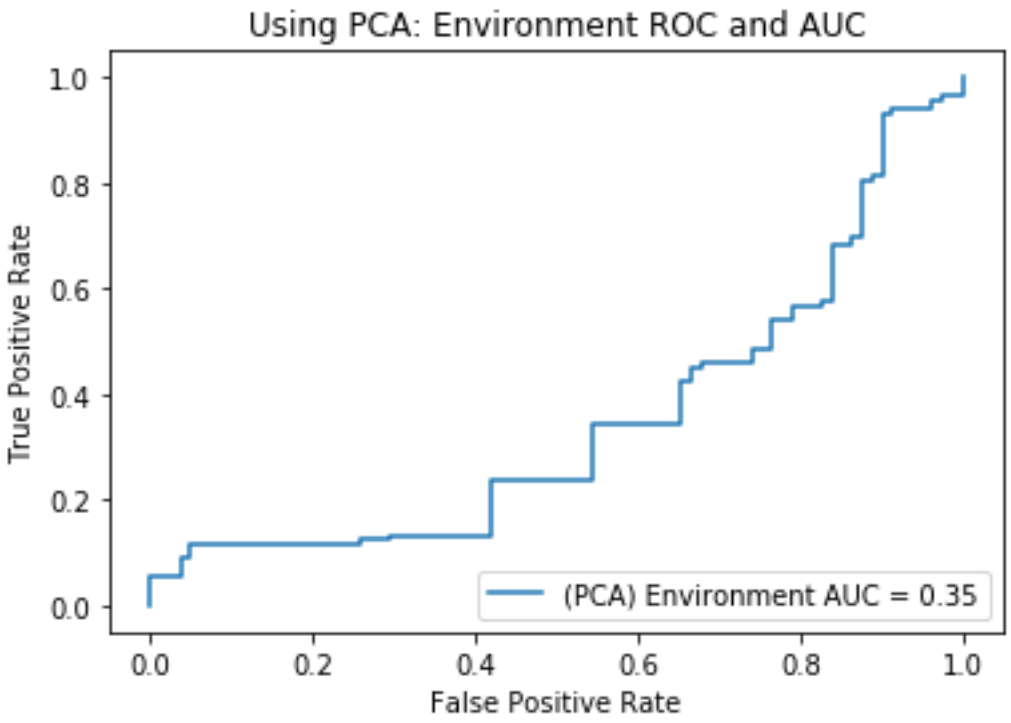
Medium using t-SNE:

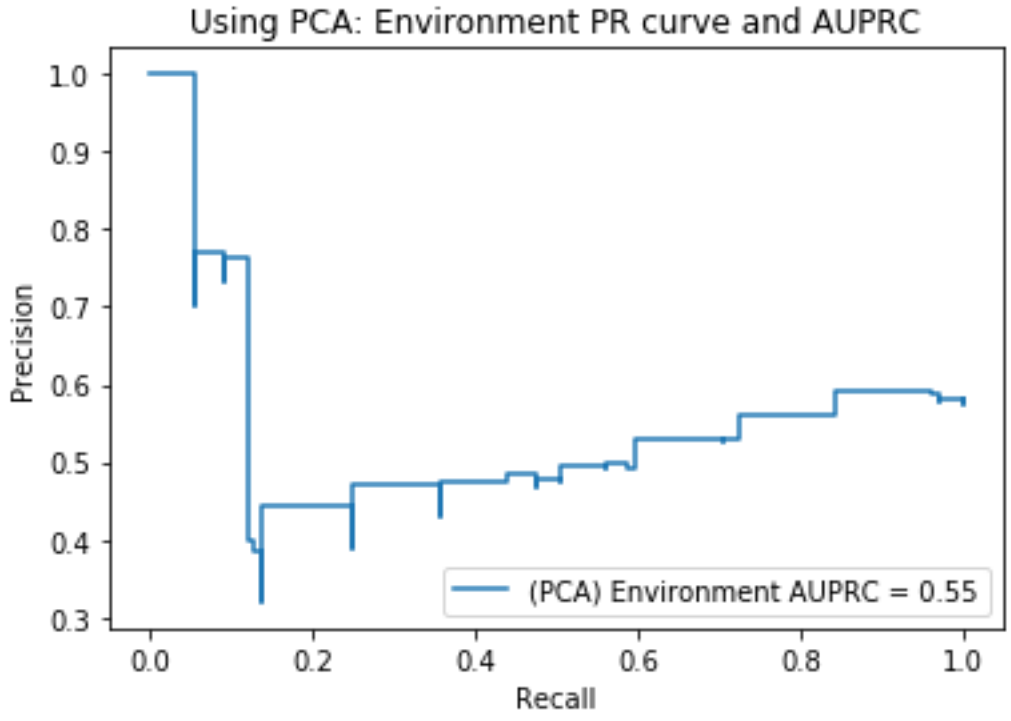




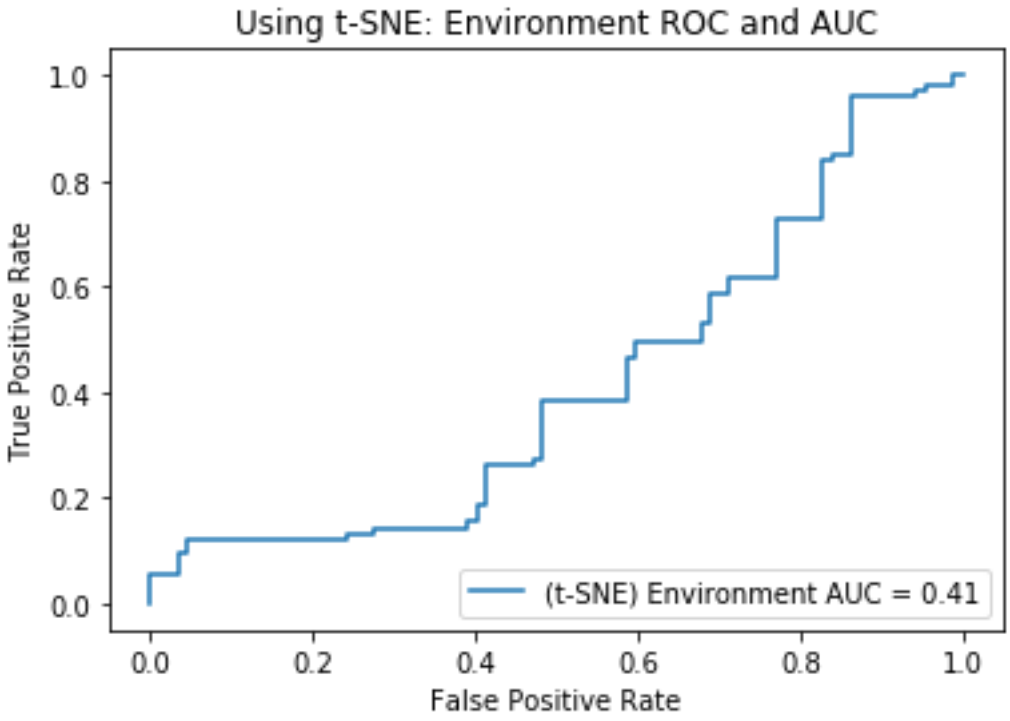
Once again, we note that classifying Medium using PCA achieved a greater AUC and AUPRC than it did with t-SNE. This time, the AUC when using PCA was the exact same as the AUC when using feature selection (problem 4). However, the PCA AUPRC was 0.4, much less than the AUPRC of 0.73 calculated in problem 4. Since the precision of the classifier was much higher in problem 4, we can safely assess that feature selection is (once again) the best pre-processing approach for the Medium Ridge classifier.

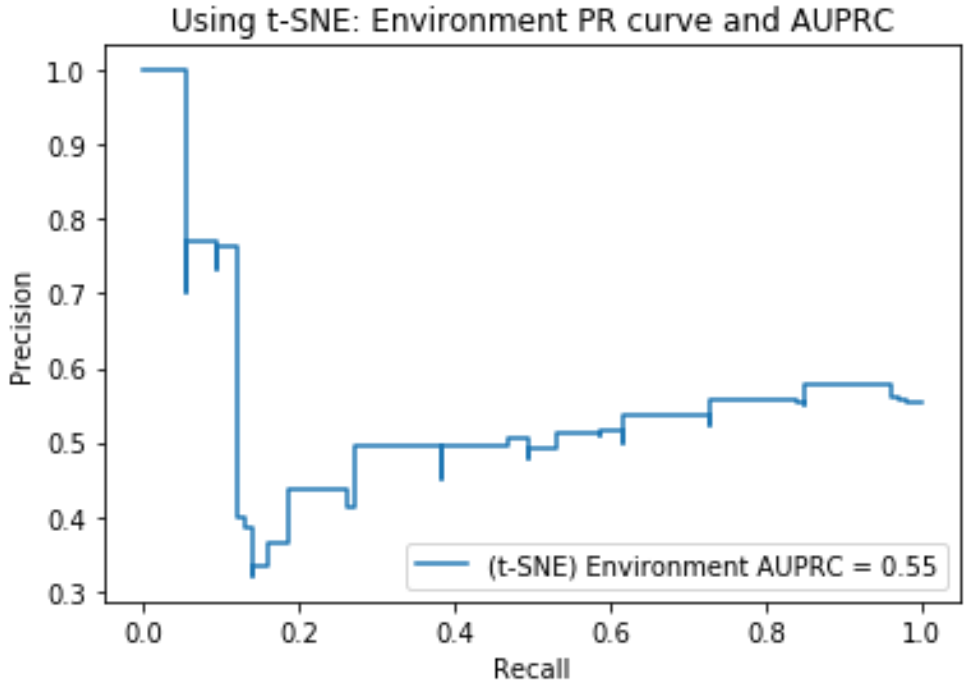
Environment using PCA:





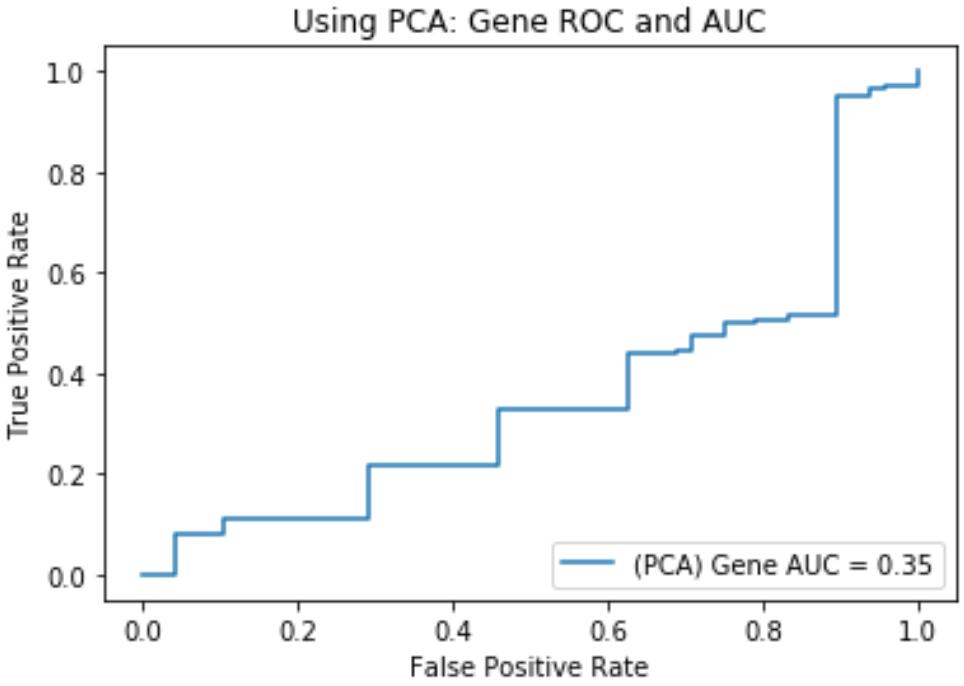
Environment using t-SNE:

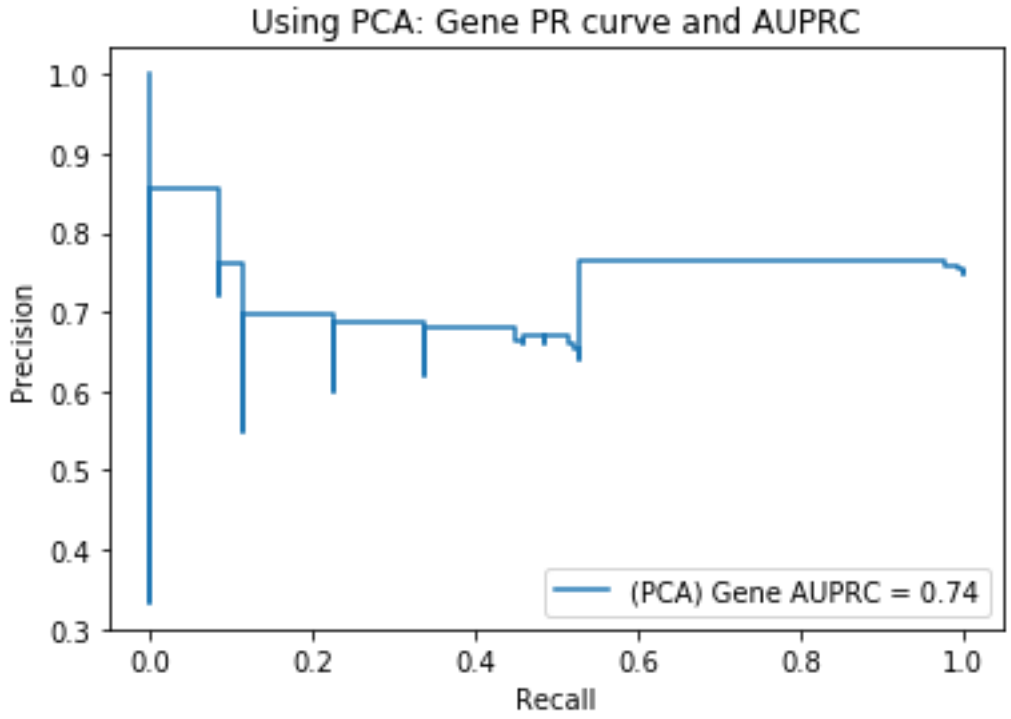




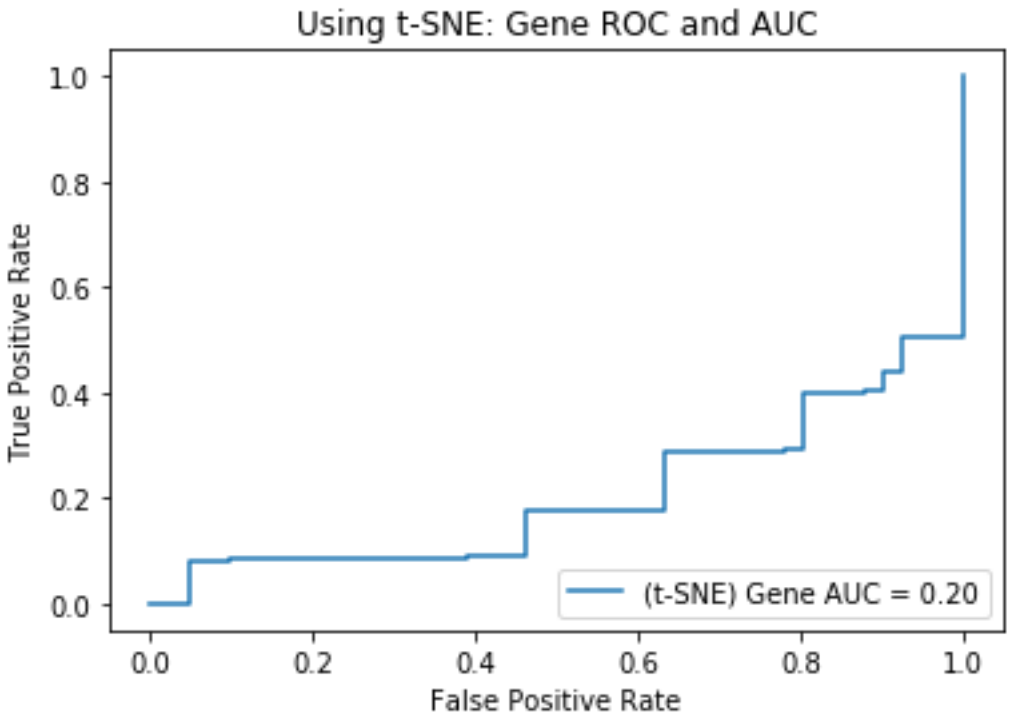
This time, t-SNE provided the superior results among the two preprocessing approaches (PCA, t-SNE) to classifying environment. When comparing the AUC and AUPRC to the feature selection approach (problem 4), the t-SNE results fall slightly short of the AUC and AUPRC that were calculated for Environment using feature selection. Once again, feature selection is the superior preprocessing approach.

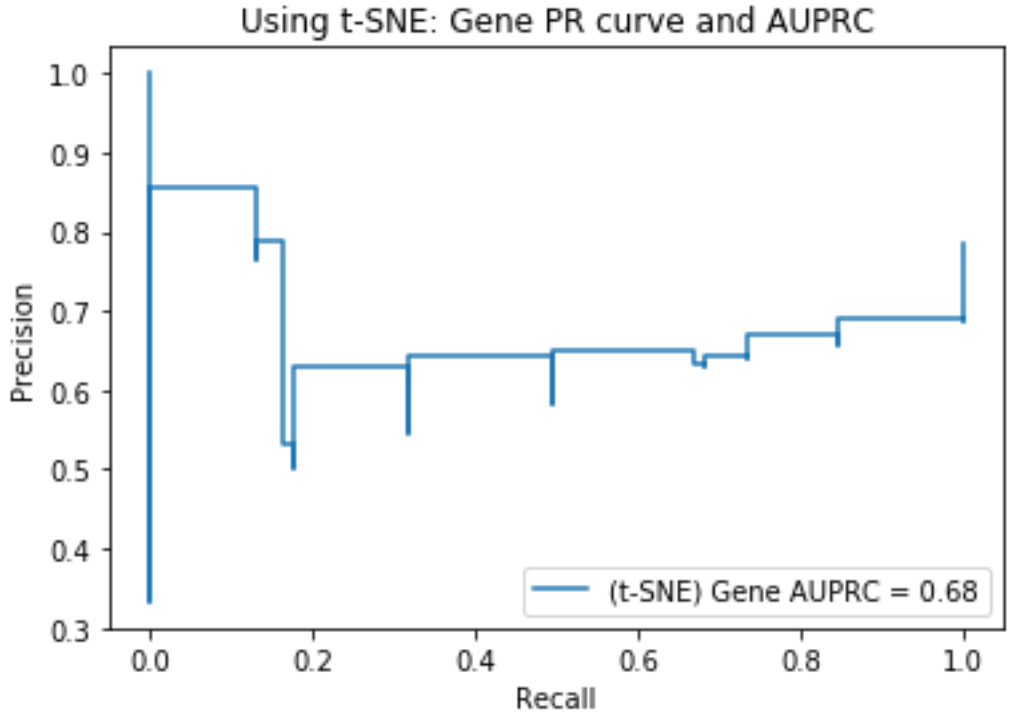
Gene using PCA:





Gene using t-SNE:





Among PCA and t-SNE, PCA performs much better when classifying based on Gene. Comparing the results of PCA to the results we found in problem 4, we see that the AUC and AUPRC found in problem 4 are much greater than those found with PCA. Therefore, we

conclude that that feature selection is the best preprocessing approach for classifying Gene using Ridge regression.

For each of our four classifiers, we found that the best pre-processing approach was feature selection, every time. There were some instances where the AUC/AUPRC of the PCA and the feature selection preprocessing approaches were quite close. An example of this was our use of t-SNE for classifying Environment, where we observed the AUC scores to be almost identical. In these cases, the loss in performance may be worth the significant simplifications to the dataset (a reduction from 4500 to only 2 columns).

References:

https://www.statisticshowto.datasciencecentral.com/ridge-regression/

https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60

https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/

https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

https://www.datacamp.com/community/tutorials/principal-component-analysis-in-python

https://towardsdatascience.com/pca-using-python-scikit-learn-e653f8989e60