CS 5013 Homework 4

Task 1:

We can see normal performance in the relationship between testing and training until around 60% of the data is used for training. At this point overfitting occurs, as we can see from the testing error actually increasing as more data is used for training.

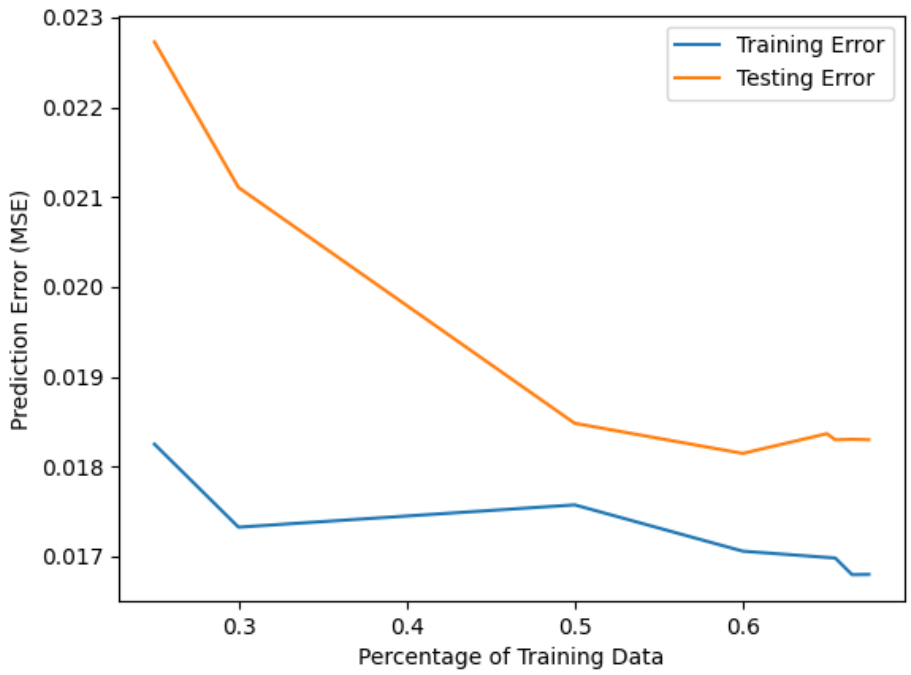

Figure 1

Task 2:

Here we can observe both underfitting and overfitting. From the high error rates in both testing and training, the model is too simple to be very effective. As alpha increases, training error increases, and testing error begins to increase slightly as well, indicating overfitting.

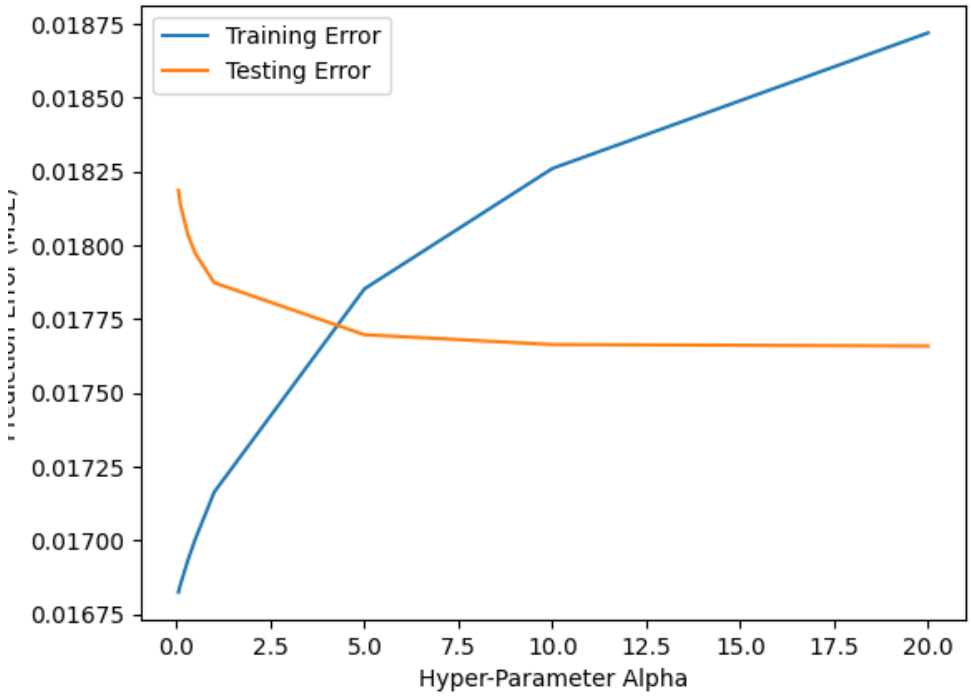

Figure 2

Task 3:

My program produced the following table.

Hyperparameter	0.05	0.1	0.5	1.0	5.0
Error	0.019816	0.019707	0.019446	0.019388	0.019496

The algorithm finds an optimal hyperparameter value somewhere around alpha=1.0, as it has the lowest error value.

Task 4:

Here, we can again observe overfitting as the percentage of data used for training increases. The classification error in testing increases over 60% of the data being used for training while the training error continually decreases.

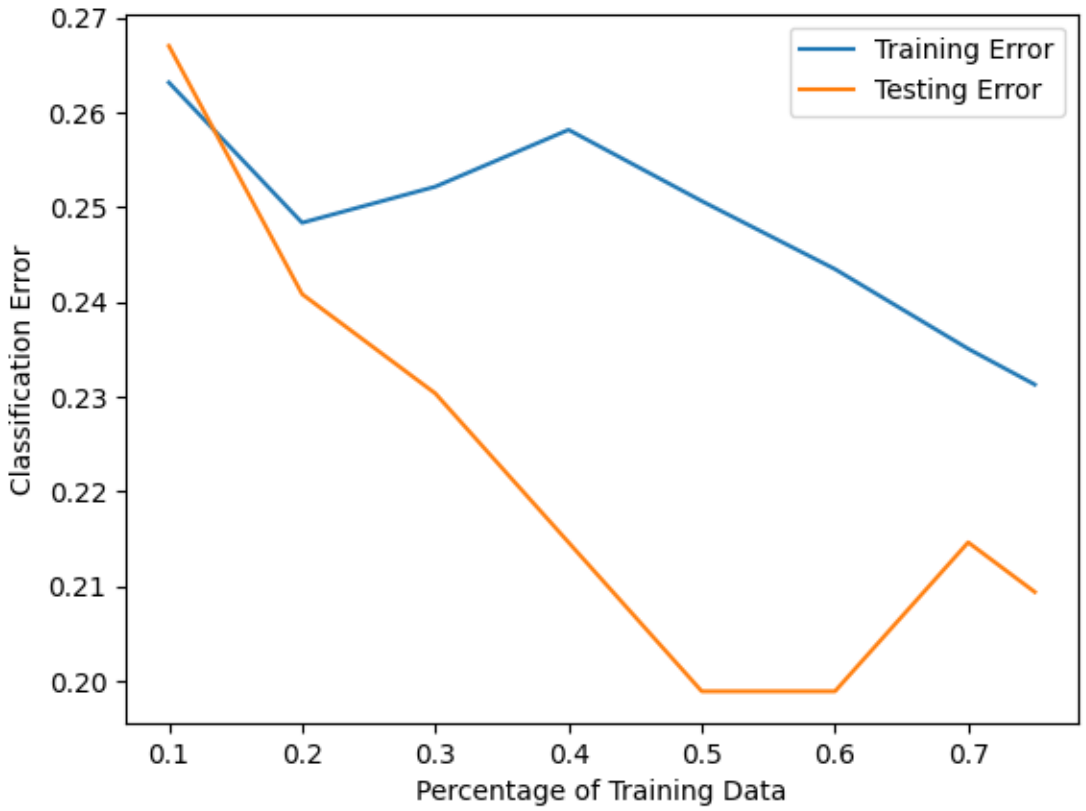

Figure 3

Task 5:

For this task, I implemented random oversampling on the unbalanced dataset, increasing the number of training records that corresponded with the minority classification. We can see that the AUC score of this method is better than that of the base method, while the accuracy is worse.

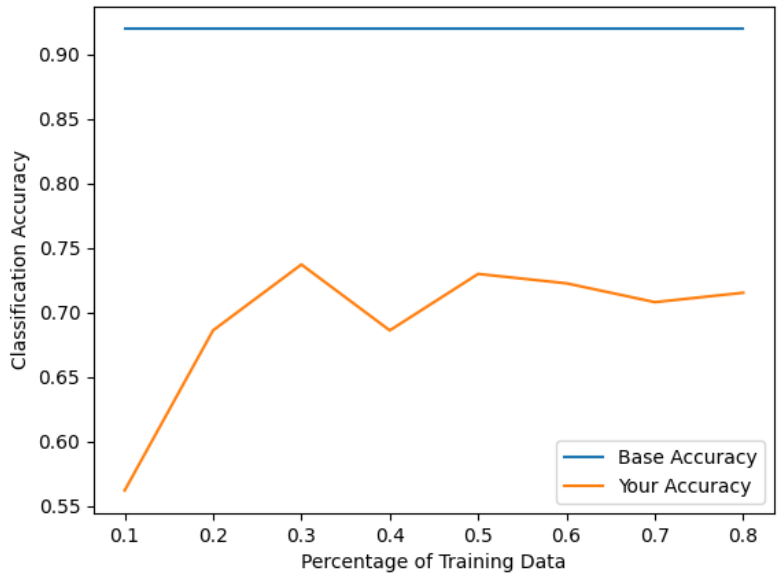

Figure 4

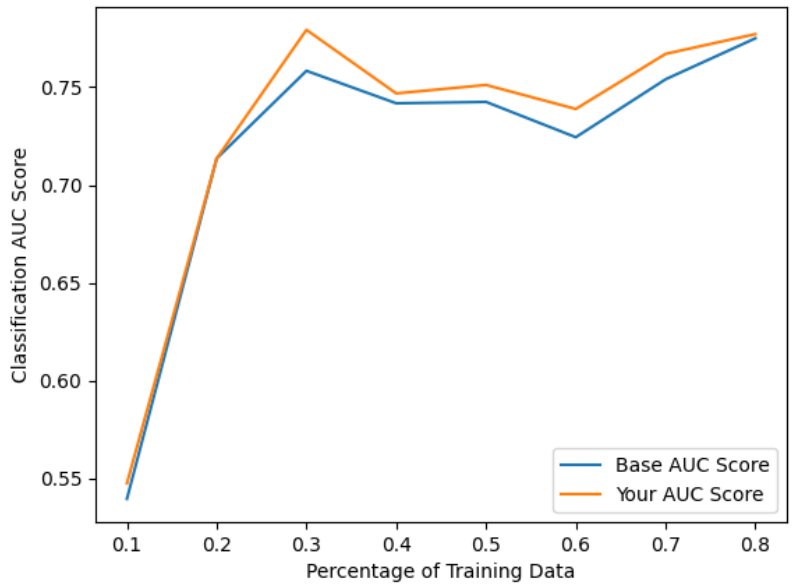

Figure 5