HW1

ZONGQI CUI

October 3, 2024

1 Problem 1

(a) As λ (the regularization parameter) increases, the penalty for large coefficients in LASSO increases. This causes the model to shrink the coefficients of the less relevant features, and some coefficients may become zero. This shrinking effect increases the bias because the model becomes simpler and less capable of capturing the complexity of the data.

(b) As λ increases, the model becomes simpler, often with fewer non-zero coefficients. This reduction in model complexity leads to less sensitivity to the training data. In simpler terms, the model becomes more robust to small variations in the training data, which reduces variance. Therefore, as λ increases, the variance tends to decrease.

(c) When $\lambda = 0$, LASSO is equivalent to ordinary least squares (OLS) regression, which fits the model to minimize the residual sum of squares without any penalty on the coefficients. In this case, the model is highly flexible, and there is no regularization. Therefore, the bias is low or close to zero because the model perfectly fits the training data.

(d) When $\lambda = \infty$, the regularization term becomes extremely large, forcing all the coefficients to be zero. This results in the simplest possible model (i.e., predicting the mean of the response variable for all inputs). In this scenario, the model has no flexibility, leading to extremely high bias but zero variance, because the model's prediction does not change regardless of the input data.

2 Problem 2

Preprocessing	**Train Accuracy**	**Train AUC**	**Test Accuracy**	**Test AUC**
Standardization	0.816000	0.890666	0.810119	0.875263
No Preprocessing	0.825333	0.946728	0.816989	0.942270
Log Transformation	0.823667	0.954344	0.815116	0.948122
Binarization	0.798667	0.949411	0.800125	0.944730

Table 1: Accuracy and AUC of Naive Bayes on Training and Test Sets Across Different Preprocessing Steps

2(c)

Preprocessing	**Train Accuracy**	**Train AUC**	**Test Accuracy**	**Test AUC**
Standardization	0.932667	0.977157	0.919425	0.969743
No Preprocessing	0.716333	0.858528	0.742036	0.859597
Log Transformation	0.948667	0.985473	0.936290	0.983044
Binarization	0.939333	0.981319	0.925671	0.978431

Table 2: Accuracy and AUC of Logistic Regression on Training and Test Sets Across Different Preprocessing Steps

2(e)

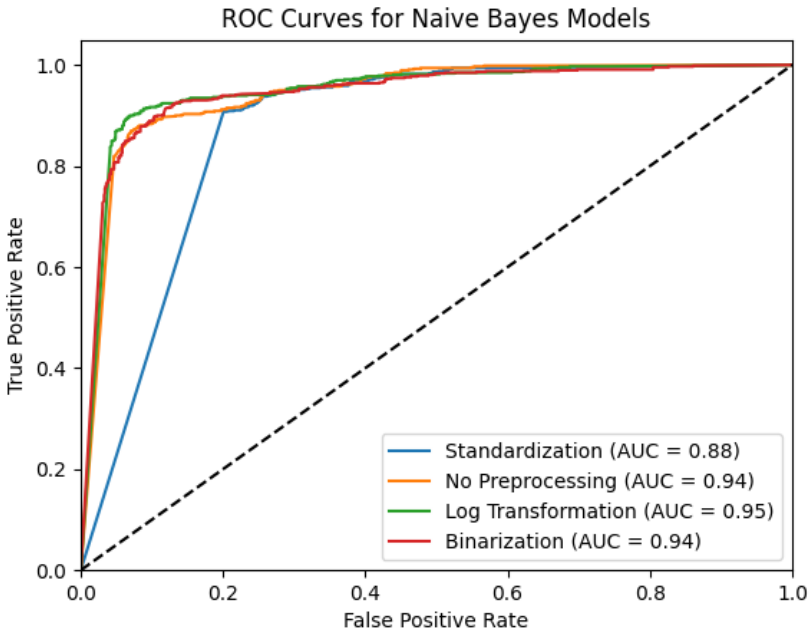

Figure 1: ROC Curves for Naive Bayes Models

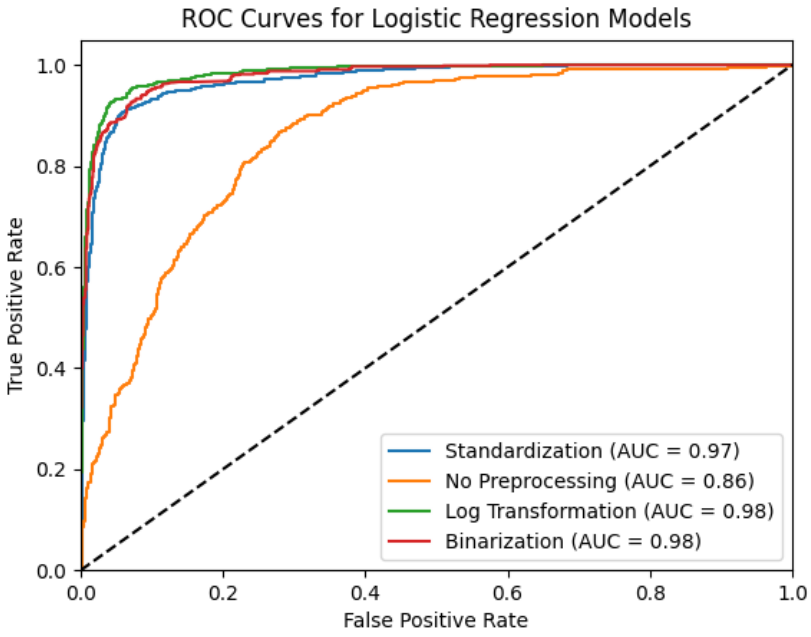

Figure 2: ROC Curves for Logistic Regression Models

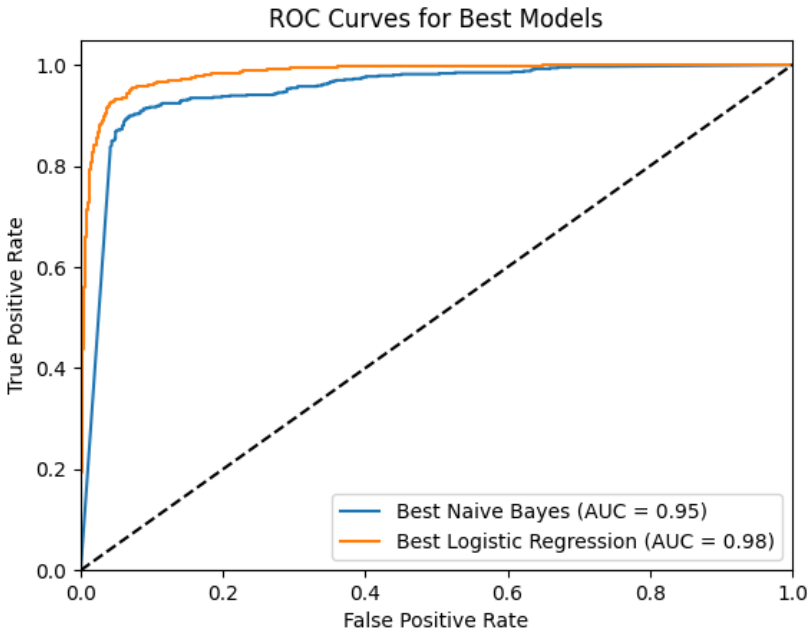

Figure 3: ROC Curves for Best Models

2(f)

2(g) **Effect of Preprocessing on Models:**

- **Naive Bayes:** - Standardization: This preprocessing step resulted in moderate performance with a test accuracy of 0.810 and a test AUC of 0.875. - No Preprocessing: This step yielded slightly better results with a test accuracy of 0.817 and a test AUC of 0.942. - Log Transformation: This step provided the best performance for Naive Bayes with a test accuracy of 0.815 and a test AUC of 0.948. - Binarization: This step also performed well with a test accuracy of 0.800 and a test AUC of 0.945.
- **Logistic Regression:** - Standardization: This preprocessing step resulted in high performance with a test accuracy of 0.919 and a test AUC of 0.970. - No Preprocessing: This step yielded the lowest performance with a test accuracy of 0.742 and a test AUC of 0.860. - Log Transformation: This step provided the best performance for Logistic Regression with a test accuracy of 0.936 and a test AUC of 0.983. - Binarization: This step also performed well with a test accuracy of 0.926 and a test AUC of 0.978.

Comparison of Naive Bayes and Logistic Regression:

- Overall, Logistic Regression outperformed Naive Bayes across all preprocessing steps in terms of both accuracy and AUC.
- The best Naive Bayes model (Log Transformation) had a test AUC of 0.948, while the best Logistic Regression model (Log Transformation) had a test AUC of 0.983.
- Logistic Regression benefited more from preprocessing steps like Standardization and Log Transformation compared to Naive Bayes.
- Naive Bayes showed relatively consistent performance across different preprocessing steps, while Logistic Regression showed significant improvement with appropriate preprocessing.

3 Problem 3

3(a) Given the results from the previous problem, we can see that the Log Transformation provided the best performance for both Naive Bayes and Logistic Regression models.
Therefore, we will use the Log Transformation preprocessing step.