Put Your Code Results Here:

Compute Confusion Matrices

Construct confusion matrices for both logistic and probit regression.

Logistic Regression Confusion Matrix:

[[11619 816]
[1739 2107]]

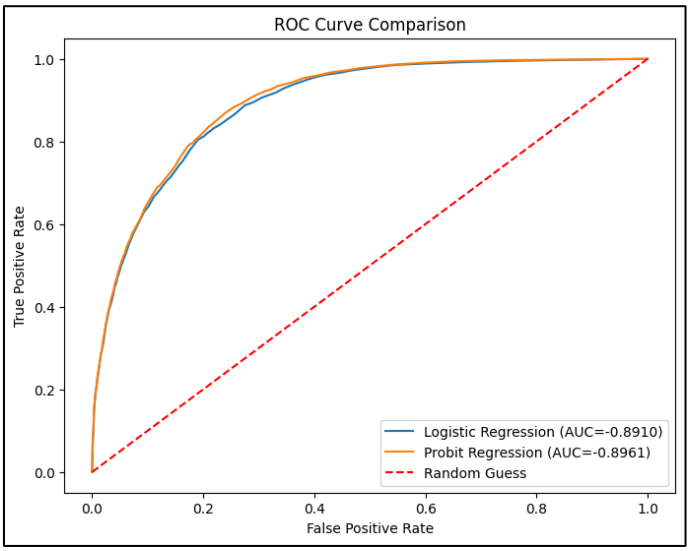
TN: 11619, FP: 816, FN: 1739, TP: 2107

Probit Regression Confusion Matrix:

[[11685 750]
[1776 2070]]

TN: 11685, FP: 750, FN: 1776, TP: 2070

Generate the ROC Curve



Compute AUC

- Logistic Regression AUC = -0.8910
- Probit Regression AUC = -0.8961

Questions (30%)

1. (5%) Show the comparison of the confusion matrices of logistic and probit regression. Do they produce similar results? Why or why not?

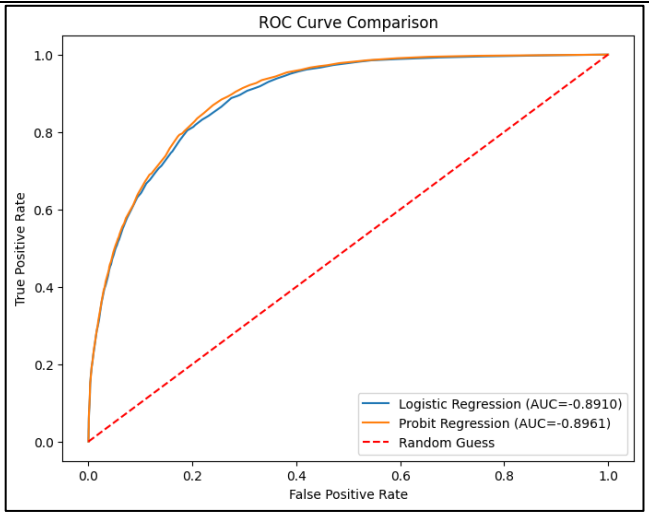
The TP, TN, FP, and FN values of the two are very similar. This may be because they are from the same dataset. So if the characteristics of the two datasets are very different today and the data is very large, you can find obvious differences.

2. (5%) How does the ROC curve and AUC of logistic regression compare to that of probit regression? Are there any key differences? Explain and show a side-by-side plot comparison.

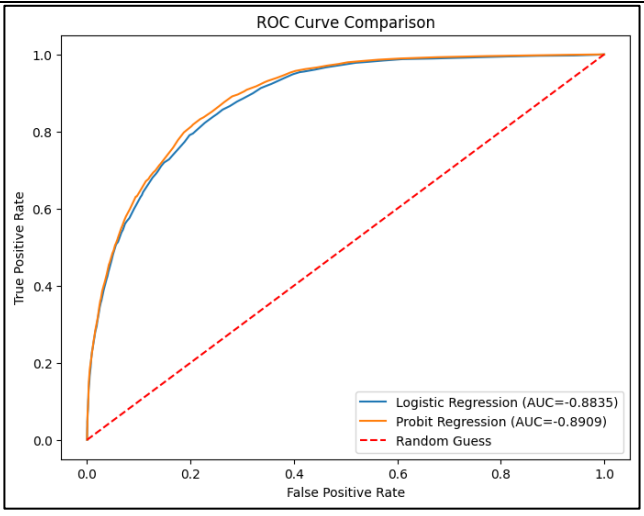
The output AUC value and the plot result are very similar. This may be because the data set does not contain strange data (e.g., special data distribution, out of SD very far)

3. (5%) Discuss the impact of different learning rates and iterations on the convergence of logistic and probit regression. How does hyperparameter tuning affect performance? Provide the results of ROC curves for different hyperparameters.

Different Hyperparameters Test



learning_rate=0.01, iterations=3000



learning_rate=0.001, iterations=10000

If set learning rate is too large, it will jump too fast and easily diverge; if it is too small, the convergence speed will be too slow. When the number of iterations is insufficient, the model will not be fully learned, resulting in a low accuracy rate. At the same time, when the number of iterations is large enough, the model result can produce the best solution, but at the same time, we must be careful not to overfit, otherwise it will be overfitting.

4. (5%) Explain the fundamental differences between logistic regression and probit regression. When might you choose one over the other?

- Logistic Regression's value is between [0,1](Logistic Distribution) and It's Sigmoid Function more usually see apply this function in cancer(medical), engaging filed. Advantage is High efficiency, Stable convergence and more common in practice
- Probit Regression is using Normal Distribution, which application in specific fields, Econometrics, sociology

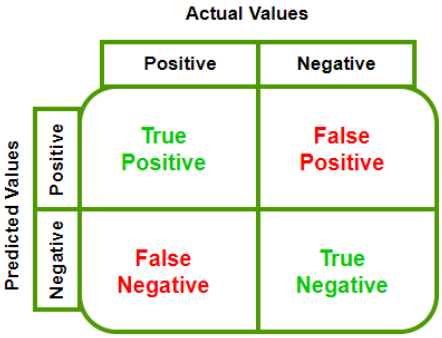
5. (5%) Discuss their activation functions: sigmoid for logistic and normal CDF for probit. How do these functions influence decision boundaries?

Both functions convert the real number domain to the probability domain of (0, 1) and Both exhibit linear decision boundaries.

- Logistic Regression (Sigmoid)：is smoother and less sensitive to both ends and is less sensitive to special extreme values or outliers in the data.
- Probit Regression (Normal CDF)：is more consistent with the data distribution when the data meet the normality assumption, but it is slightly more sensitive to special extreme values.

6. (5%) Define and explain the significance of Confusion Matrices, ROC Curves, and AUC in evaluating classification models. How do they contribute to model selection?

Confusion Matrices:



Provide detailed error classification information for the model. Like can compute Accuracy, Recall, Precision, etc,.

ROC:

The ROC curve is drawn by FPR and TPR under different classification thresholds.

It Show the model's ability to classify positive and negative values. And a higher ROC curve means that the model has better classification performance at different thresholds.

AUC:

Which is the area under the ROC curve ranges from 0.5 to 1.0 (1.0 is the best perfect classification). It use only one single number quickly measures the overall performance of the model. More closer it is to 1, means the stronger the model's ability to distinguish between positive and negative samples.