**Executive Summary:**

The received data has 100 features including numerical and categorical. It was observed that the data is imbalanced and the class weights need to be considered later in the modeling section.

The logistic regression was suggested to be used for this study which is a GLM. I choose the gradient boosting classifier as Non-GLM model. the reason I selected the XGboost are as follow:

**XGBoost**

Pros:

1) XGBoost is a decision tree-based algorithm and it takes care of imbalanced data itself.

2) XGboost is a well-known algorithm and it often outperforms other algorithms in competitions.

Cons:

1. Encoding categorical features could affect the XGBoost algorithm performance

**Logistic Regression**

Pros:

1. It is a white box. I mean we can interpret the coefficients and like any linear equation we can check out the importance of each feature just by checking the coefficients out.

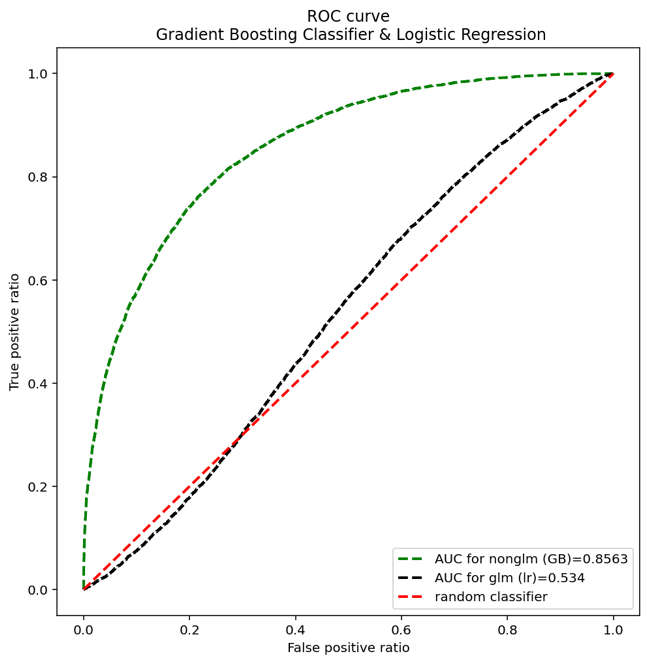
Cons:

1. It needs non-collinearity between features. The Lr can find complicated the relationship between features and output while other ML algorithms like MLP and XGBoost are incredibly tough.

Based on my study I think the XGBoost results for test data set are more reliable. The estimates for XGBoost and Logistic regression models and associated AUCs are seen below.

|  |  |
| --- | --- |
| ML algorithm | AUC |
| XGBoost | 0.8563 |
| Logistic regression | 0.5340 |

Additionally, the ROC curve is ideal for comparing classification algorithms. From below curve, it is seen that XGBoost outperforms the LR algorithm.



The above plot is a what I would show a business partner to show how good is the model and to compare the performance of both models. The XGboost would outperform the LR model, having larger area under the curve.

Bests

Amin