Pros and Cons of the two models

Logistic regression: Logistic regression is a linear model for classification. It is simple to implement, efficient and fast

Pros

- Simple algorithm that is easy to implement, does not require high computation power.
- Performs extremely well when the data is linearly separable.
- Less prone to over-fitting, with low-dimensional data.

Cons

- Poor performance on non-linear data
- Poor performance with highly correlated features.

Gradient boosting

The gradient boosted regression tree is an ensemble method that combines multiple decision trees in series to create a more powerful model. Each tree tries to correct the mistakes of the previous one using the parameter learning rate.

Pros

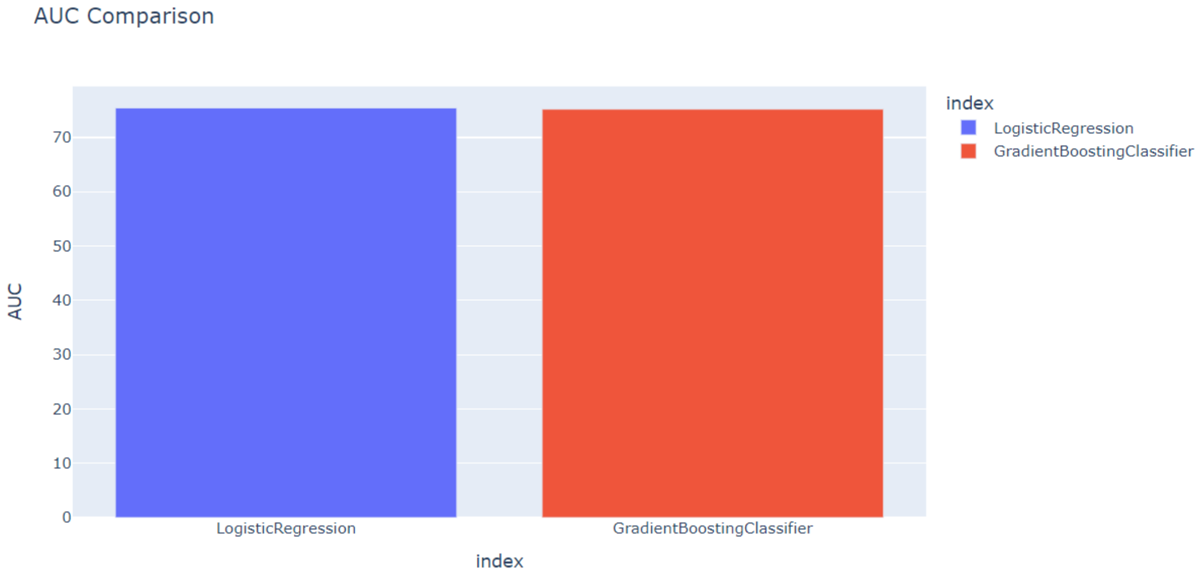
- Less feature engineering required (No need for scaling, normalizing data, can also handle missing values well.
- Fast to interpret
- Outliers have minimal impact.
- Good model performance
- Less prone to overfitting

Cons

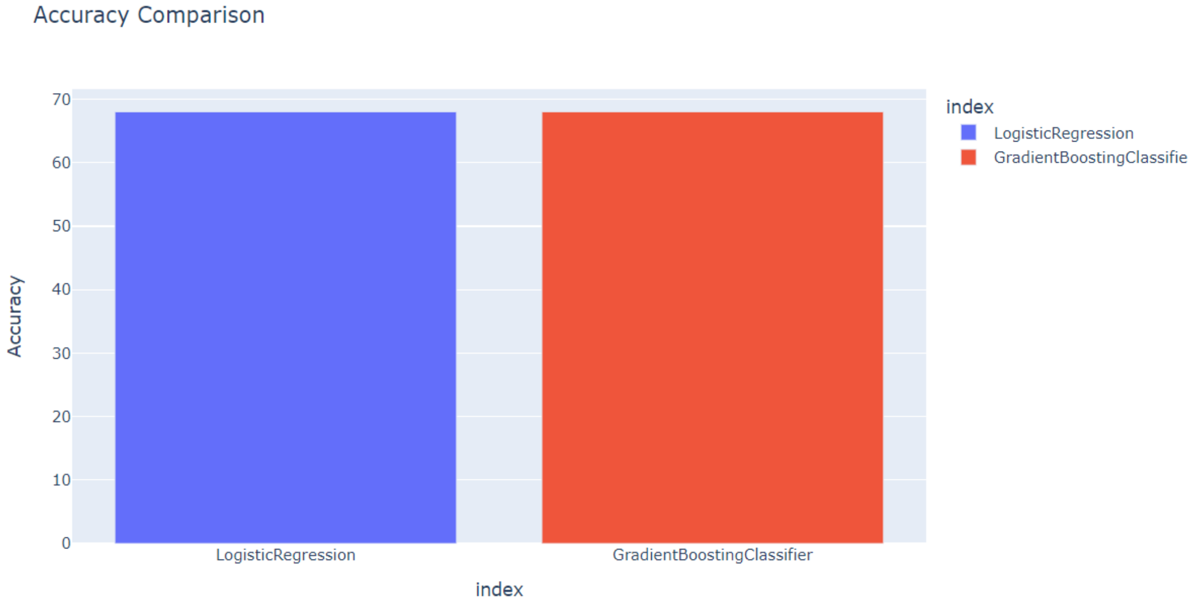
- Overfitting possible if parameters not tuned properly

Graphical comparison of the two models

AUC: It represents the area under an ROC curve. An ROC curve is a plot of false positive rate on the x-axis and true positive rate on the y-axis.



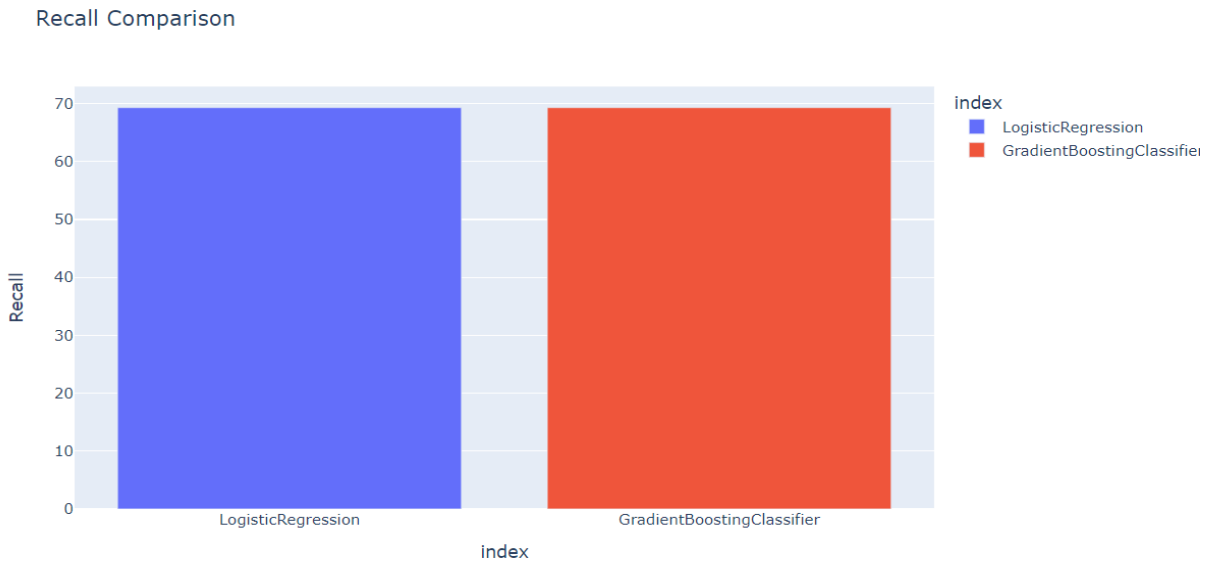
Accuracy: Accuracy is the total number of correct predictions divide by the total predictions



$Accuracy = (TP+TN) / (TP+TN+FN+FP)$

Recall

Recall or sensitivity: Recall is the true positive predictions out of the total number of actual positive $Recall = TP/ (TP +FN)$



Comparison

- Gradient boosting classifier often does not work well on high-dimensional sparse data unlike logistic regression which works well with sparse data.
- Gradient boosting often require a longer time to tune compared to logistic regression which is fast to train and predict
- Logistic regression tends to underperform when there are multiple or non-linear decision boundaries unlike gradient boosting machines which is reliable and ability to handle nonlinearities in a dataset
- Gradient boosting works well without scaling and can handle outliers unlike logistic regression. Logistic regression model can be affected by outliers.

Which will perform better on test set

From both models, gradient boosting performs better than logistic regression in terms of AUC. They are immensely powerful. They are made up of multiple decision trees connected in series. Each decision tree corrects the error of the previous decision tree, often work well without heavy tuning of the parameters, and do not require scaling of the data. They can handle outliers. Gradient boosting will perform better on the test set compared to logistic regression