Higgs Boson Challenge

Team

Amey Bhatuse

Swarnava Bhattacharjee

Venkatesh BY

Univ.AI AI 1 cohort 5

Problem Statement

The events detected by the detector at the LHC in large majority, represent known processes(called background) that are mostly produced by the decay of exotic particles.The search for new physics provided by unknown events (called signal) is difficult because these processes are rare.

Goal

Classification of events as signal and background dataset along with predicting the weights for the simulator and calculating the AMS(Approximate Median Significance)

Approach

To train a classifier on a simulated set of signal and background events that will be able to classify the events as signal and background and train a regression model for predicting the weights

DATASET

Simulated events from the official ATLAS full detector simulator containing the signal and background events

Types of variables

- -Primitive (raw quantities measured by detector)
- -Derived (quantities computed from raw quantities)

Outliers in data:

Variables are floating point and have value -999.0 if they are indicated as may be undefined

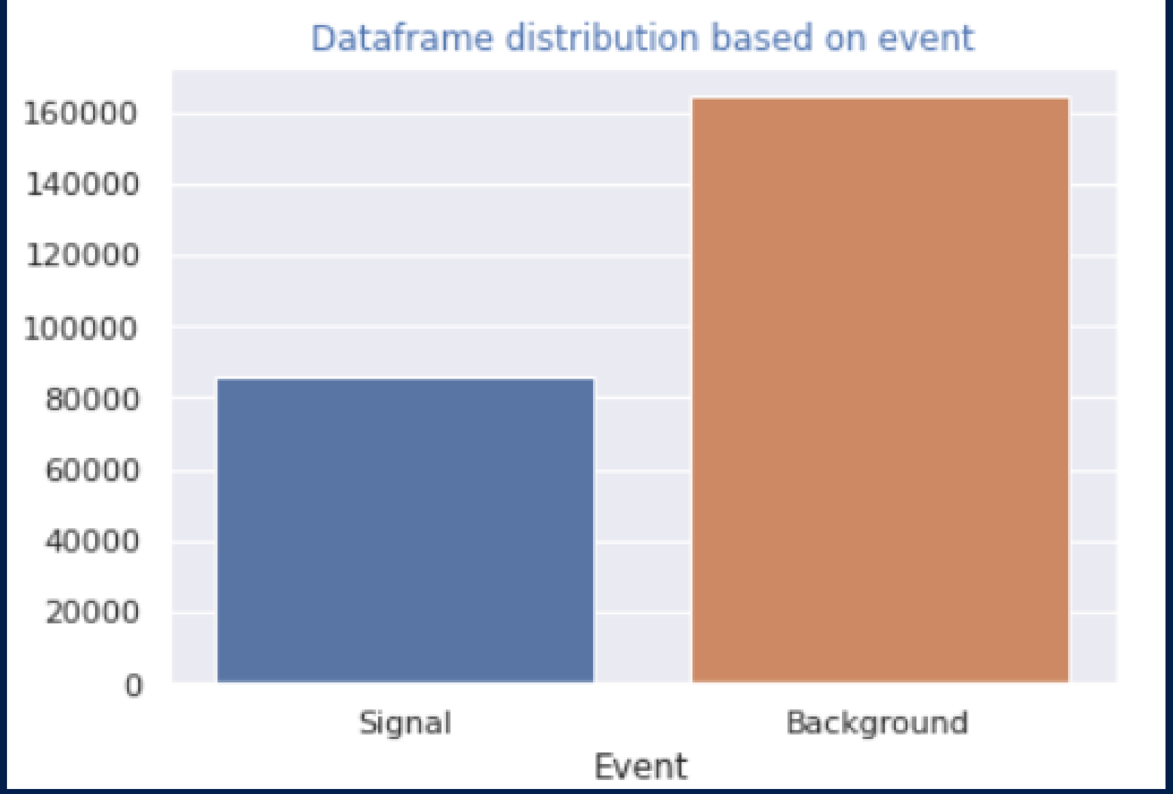
Samples

Training data samples = 250000
Testing data samples = 550000

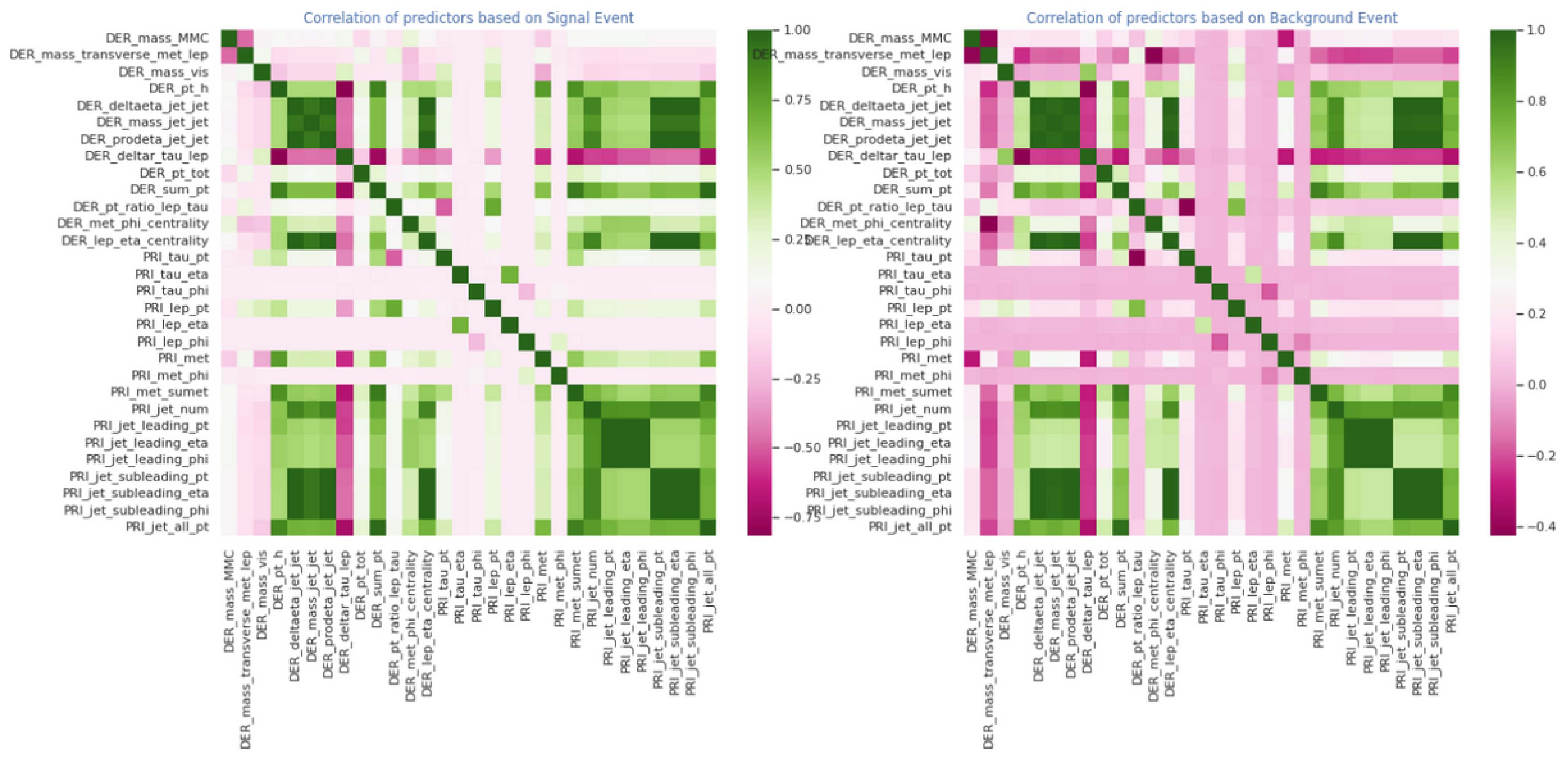
Weights- describe the way the simulator works , so not to be used as features

EDA(EXPLORATORY DATA ANALYSIS)

Separating the dataset shows that there is a significant class imbalance(~1:3) between signal and background events



Correlation of predictors based on signal and background events

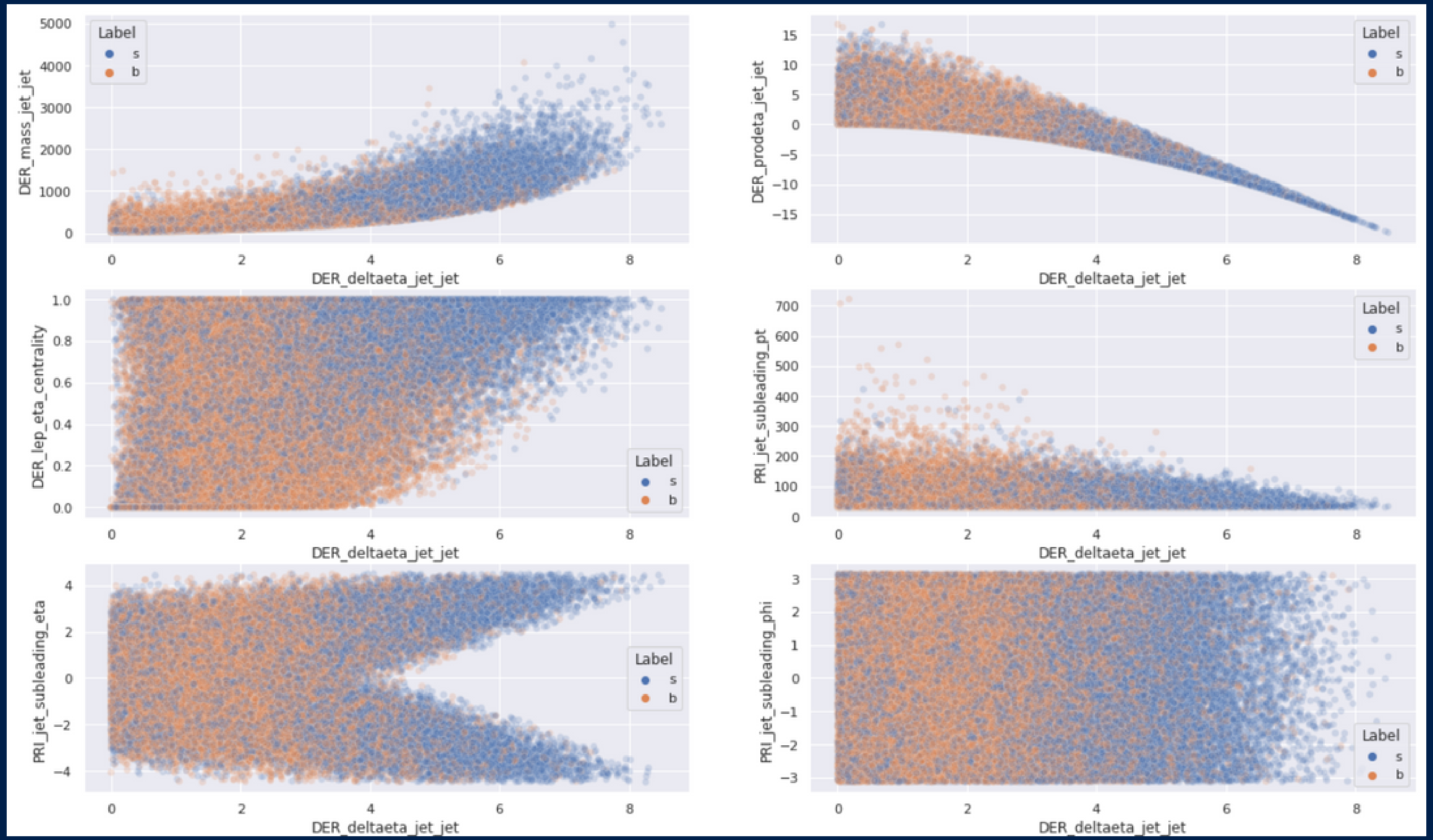


DATA CLEANING

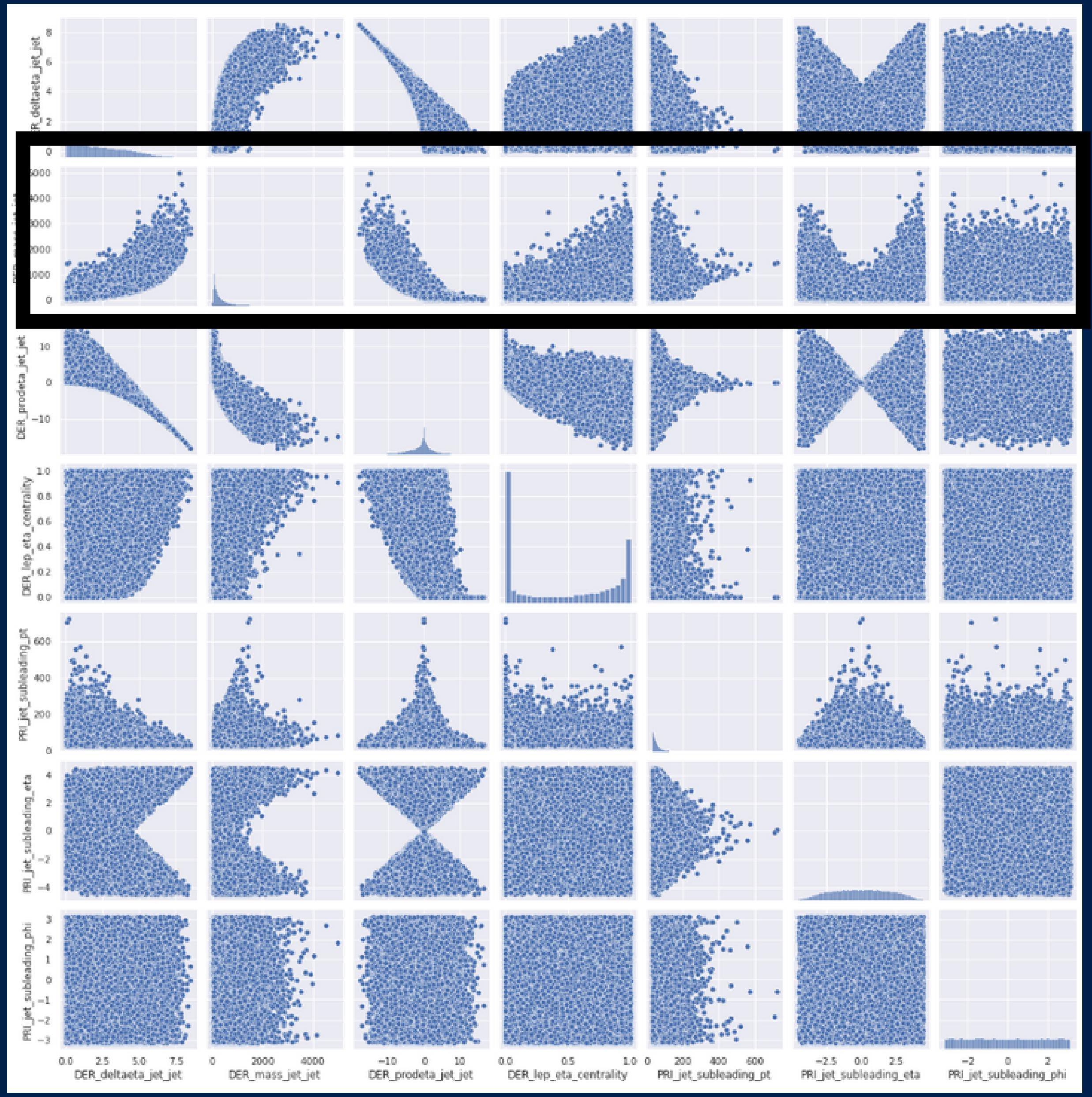
-We are dropping those columns that have more than 60% outliers, i.e. have value -999.0 for 60% of the events

We trained the baseline model along with the dropped predictors and found no improvement, thus they are insignificant

-We standardize predictor variables and choose upscaling(SMOTE) as a method for imbalance treatment



Correlation between dropped predictors



Collinearity between predictors

Model

We train four different models and try to tune their hyperparameters

-Logistic Regression(Baseline Model)

-Decision Trees

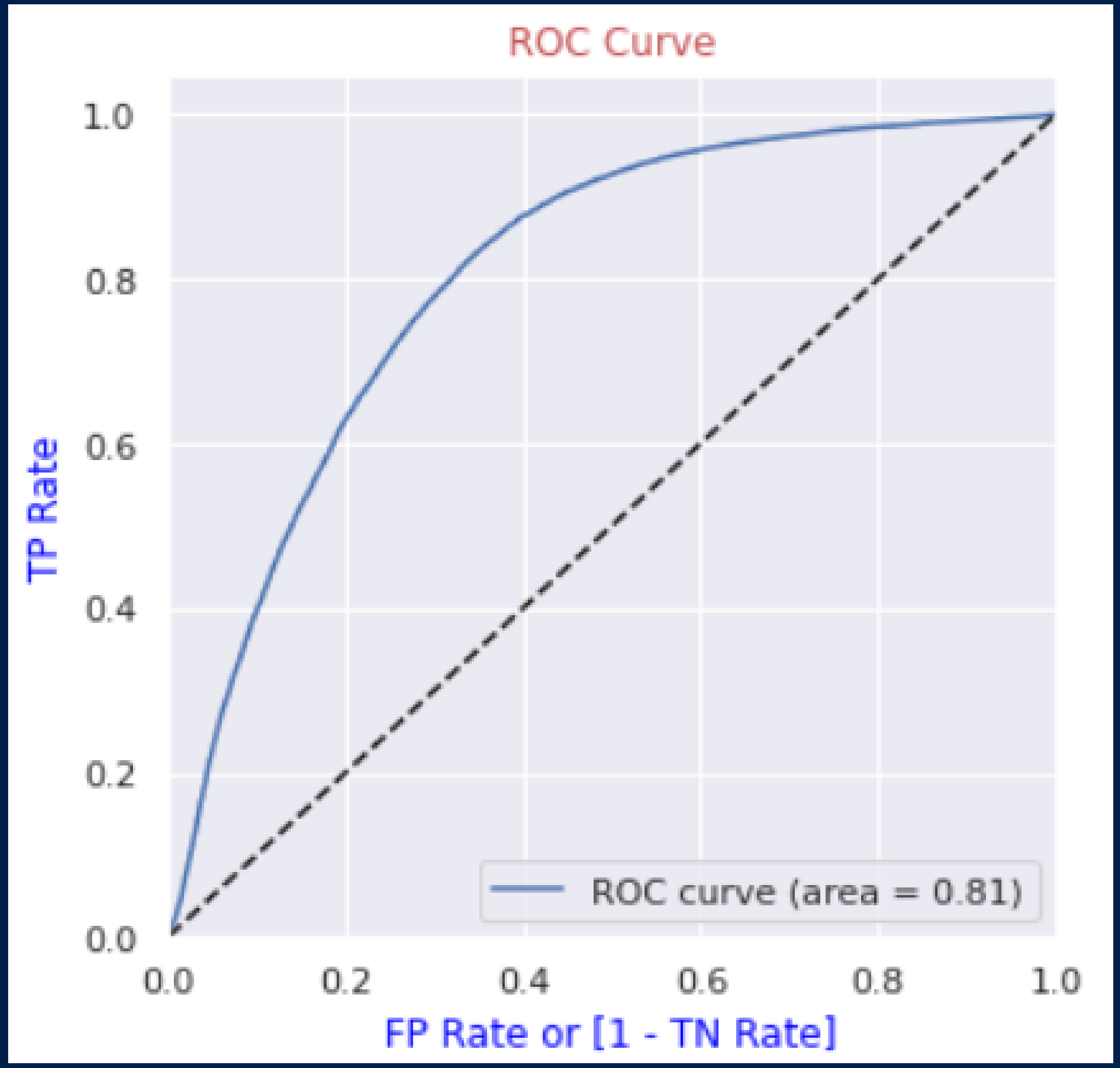
-Random Forest(Ensemble Model)

-AdaBoost

1.**Logistic Regression**(Baseline Model)

-F1 score = 0.75

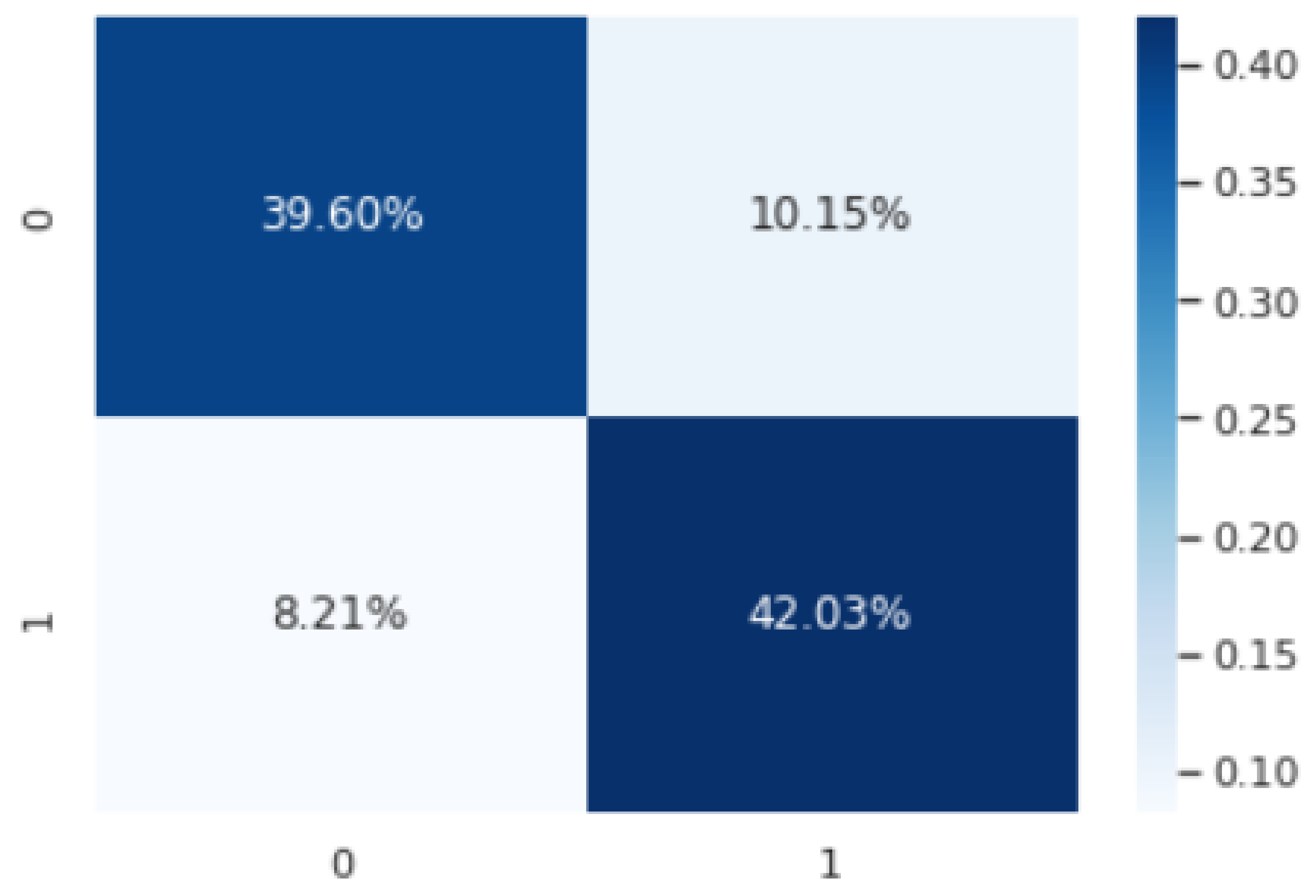
-Area under ROC curve = 0.81



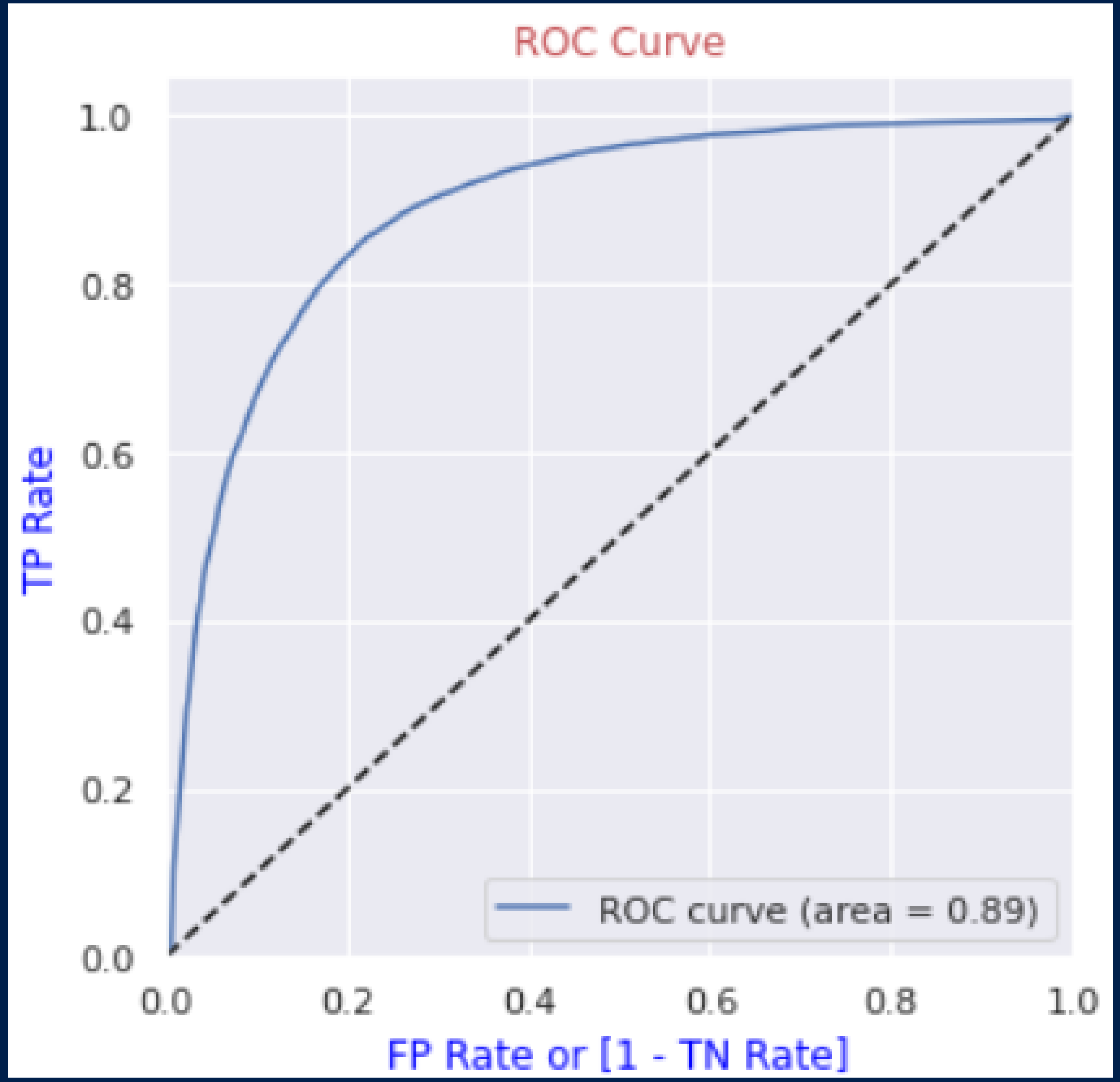
ROC curve for Logistic Regression

2.Decision Tree

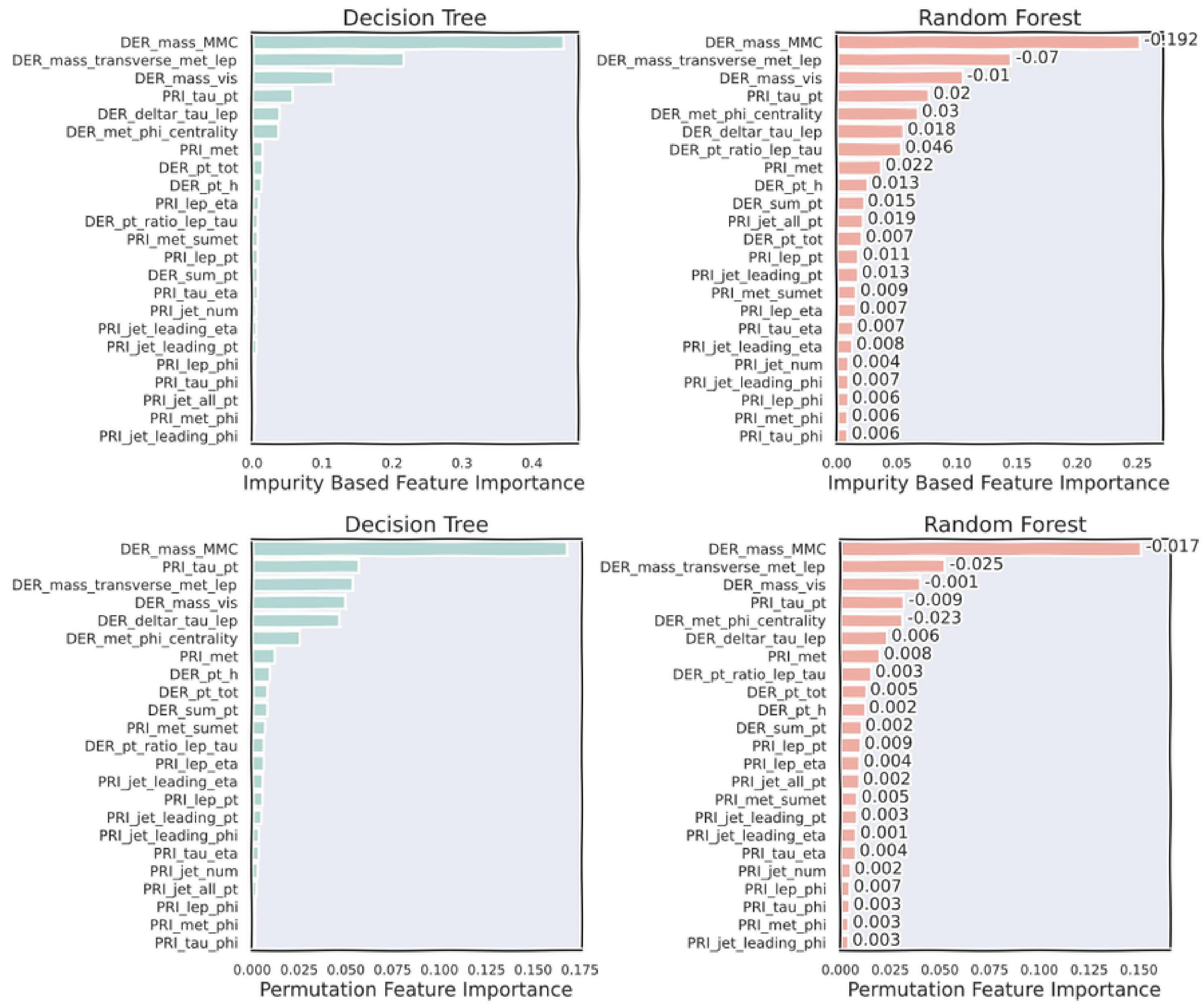
-F1 score = 0.82
-Area under ROC curve = 0.89

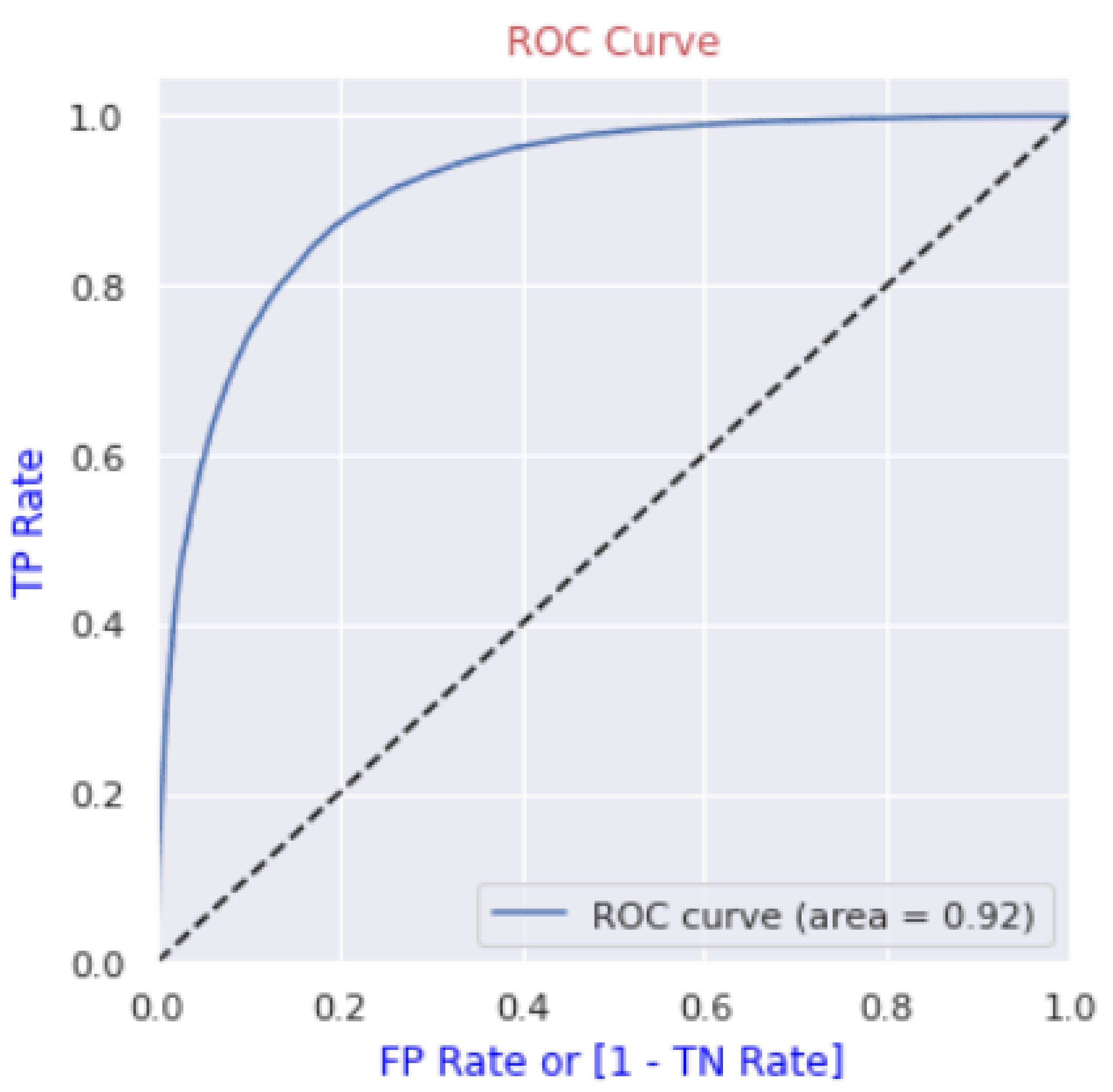


After Hyperparameter tuning with GridSearch we found 11 as the ideal depth with minimum samples as 5



3.Random Forest (Ensemble Model) with Grid Search





Most of the features are unimportant

Best F1 score = 0.838

Best Configuration:

Maximum Depth: 15

Minimum Samples: 5

Number of trees: 50

Best AUC for RUC curve: 0.92

4.AdaBoost

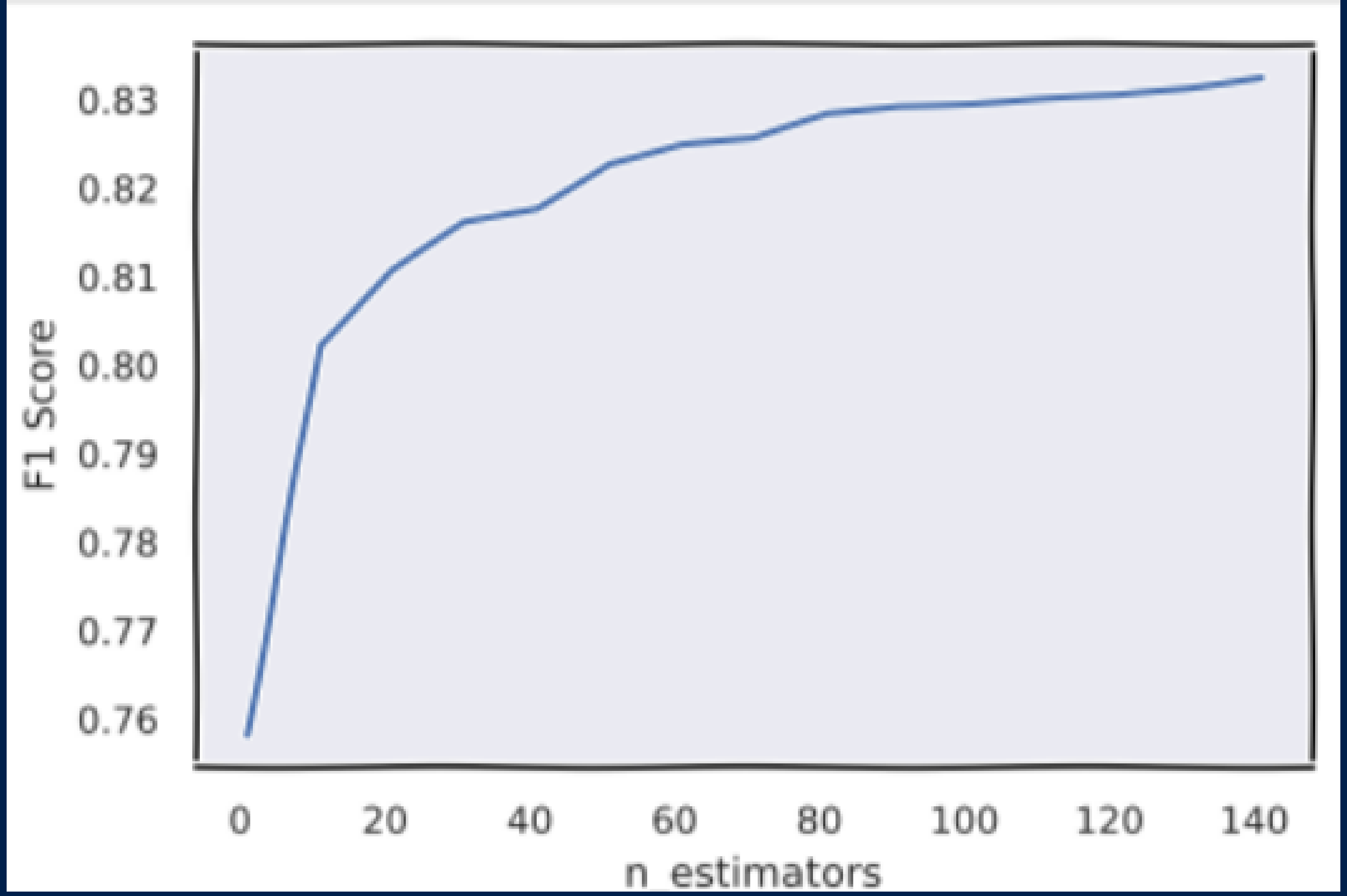
Decision Tree with maximum depth of 2 as a weak learner

F1 score for weak learner = 0.758

We encounter an almost plateau region for the F1 score when the number of estimators become more than 100

f1_score after boosting the weak learner = 0.83

AUC for ROC = 0.91



Experimention/Observation

- Label

ROC(AUC) for ADABoost with Hyperparameter tuning is 0.91 which is nearly same for hyperparameter tuned Random Forest model. However, the computational cost for AdaBoost is considerably higher than Random Forest

- Weights

Linear Regression(Baseline Model)
-R2 score on Train data : 0.4616

Decision Tree Regressor
-R2 score on Train data : 0.6513

Model Selection

- Label

We choose Random Forest model with AUC for ROC as 0.92 as our final model for the classification of signal and background events

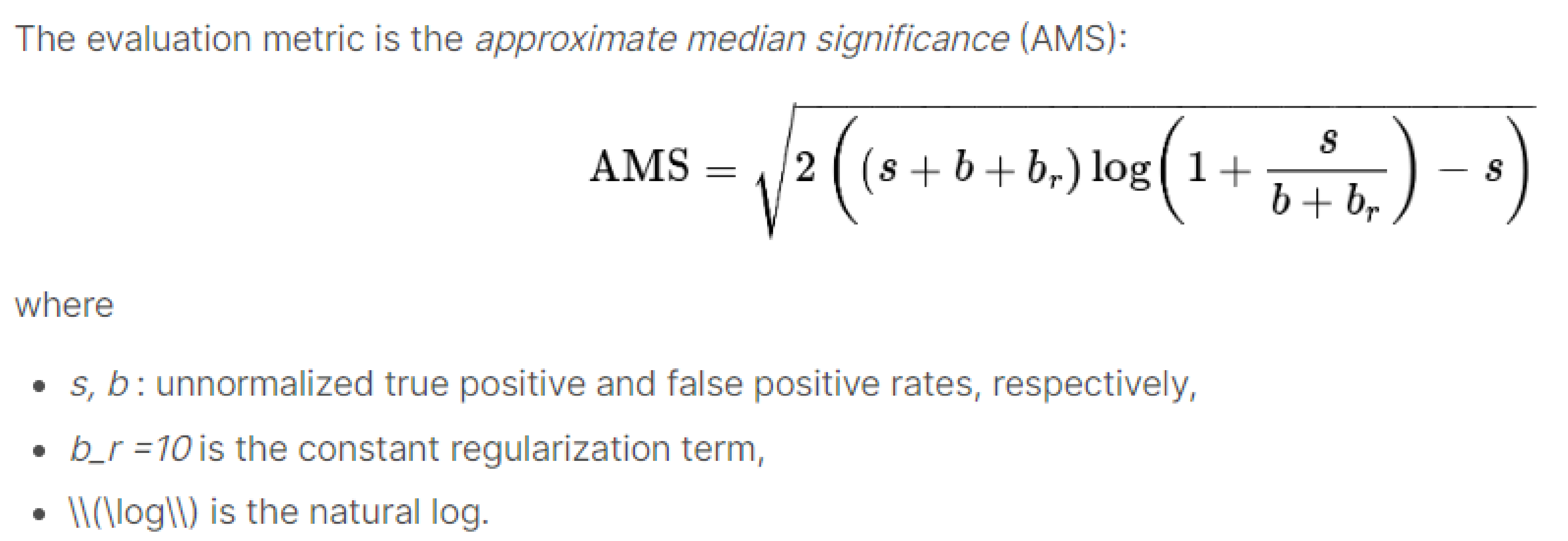
- Weights

We choose the Decision Tree Regressor with R2 score 0.6513 as the best model for predicting weights

Prediction on Unseen data

- We performed the same preprocessing on the test dataset as the train dataset
- We classified the events as signal and background events for the test dataset

AMS(Approximating Median Significance)



The evaluation metric is the *approximate median significance* (AMS):

$$\text{AMS} = \sqrt{2\left((s+b+b_r)\log\left(1+\frac{s}{b+b_r}\right)-s\right)}$$

where

- *s, b* : unnormalized true positive and false positive rates, respectively,
- *b_r =10* is the constant regularization term,
- \\(\log\\) is the natural log.

We calculated AMS from predicted label and merged with the test dataset and submitted the .csv file as output

Thank you!