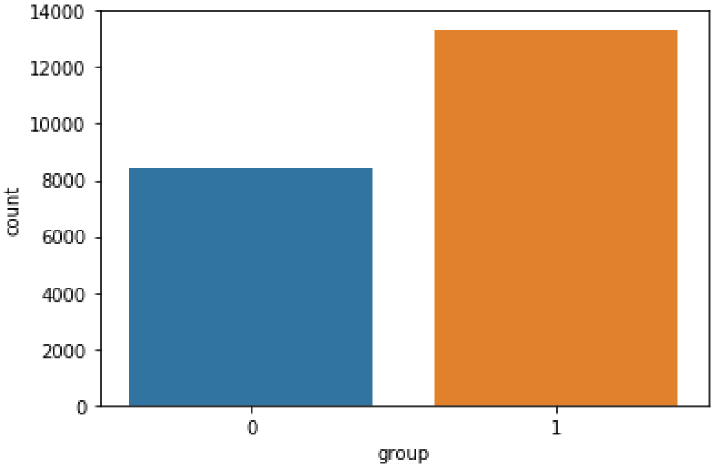
Assignment 3

1. Selection of Models

First, we performed a brief exploratory analysis on the data and found that there was a class imbalance in the dataset. This led us to choose Machine Learning algorithms that could deal well with imbalanced classes, namely Gradient Boosting, Random Forest and Support Vector Machines



1.1 Gradient Boosting

Boosting is a sequential ensemble learning algorithm which converts weak learners into strong learners. Subsequent learners learn from misclassified instances of previous learners. This is done by iteratively fitting trees to the current residuals of the model, until fitting can no longer occur without resulting in overfitting. At each iteration, a subsample of the original dataset is randomly selected without replacement. In Gradient Boosting, a loss function is optimised using Gradient Descent

1.2 Random Forest

Random forest is a parallel ensemble learning algorithm which from randomly selected subset of training data creates a set of decision trees. Random Forest without increasing the variance reduces the correlation between the trees. This is done by performing the random selection of the input variable in the tree-growing process. The final class of the test object is created by aggregating the votes from different decision trees.

1.3 Support Vector Machines

The Support Vector Machine (svm) separates the data into classes by finding a hyperplane that best divides the data - a line in 2D and a plane in 3D. support vectors are the data points which are nearest to the hyperplane, support vectors can alter the position of the hyperplane. When the support vector classifier is combined with the non-linear kernel then this results in support vector machine. SVM is a penalization method i.e., svm gives zero penalty to the points which are inside their margin, and a linear penalty to the points which are on the wrong side of margin and far away

2. Method

Scikit-learn’s implementation of Recursive Feature Elimination Cross Validation and Grid Search Cross Validation was used to find the best parameters and attributes for each of the three Machine Learning algorithms. The number of folds was set to 5, and the scoring metric was set to “average_precision”, which makes use of AUPRC as the scoring metric.

2.1 Range of hyper-parameters for each algorithm

Gradient Boosting	`{`
	`"n_estimators":[50,100,200,300],`
	`"max_depth":[3,4,5],`
	`"max_features":["log2","sqrt"],`
	`}`

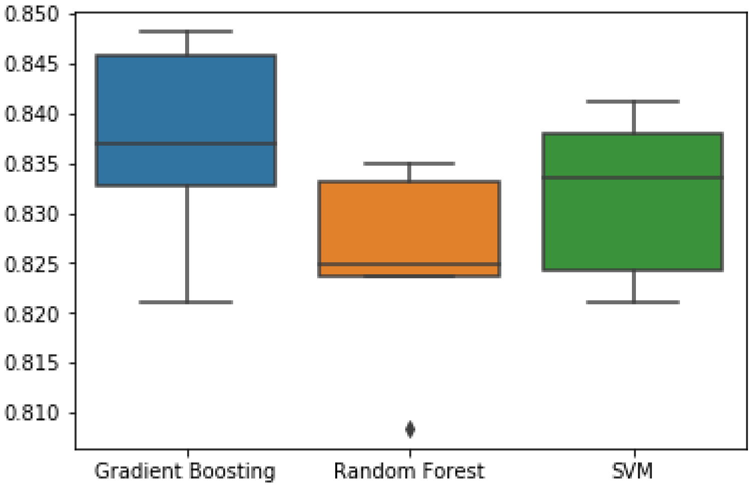
Random Forest	`{`
	`"max_depth":[2,3,5],`
	`"n_estimators":[50,100,200,300],`
	`"criterion":["gini","entropy"]`
	`}`

SVM	`{`
	`"C":[10,50,100],`
	`"gamma":[0.001,0.0001],`
	`"kernel":['linear','rbf']`
	`}`

3. Results

The mean of AUPRC across folds was used to determine the algorithm which performed the best. Based on the results below, Gradient Boosting performed better than Random Forest and SVM, and was used for the prediction of the test data.

	Mean	Standard Deviation
Gradient Boosting	**0.837**	**0.010**
Random Forest	0.825	0.009
SVM	0.832	0.008



0.850
0.845
0.840
0.835
0.830
0.825
0.820
0.815
0.810
Gradient Boosting
Random Forest
SVM