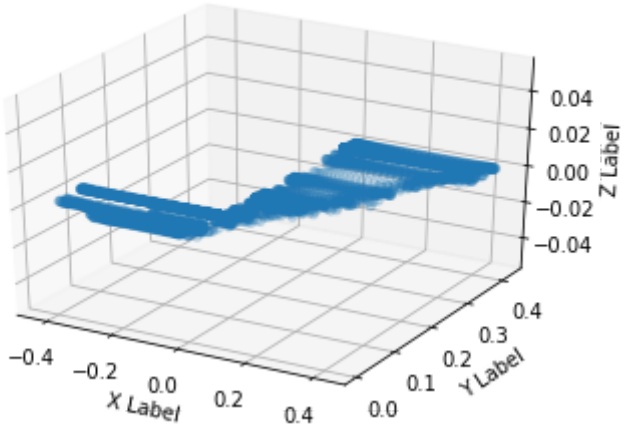


Conclusion and Inference:

After fitting Linear Regression, Kernal Regression and Random Forest models I have plotted the colorbar and Scatter plots. With the Mean Square error and testing accuracy values i conclude that the Kernel regression model fits better and gives better results.

Problem 2: Click Through Rate Prediction

Part 2.1: SVM

Let's Start with SVM. Please use svm.LinearSVC, Let's try to add balanced weight to handle the class-imbalance issue.

1. please compute the precision/recall, f1-score, and confusion matrix.
2. Please run the algorithm for multiple times and observe the result.

Let's Start with SVM. Please use svm.LinearSVC, Let's try to add balanced weight to handle the class-imbalance issue. (a) please compute the precision/recall, f1-score, and confusion matrix. (b) Please run the algorithm for multiple times and observe the result.

Iteration-1 Confusion Matrix

	precision	recall	f1-score	support
0	0.92	0.52	0.66	2500
1	0.24	0.76	0.37	500

avg / total 0.80 0.56 0.61 3000

The Mathews Corelation coefficient is 0.208997129935

Confusion Matrix

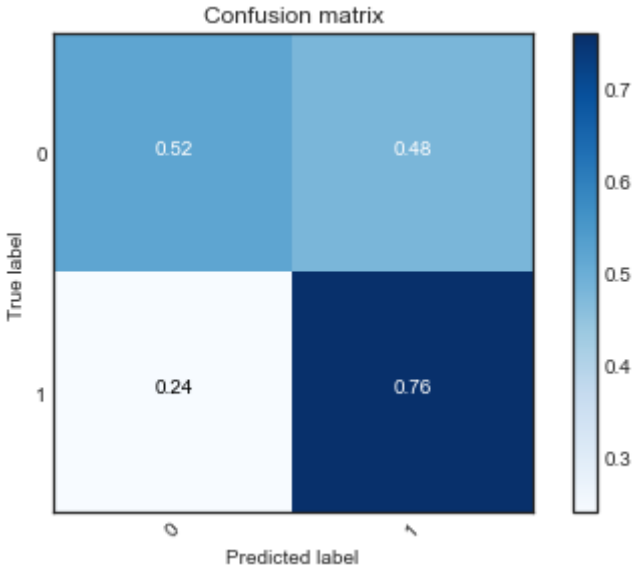
[[1300 1200]

[120 380]]

Normalized confusion matrix

[[0.52 0.48]

[0.24 0.76]]



Iteration-2 Confusion Matrix

	precision	recall	f1-score	support
0	0.91	0.52	0.66	2500
1	0.24	0.74	0.36	500

avg / total 0.80 0.56 0.61 3000

The Mathews corelation coefficient is 0.192798201214

Confusion Matrix

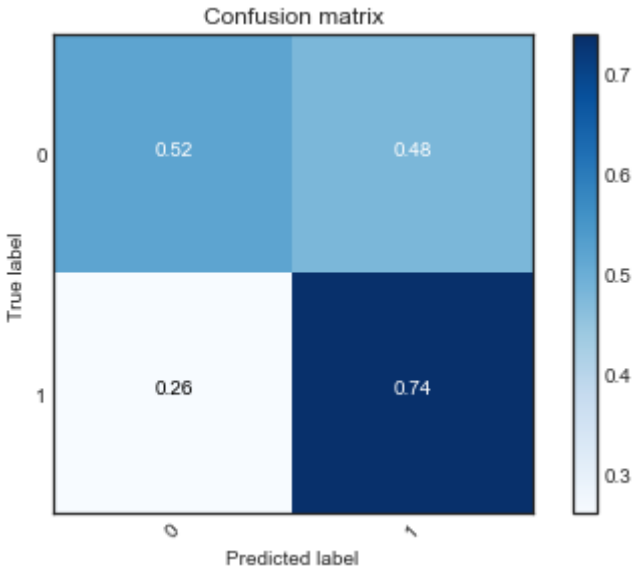
[[1301 1199]

[131 369]]

Normalized confusion matrix

[[0.5204 0.4796]

[0.262 0.738]]



Iteration-3 Confusion Matrix

	precision	recall	f1-score	support
0	0.91	0.53	0.67	2500
1	0.24	0.73	0.36	500

avg / total 0.80 0.56 0.62 3000

The Matthews corelation coefficient is 0.194496926724

Confusion Matrix

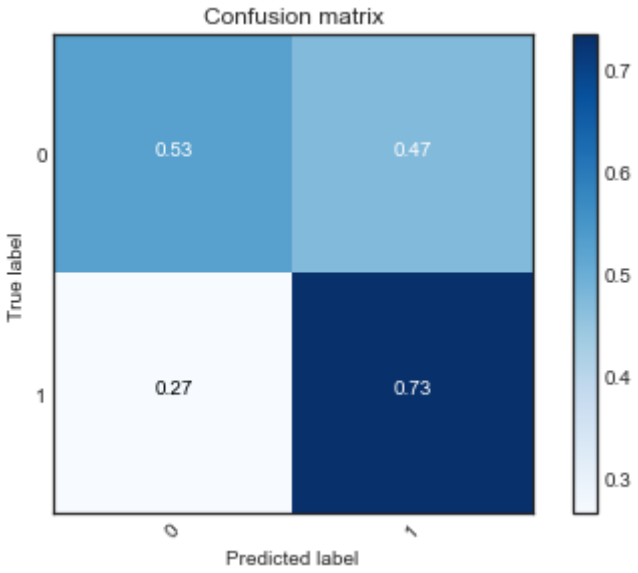
[[1317 1183]

[133 367]]

Normalized confusion matrix

[[0.5268 0.4732]

[0.266 0.734]]



Part 2.2 Regularized SVM

1. Let's try to add penalty, please exlpore the use of the 'l1' and'l2' penalty in Scikit-learn, Please also use cross validation to select the best tuning parameters C.
2. please compute the precision/recall, f1-score, and confusion matrix for 'l1' and 'l2' model with the best tuning paramter C.

L1 Regularization

best parameters: {'C': 0.01} ,testing accuracy: 0.827666666667

	precision	recall	f1-score	support
0	0.92	0.52	0.66	2500
1	0.24	0.77	0.37	500

avg / total 0.81 0.56 0.61 3000

The Matthews Corelation coefficient is 0.216210742298

Confusion matrix

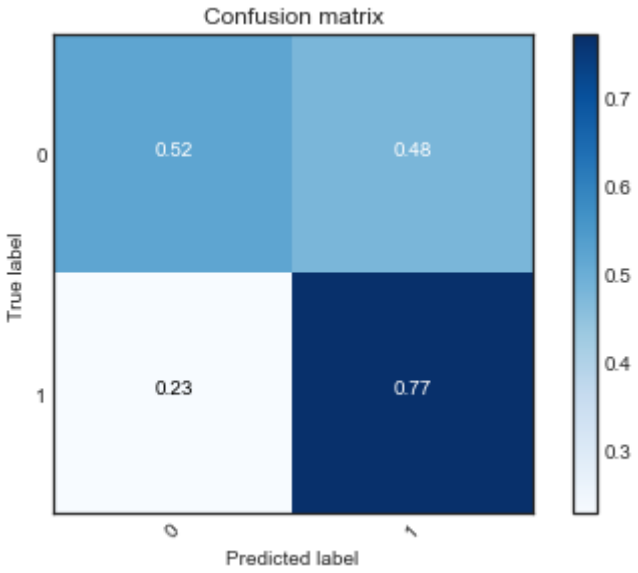
[[1299 1201]

[115 385]]

Normalized confusion matrix

[[0.5196 0.4804]

[0.23 0.77]]



L2 Regularization

best parameters: {'C': 0.01} ,testing accuracy: 0.827

	precision	recall	f1-score	support
0	0.84	0.98	0.90	2500
1	0.40	0.07	0.12	500

avg / total 0.77 0.83 0.77 3000

The Matthews Corelation coefficient is 0.106913664854

Confusion Matrix

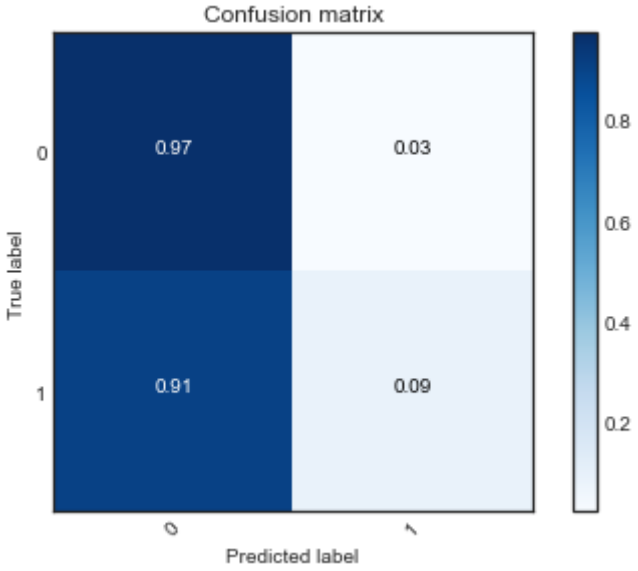
[[2435 65]

[454 46]]

Normalized confusion matrix

[[0.974 0.026]

[0.908 0.092]]



Part 3: Logistic Regression

Please also explore using Logistic Regression on this problem and report the result.

1. Please plot the ROC curve and compute the area under the ROC curve. (You don't need to explore the use of penalty since the cross validation can be very slow)
2. Please plot the precision recall curve and compute the average precision
3. Please compute the F1-score and confusion matrix.

Confusion matrix

[[2434 66]

[454 46]]

Normalized confusion matrix

[[0.9736 0.0264]

[0.908 0.092]]

Area under the ROC curve : 0.699238

	precision	recall	f1-score	support
0	0.84	0.97	0.90	2500
1	0.41	0.09	0.15	500

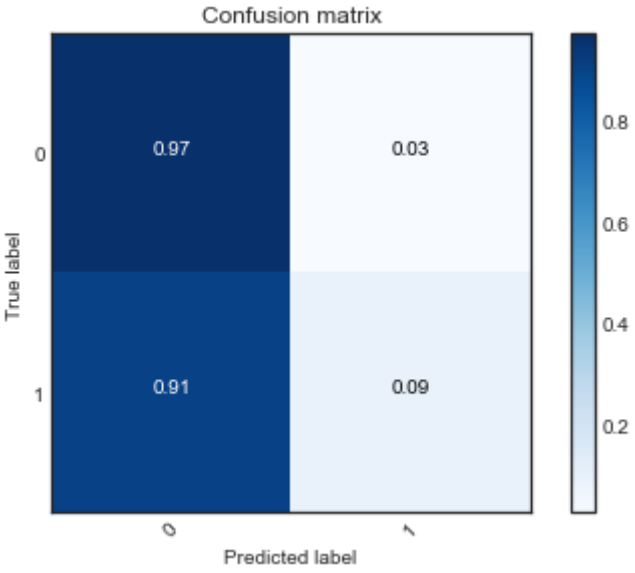
avg / total 0.77 0.83 0.78 3000

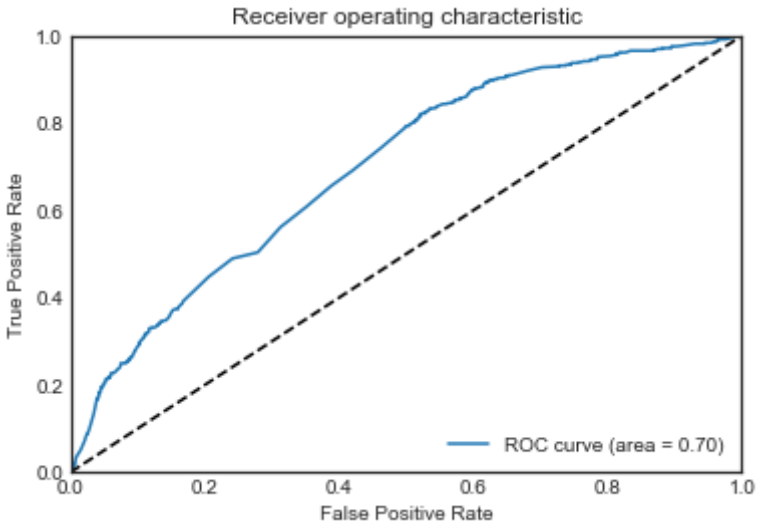
Matthews Correlation Coefficient: 0.12895887627

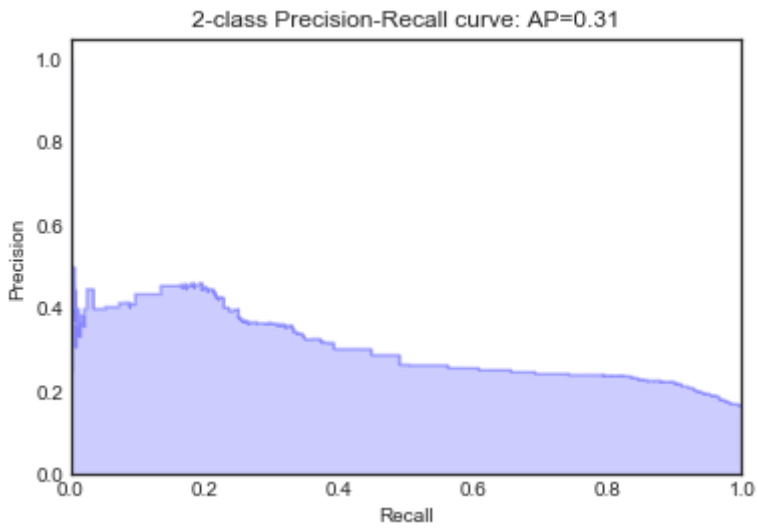
Average precision-recall score: 0.31

Text(0.5,1,'2-class Precision-Recall curve: AP=0.31')

The confusion matrix, ROC Curve and Precision recall curve is plotted below:







Part 4: Random Forest

Please also explore using Random Forest on this problem and report the result.

1. Please use cross-validation to select the best tuning parameters
2. Please plot the ROC curve and compute the area under the ROC curve.
3. Please plot the precision recall curve and compute the average precision
4. Please compute the F1-score and confusion matrix

Accuracy:0.8195

Best parameters:{'max_depth': 16, 'max_features': 0.3, 'n_estimators': 32}

Confusion Matrix

[[3230 104]

[591 75]]

Normalized confusion matrix

[[0.96880624 0.03119376]

[0.88738739 0.11261261]]

Area under the ROC curve : 0.686444

	precision	recall	f1-score	support
0	0.85	0.97	0.90	3334
1	0.42	0.11	0.18	666

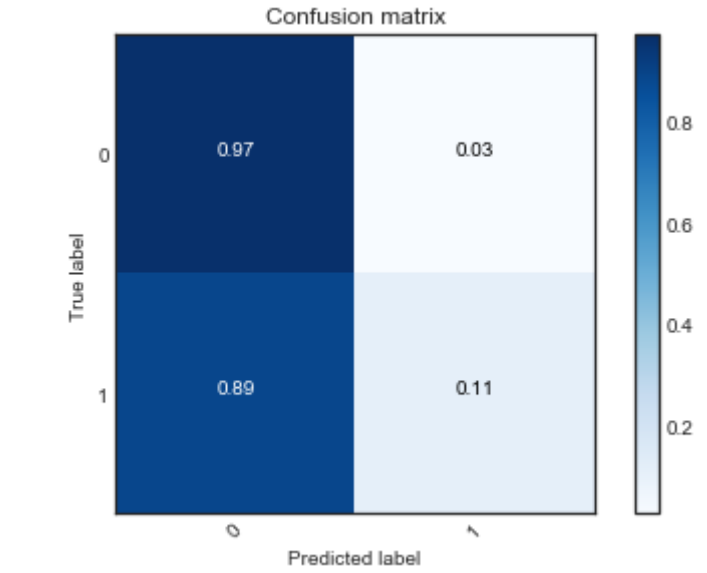
avg / total 0.77 0.83 0.78 4000

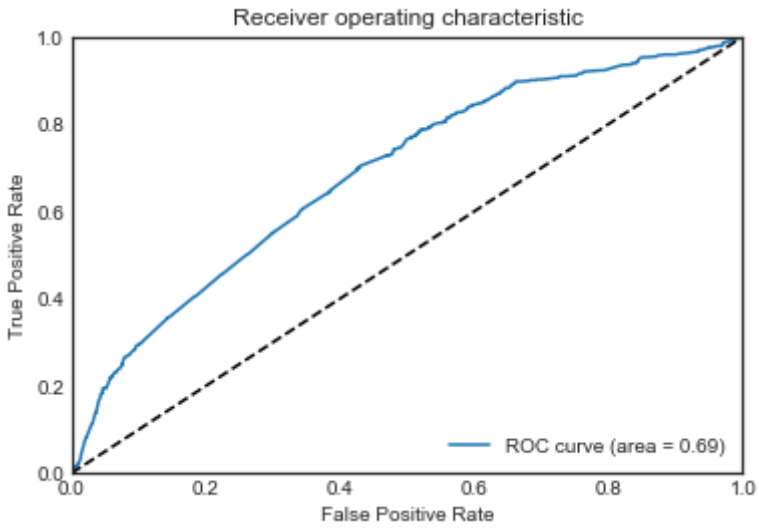
The Matthews corelation coefficient is: 0.146699910598

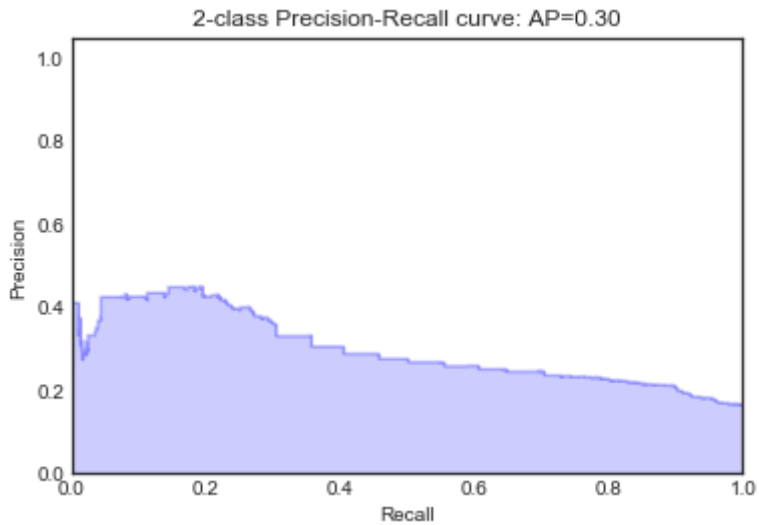
Average precision-recall score: 0.30

Text(0.5,1,'2-class Precision-Recall curve: AP=0.30')

The confusion matrix, ROC Curve and Precision recall curve is plotted below:







Part 5: Gradient Boosting Tree

Please try to implement the xgboost library to this dataset.

1. Please use cross-validation to select the best tuning parameters
2. Please plot the ROC curve and compute the area under the ROC curve.
3. Please plot the precision recall curve and compute the average precision

4. Please compute the F1-score and confusion matrix

Best Tuning Parameters

{'max_depth': 5, 'min_child_weight': 1}

{'gamma': 0.4}

Confusion Matrix

[[3266 68]

[608 58]]

Normalized confusion matrix

[[0.97960408 0.02039592]

[0.91291291 0.08708709]]

Area under the ROC curve : 0.693287

	precision	recall	f1-score	support
0	0.84	0.98	0.91	3334
1	0.46	0.09	0.15	666

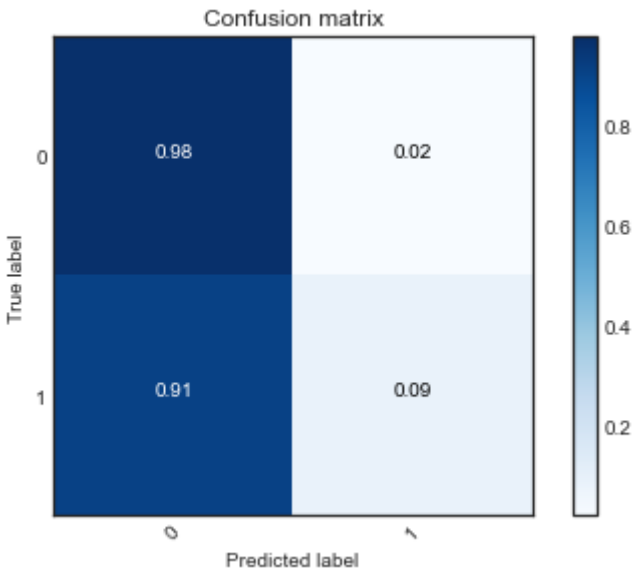
avg / total 0.78 0.83 0.78 4000

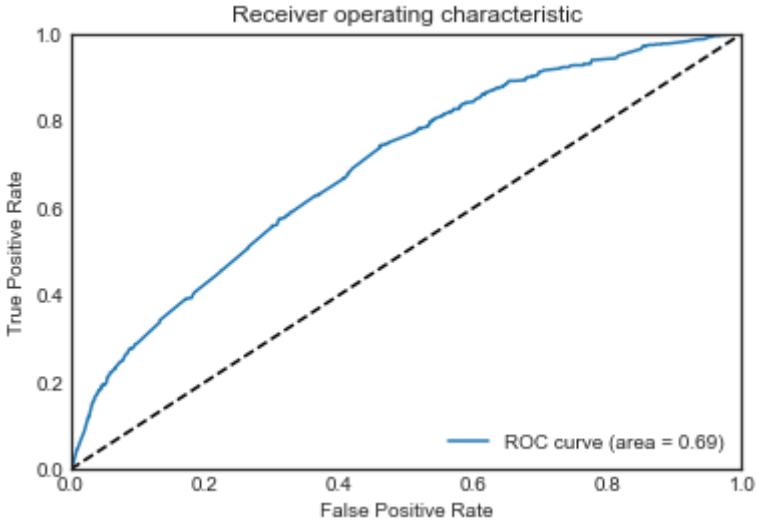
The Matthews Corelation coefficient : 0.142240458602

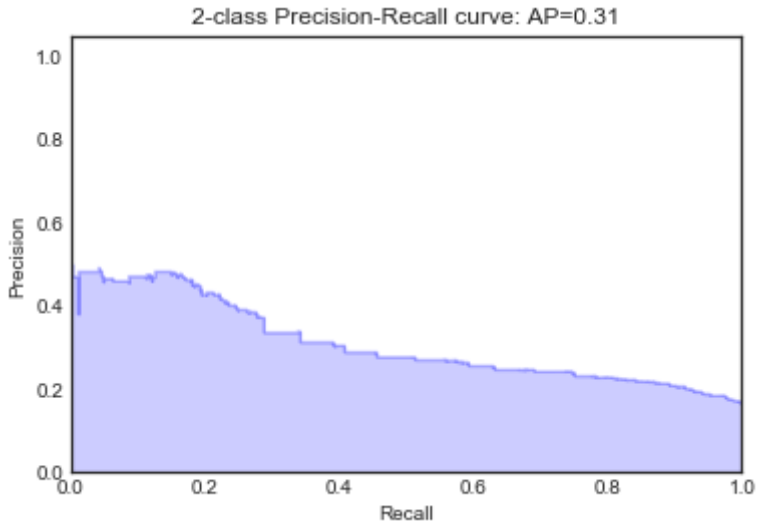
Average precision-recall score: 0.31

Text(0.5,1,'2-class Precision-Recall curve: AP=0.31')

The confusion matrix, ROC Curve and Precision recall curve is plotted below:




Receiver operating characteristic
True Positive Rate
False Positive Rate
ROC curve (area = 0.69)
0.0
0.2
0.4
0.6
0.8
1.0


2-class Precision-Recall curve: AP=0.31
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0