Machine Learning for Data Certification

by Humza Khan

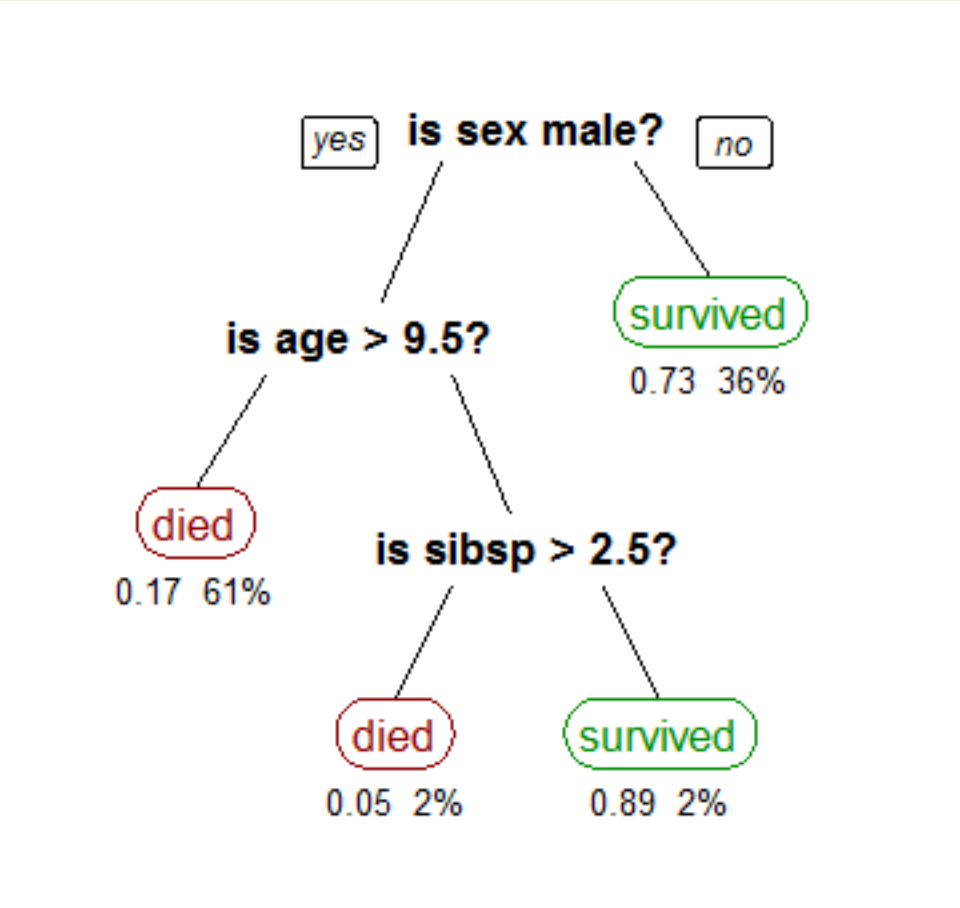
Mentors: Federico de Guio and Nural Akchurin

Recap

- Data at CMS needs to be certified
- A lot is removed with preliminary filters
- Remaining data is hand-checked by detector experts
- Minimize data experts need to check
- Use machine learning

Boosted Decision Trees (BDT)

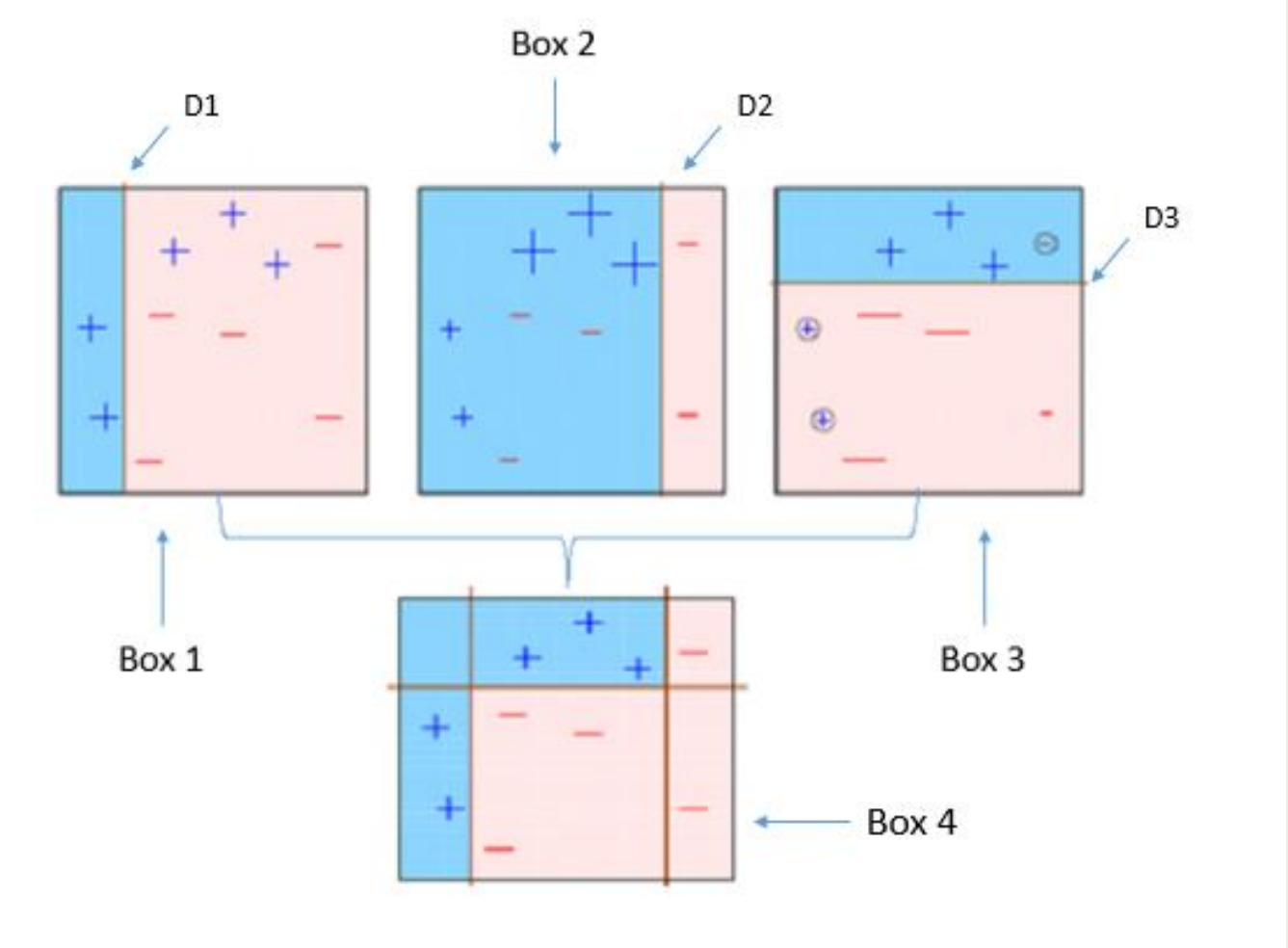
- Uses a binary tree to learn
- Each interior node has a feature on it
- Each leaf tells you the probability of that outcome



Boosted Decision Trees (continued)

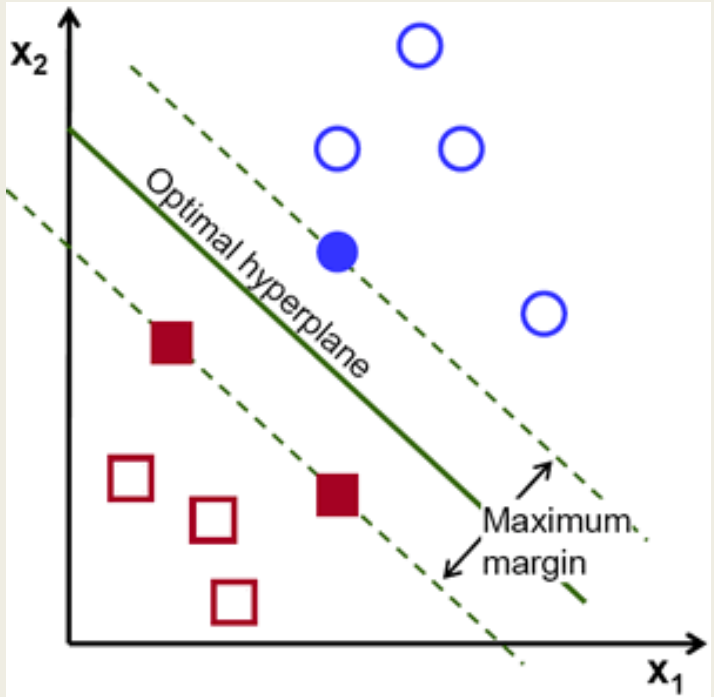
- Takes weak learner, weights samples equally
- Fits to data
- Re-weights samples based on error
- Iterates through until desired accuracy is reached
- Ensemble of weak learners turns into strong learner

Boosted Decision Trees (continued)



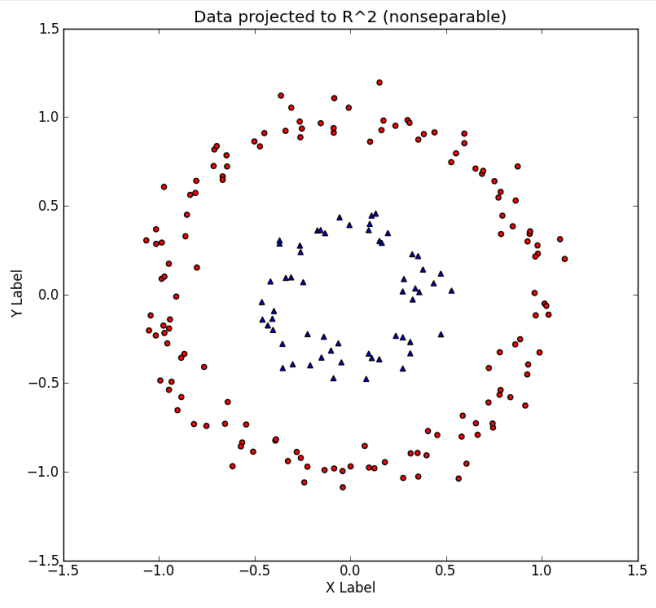
Support Vector Machines (SVM)

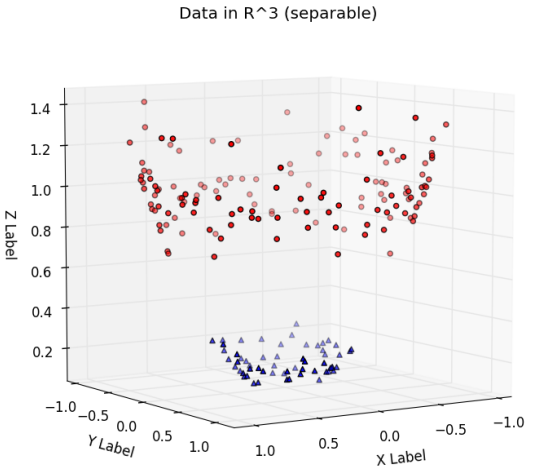
- Data with n features is in n-dimensional space
- Find $n - 1$ dimensional hyperplane to divide data
- Maximize distance between hyperplane and points
- Not all data is linearly separable



Support Vector Machines (continued)

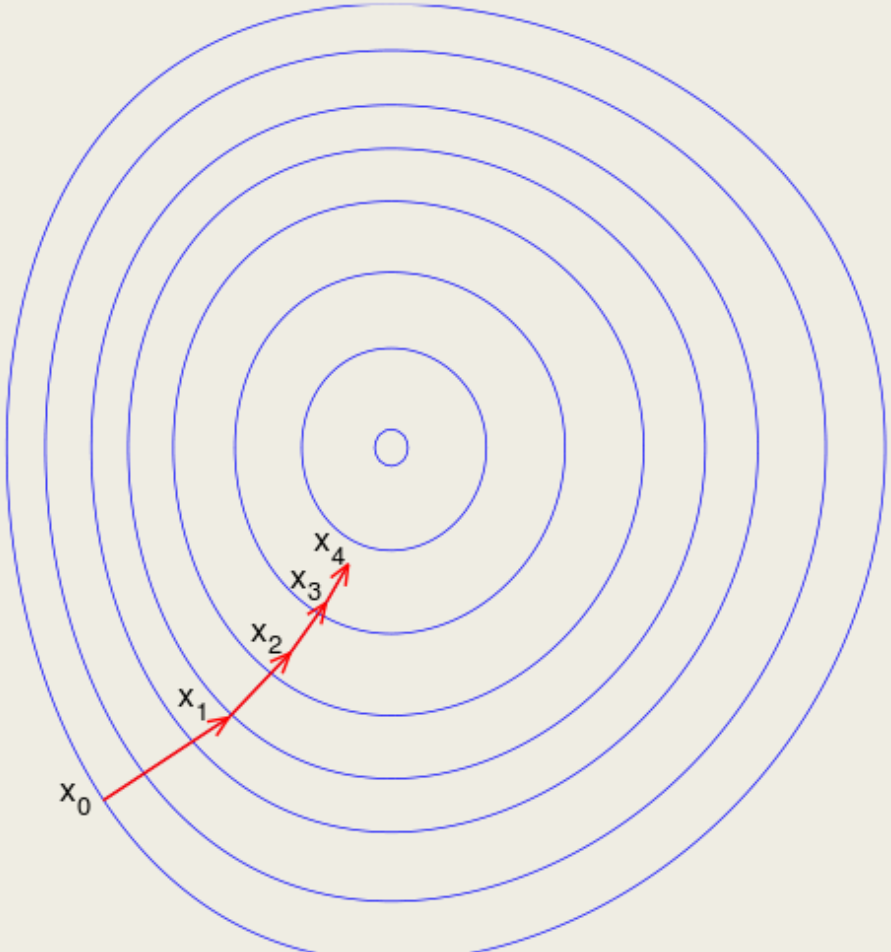
- Kernel function maps feature space to higher dimensional space
- $\varphi(x_1, x_2) = (x_1, x_2, x_1^2 + x_2^2)$
- Kernel function computes inner product in lower dimensional space





Stochastic Gradient Descent (SGD)

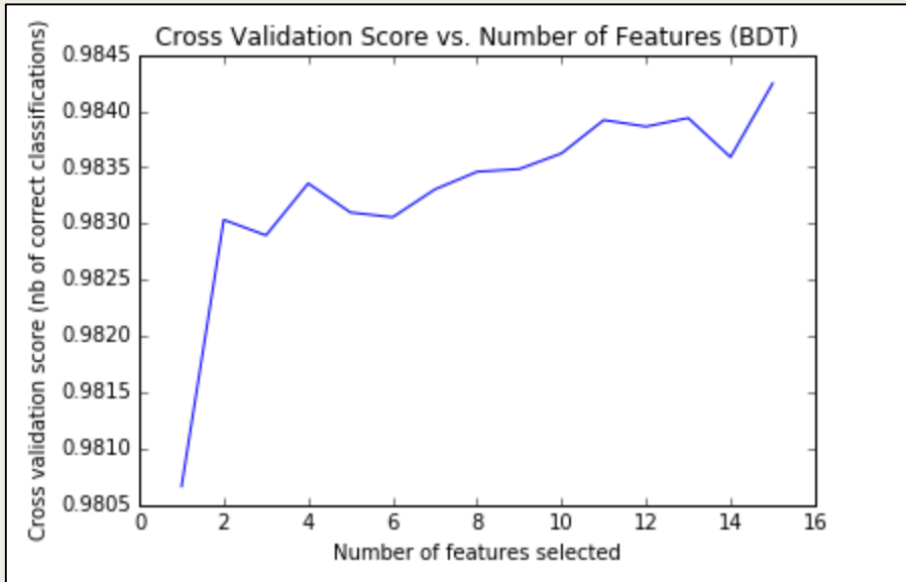
- Finds greatest derivative at point and moves that way
- Good at finding minima quickly
- Can get stuck at local minimum instead of global
- SGD only updates based on one sample instead of all



Feature Selection

- Not all features might be useful
- Good to eliminate non-discriminatory features
- Reduces overfitting and training time
- Variance threshold
- Recursive Feature Elimination

```
[ True  True False  True  True  True  True False False False False False
  True False False False False False False False False  True  True  True
  True  True  True  True  True]
```



BDT Results

Classification report for BDT, Untuned, No Weights AdaBoostClassifier(algorithm='SAMME',
base_estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
max_features=None, max_leaf_nodes=None,
min_impurity_split=1e-07, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0,
presort=False, random_state=None, splitter='best'),
learning_rate=1, n_estimators=200, random_state=None):

	precision	recall	f1-score	support
0.0	0.94	0.73	0.82	6004
1.0	0.99	1.00	0.99	110215
avg / total	0.98	0.98	0.98	116219

Classification report for BDT, Untuned, Weights AdaBoostClassifier(algorithm='SAMME',
base_estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=3,
max_features=None, max_leaf_nodes=None,
min_impurity_split=1e-07, min_samples_leaf=1,
min_samples_split=2, min_weight_fraction_leaf=0.0,
presort=False, random_state=None, splitter='best'),
learning_rate=1, n_estimators=200, random_state=None):

	precision	recall	f1-score	support
0.0	0.94	0.67	0.78	6004
1.0	0.98	1.00	0.99	110215
avg / total	0.98	0.98	0.98	116219

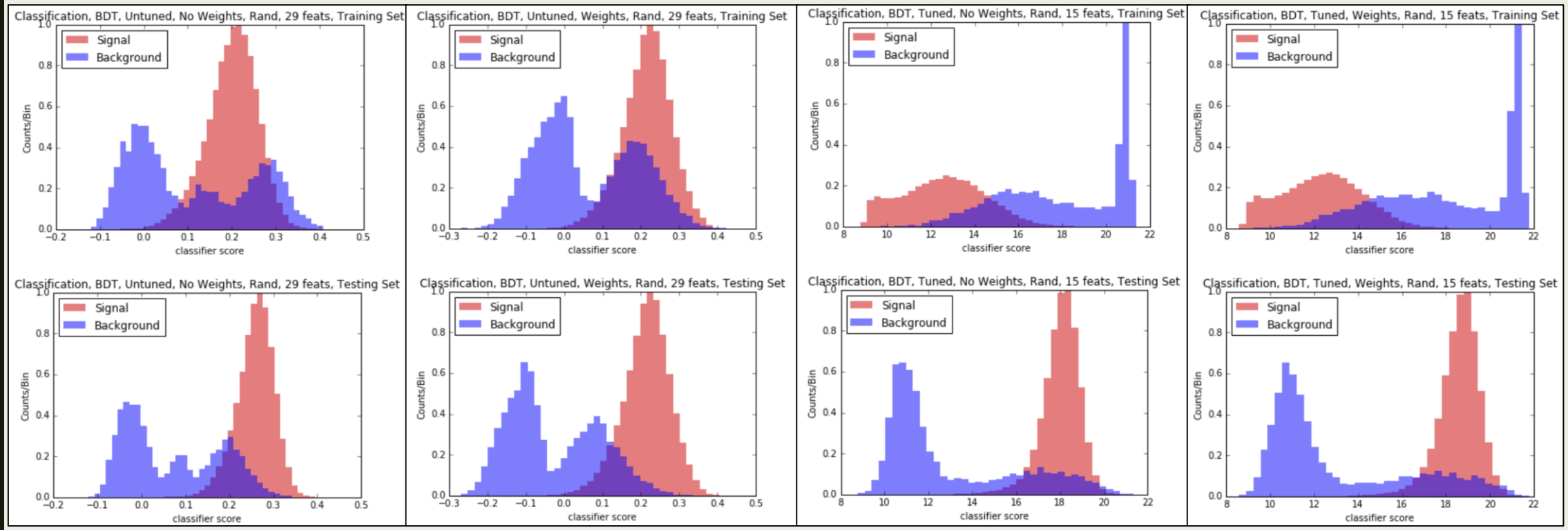
Classification report for BDT, Tuned, No Weights AdaBoostClassifier(algorithm='SAMME.R',
base_estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=50,
max_features=None, max_leaf_nodes=None,
min_impurity_split=1e-07, min_samples_leaf=10,
min_samples_split=50, min_weight_fraction_leaf=0.0,
presort=False, random_state=None, splitter='best'),
learning_rate=1, n_estimators=200, random_state=None):

	precision	recall	f1-score	support
0.0	0.98	0.76	0.85	6004
1.0	0.99	1.00	0.99	110215
avg / total	0.99	0.99	0.99	116219

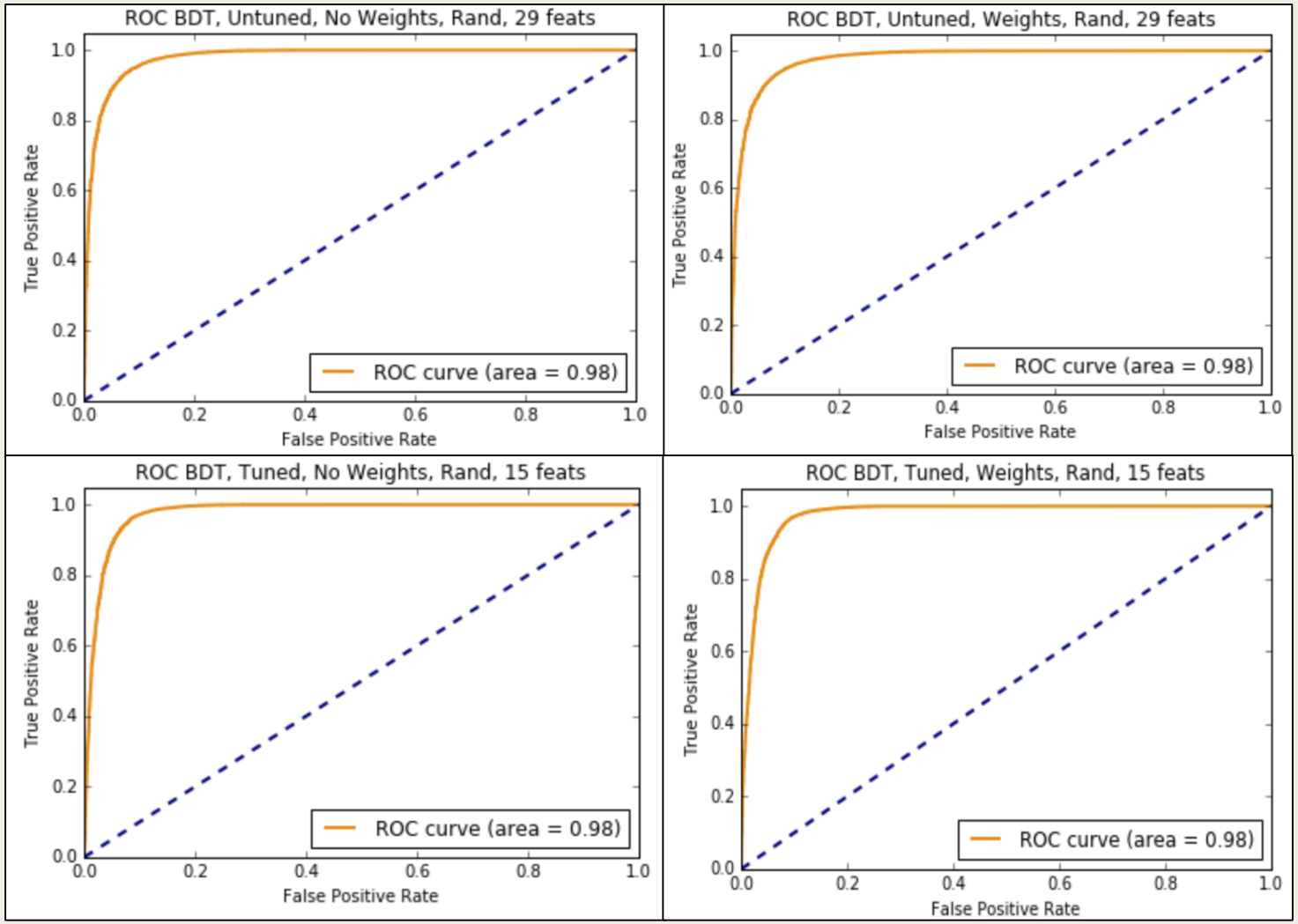
Classification report for BDT, Tuned, Weights AdaBoostClassifier(algorithm='SAMME.R',
base_estimator=DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=50,
max_features=None, max_leaf_nodes=None,
min_impurity_split=1e-07, min_samples_leaf=10,
min_samples_split=50, min_weight_fraction_leaf=0.0,
presort=False, random_state=None, splitter='best'),
learning_rate=1, n_estimators=200, random_state=None):

	precision	recall	f1-score	support
0.0	0.98	0.76	0.85	6004
1.0	0.99	1.00	0.99	110215
avg / total	0.99	0.99	0.99	116219

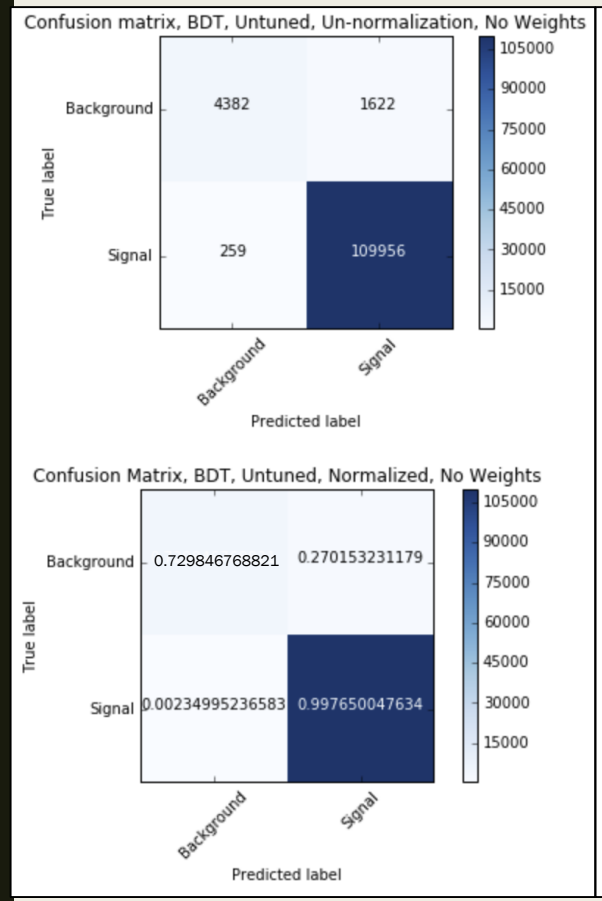
BDT Results (continued)

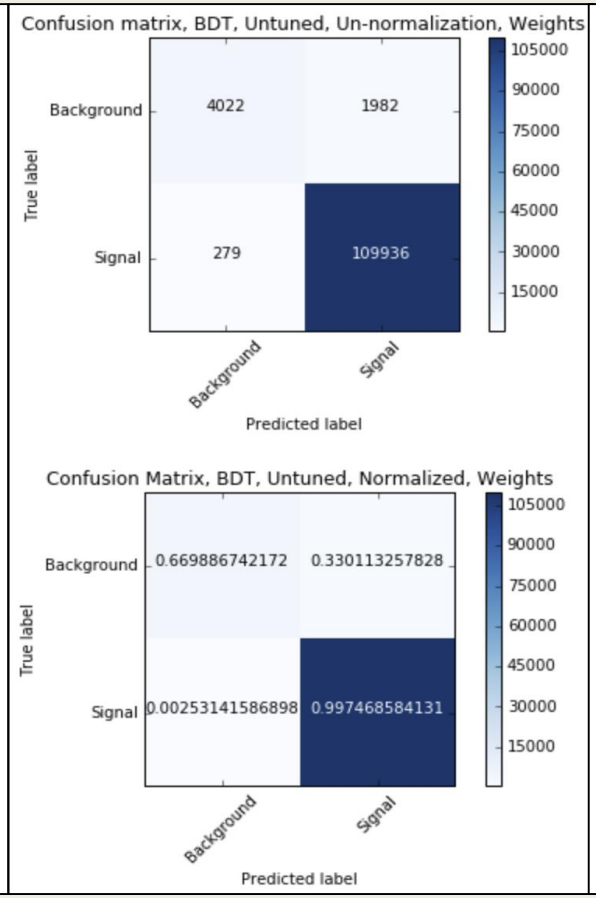


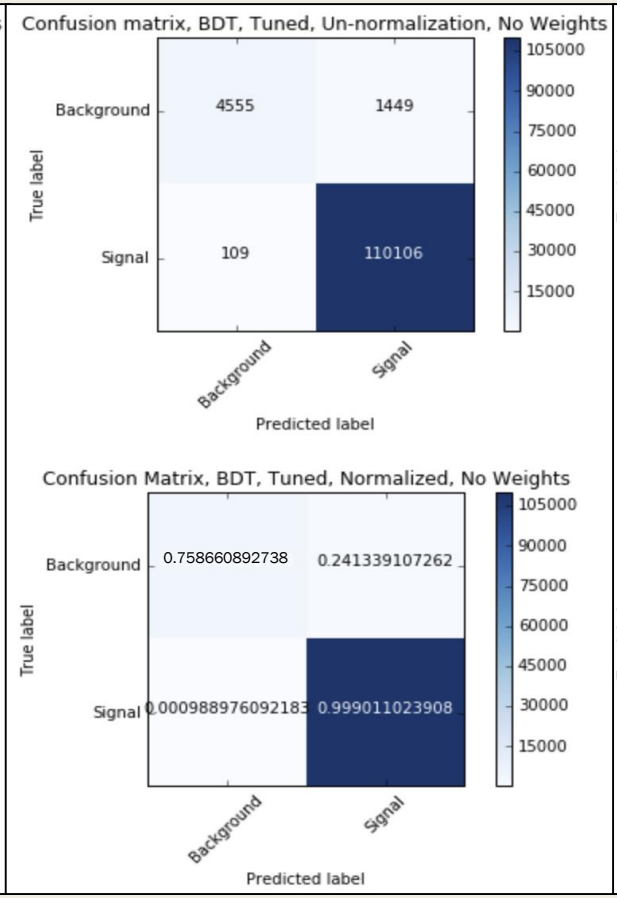
BDT (continued)

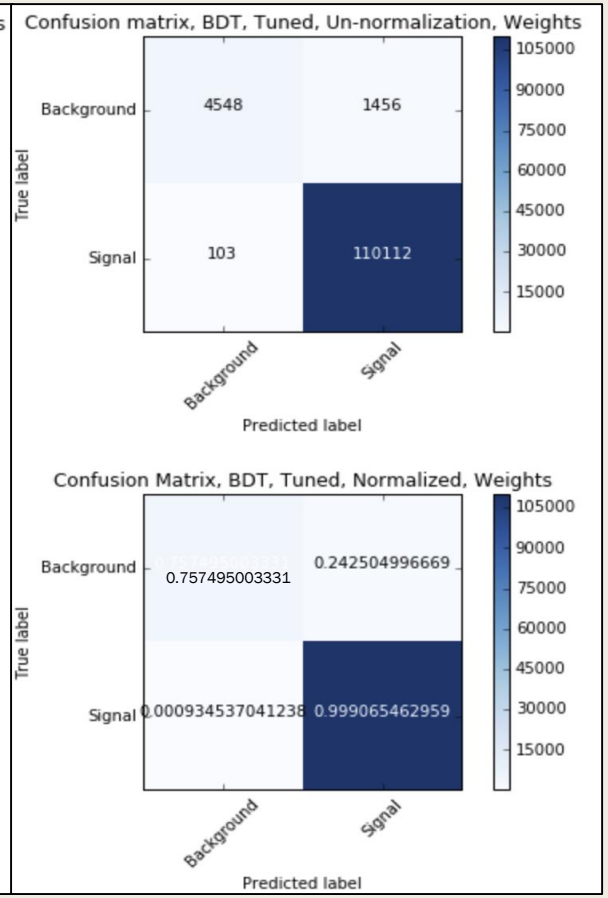


BDT Results (continued)









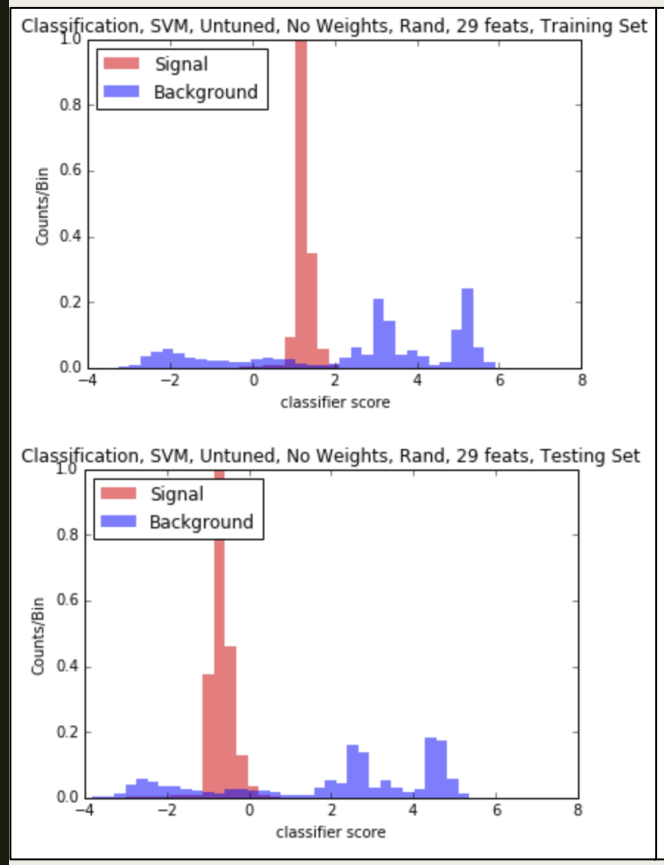
SVM Results

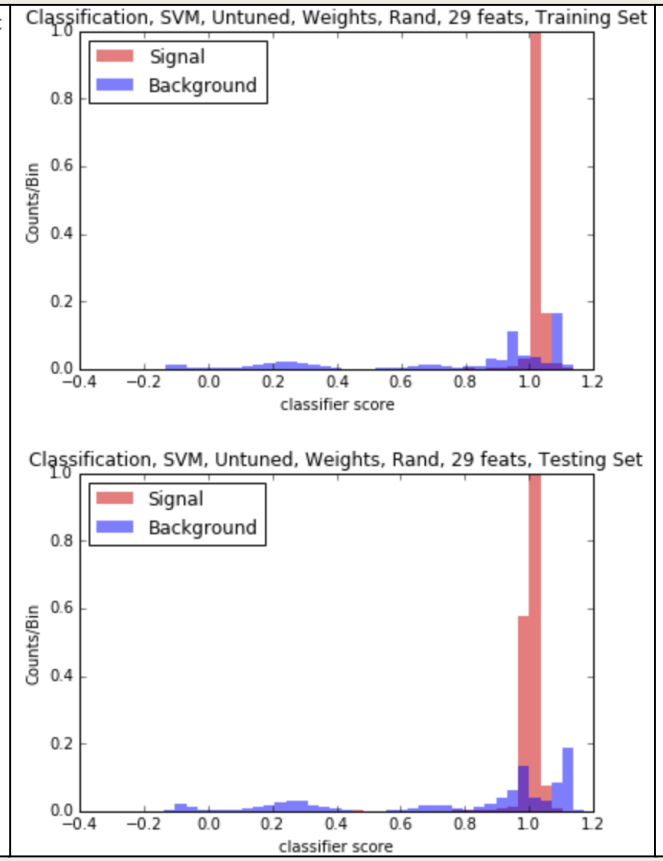
```
Classification report for SVM, Untuned, No Weights SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape=None, degree=3, gamma=0.005, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False):
             precision    recall  f1-score   support

        0.0       0.91      0.66      0.77      6004
        1.0       0.98      1.00      0.99    110215

avg / total       0.98      0.98      0.98    116219


CPU times: user 2min 29s, sys: 977 ms, total: 2min 30s
Wall time: 2min 31s
```

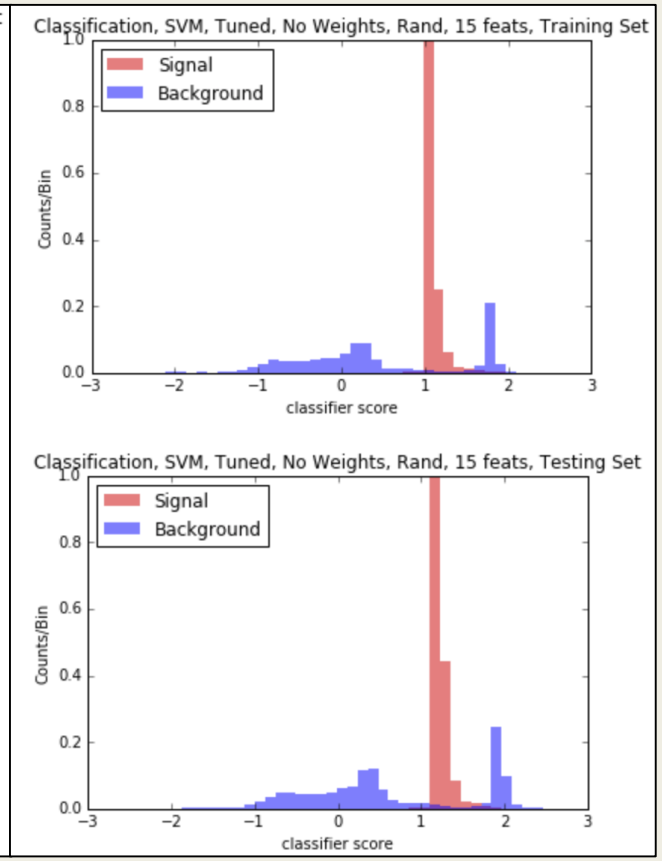
```
Classification report for SVM, Untuned, Weights SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
  decision_function_shape=None, degree=3, gamma=0.005, kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False):
             precision    recall  f1-score   support

        0.0       1.00      0.24      0.38      6004
        1.0       0.96      1.00      0.98    110215

avg / total       0.96      0.96      0.95    116219


CPU times: user 6min 10s, sys: 1.86 s, total: 6min 12s
Wall time: 6min 14s
```

```
Classification report for SVM, Tuned, No Weights SVC(C=1, cache_size=200,
  decision_function_shape=None, degree=3, gamma='auto', kernel='rbf',
  max_iter=-1, probability=False, random_state=None, shrinking=True,
  tol=0.001, verbose=False):
             precision    recall  f1-score   support

        0.0       0.98      0.69      0.81      6004
        1.0       0.98      1.00      0.99    110215

avg / total       0.98      0.98      0.98    116219
```

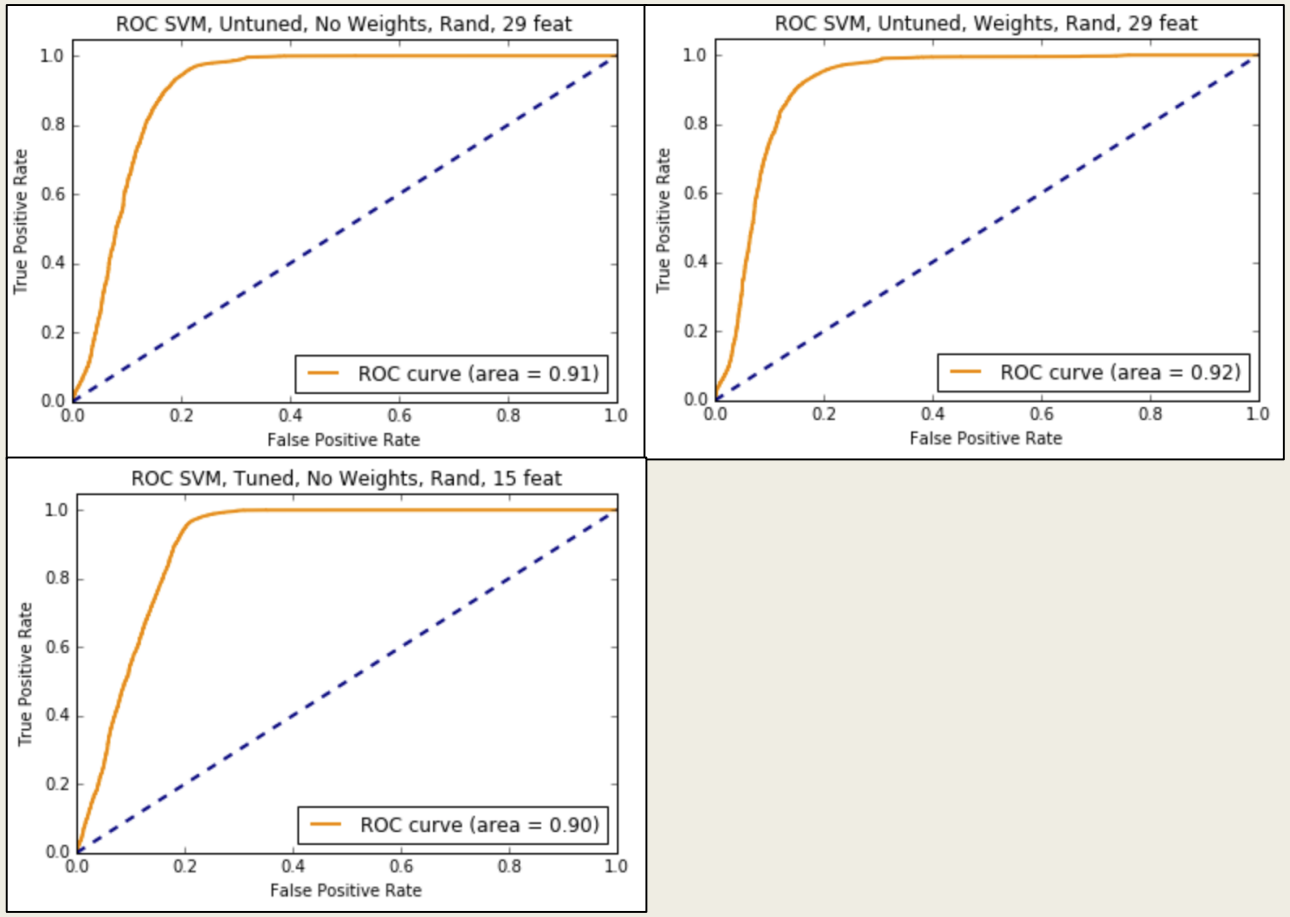
SVM Results (continued)



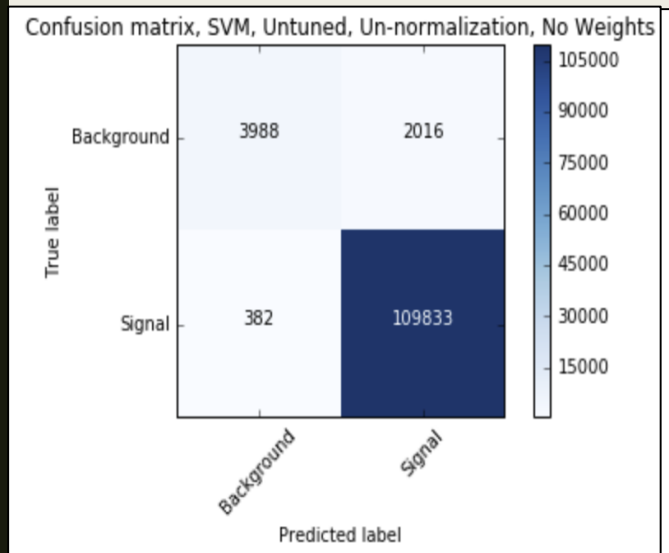


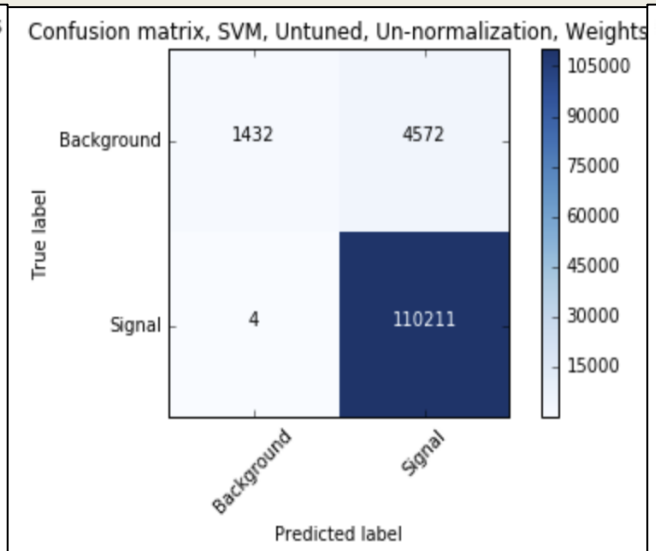


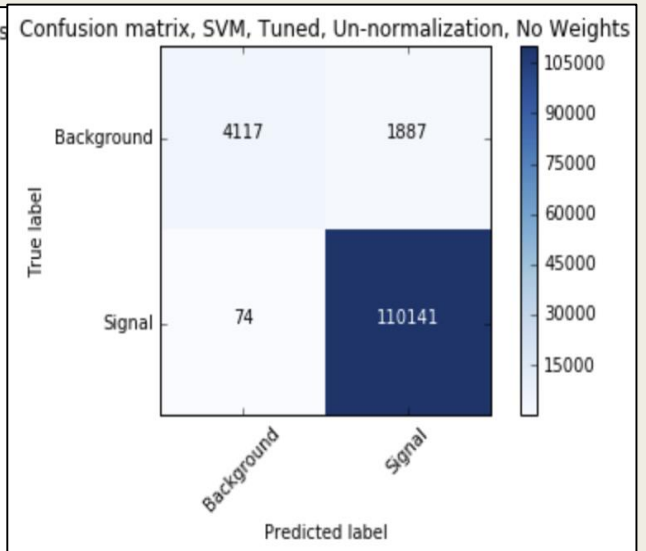
SVM Results (continued)

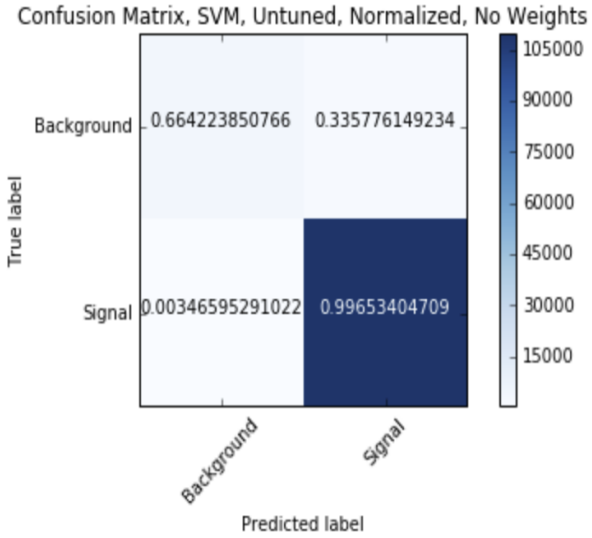


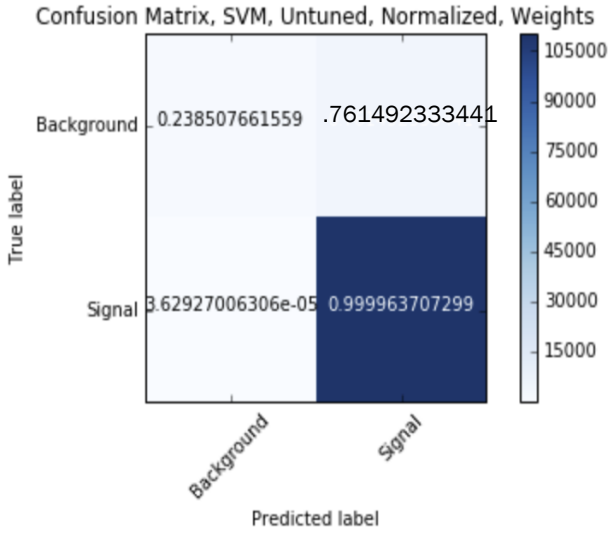
SVM Results (continued)

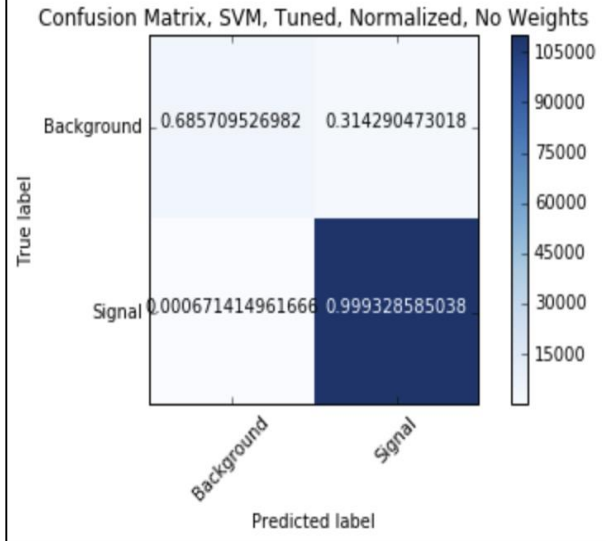












SGD Results

Classification report for SGD, Untuned, No Weights, SGDClassifier(alpha=0.0001, silon=0.1,
eta0=0.0, fit_intercept=True, l1_ratio=0.15,
learning_rate='optimal', loss='hinge', n_iter=5, n_jobs=1,
penalty='l2', power_t=0.5, random_state=None, shuffle=True,
verbose=0, warm_start=False):

	precision	recall	f1-score	support
-1.0	0.84	0.32	0.46	6004
1.0	0.96	1.00	0.98	110215
avg / total	0.96	0.96	0.95	116219

Classification report for SGD, Untuned, Weights, SGDClassifier(alpha=0.0001, on=0.1,
eta0=0.0, fit_intercept=True, l1_ratio=0.15,
learning_rate='optimal', loss='hinge', n_iter=5, n_jobs=1,
penalty='l2', power_t=0.5, random_state=None, shuffle=True,
verbose=0, warm_start=False):

	precision	recall	f1-score	support
-1.0	0.83	0.32	0.46	6004
1.0	0.96	1.00	0.98	110215
avg / total	0.96	0.96	0.95	116219

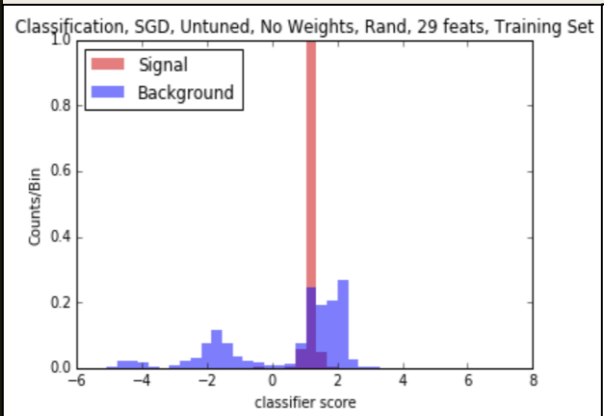
Classification report for SGD, Tuned, No Weights, SGDClassifier(alpha=0.0001,
epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15,
learning_rate='optimal', loss='huber', n_iter=9, n_jobs=1,
penalty='l2', power_t=0.5, random_state=None, shuffle=True,
verbose=0, warm_start=False):

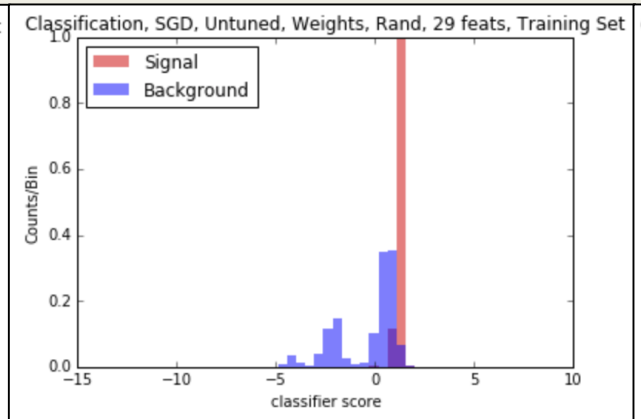
	precision	recall	f1-score	support
-1.0	0.41	0.70	0.52	6004
1.0	0.98	0.94	0.96	110215
avg / total	0.95	0.93	0.94	116219

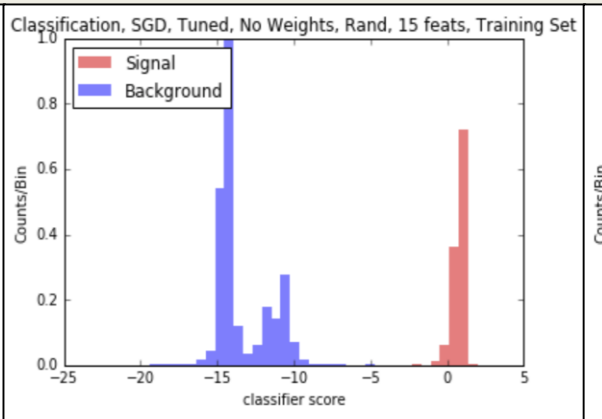
Classification report for SGD, Tuned, Weights, SGDClassifier(alpha=0.0001,
epsilon=0.1, eta0=0.0, fit_intercept=True, l1_ratio=0.15,
learning_rate='optimal', loss='huber', n_iter=9, n_jobs=1,
penalty='l2', power_t=0.5, random_state=None, shuffle=True,
verbose=0, warm_start=False):

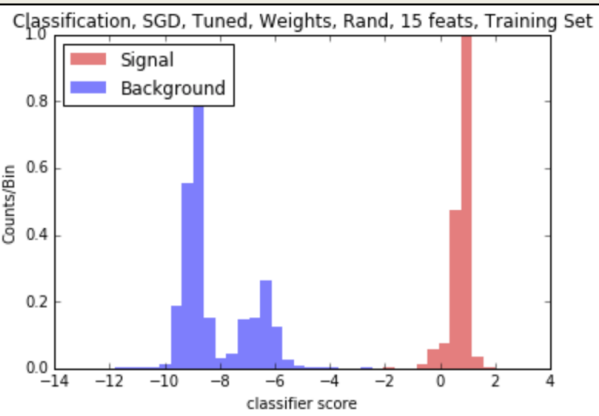
	precision	recall	f1-score	support
-1.0	0.42	0.70	0.53	6004
1.0	0.98	0.95	0.97	110215
avg / total	0.95	0.94	0.94	116219

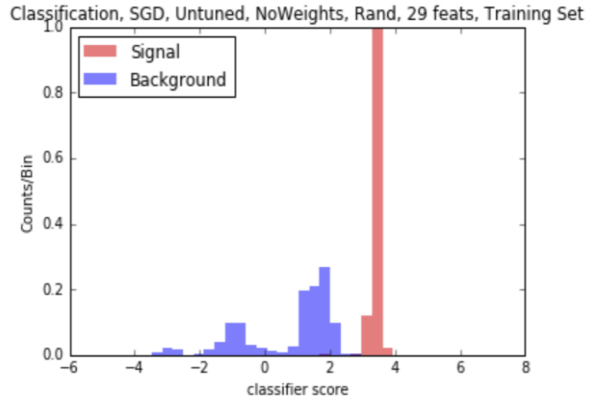
SGD Results (continued)

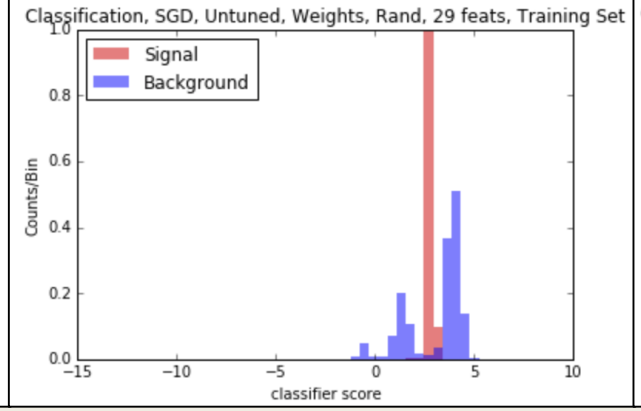


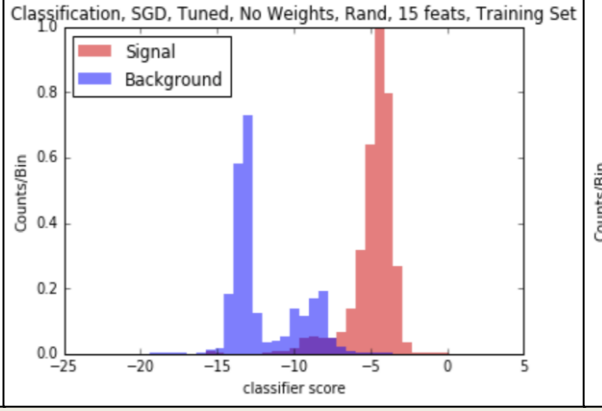


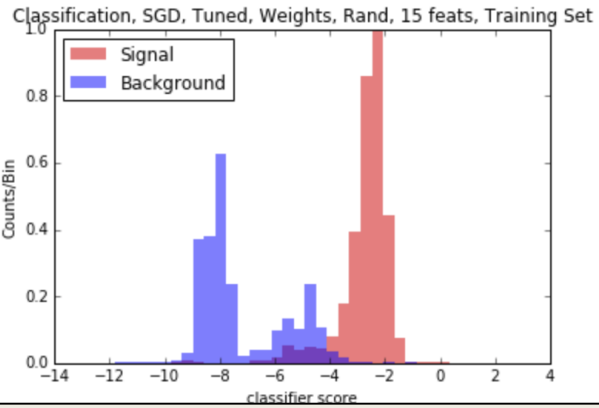




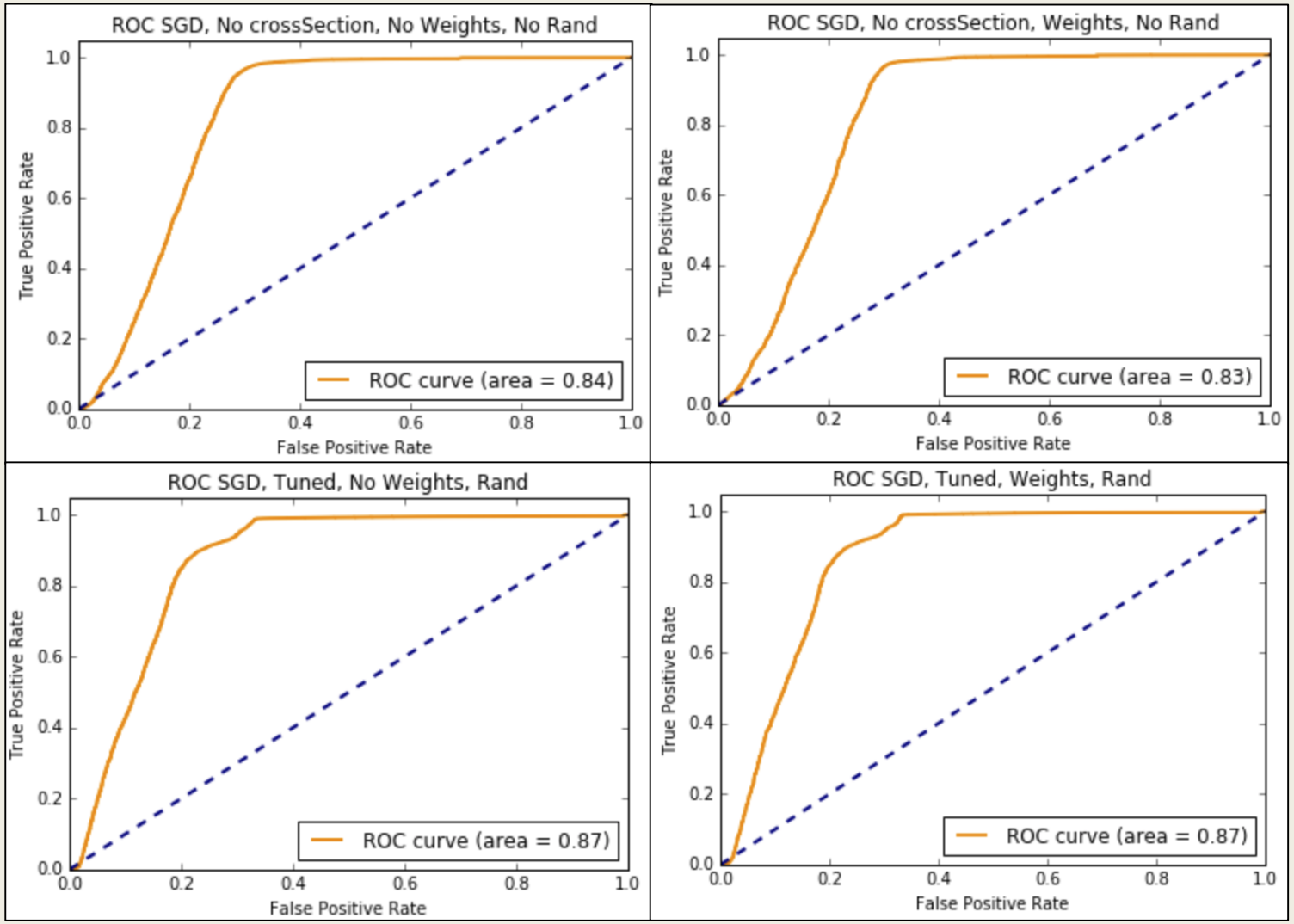




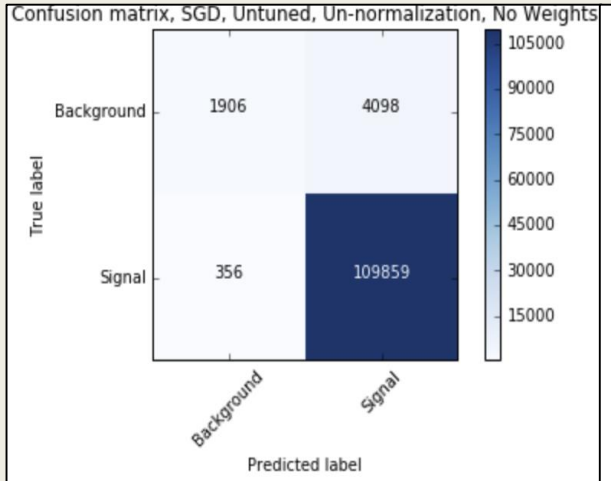


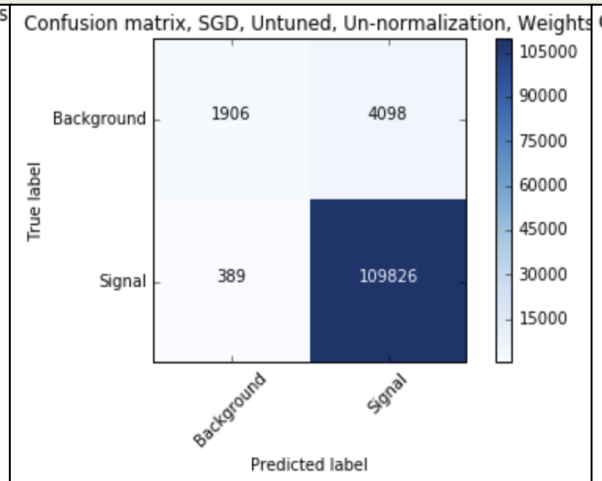


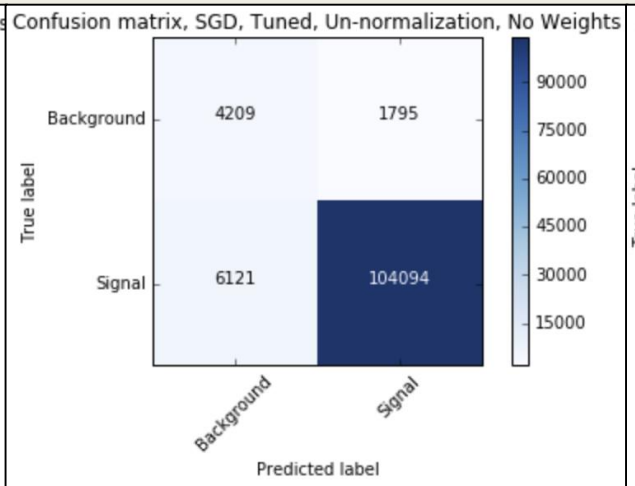
SGD Results (continued)

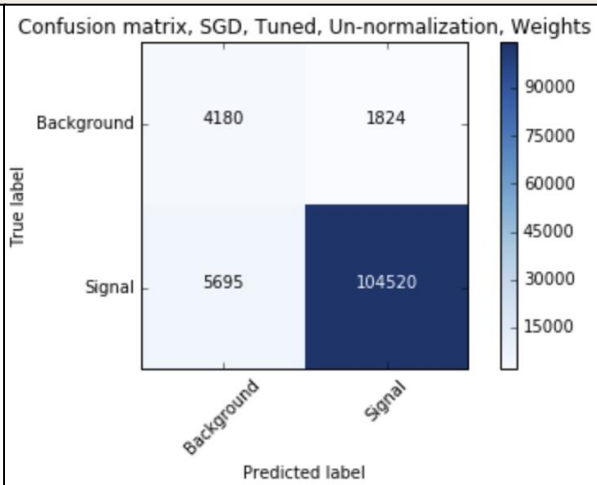


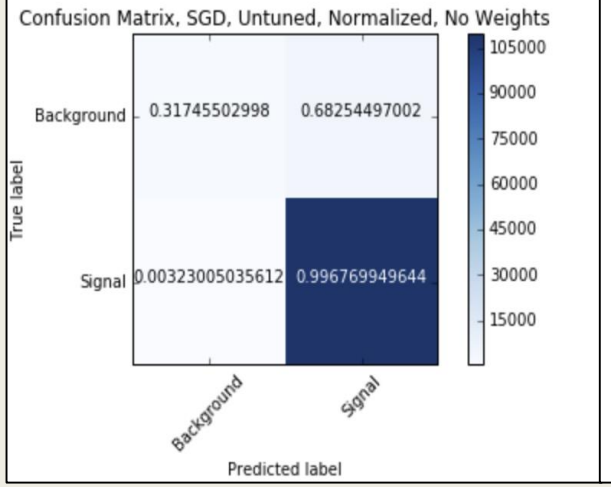
SGD Results (continued)

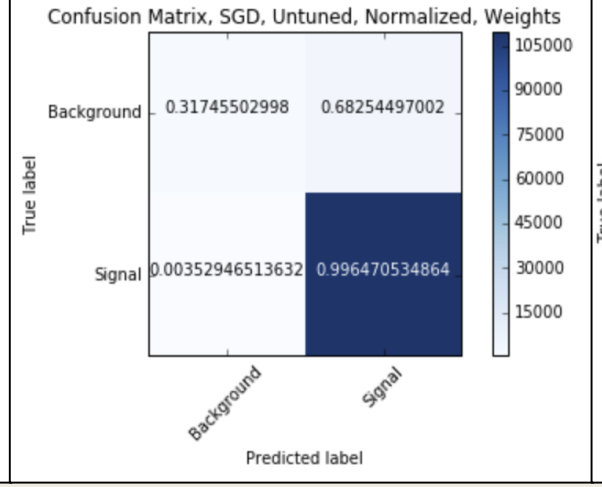


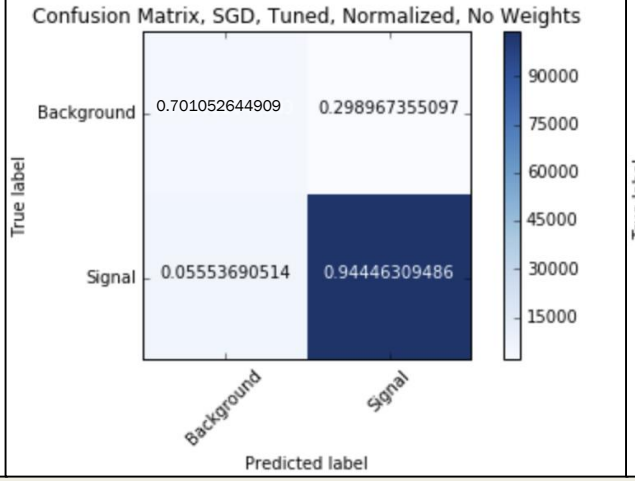


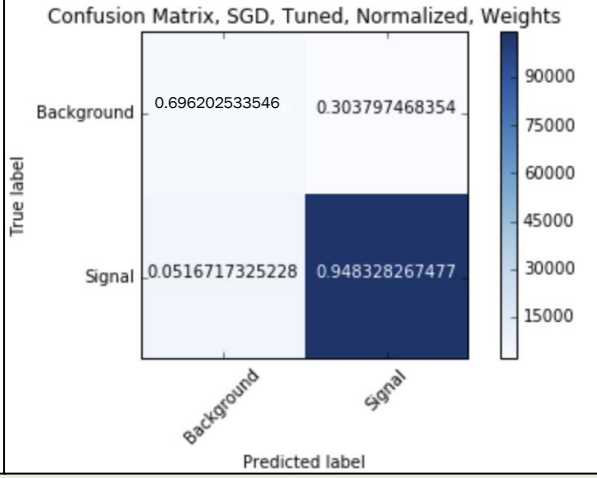












Matthews Correlation Coefficients

	SVM	BDT	SGD
Untuned, no weights	0.769	0.811	0.503
Untuned, weights	0.478	0.769	0.499
Tuned, no weights	0.813	0.854	0.502
Tuned, weights	N/A	0.856	0.702

Summary

- BDT seems to outperform the other two algorithms
- Need to optimize some more
- Implement realistic workflow where data is sent in and value is spat out

Sources

- http://docs.opencv.org/2.4/_images/optimal-hyperplane.png
- https://upload.wikimedia.org/wikipedia/commons/f/f3/CART_tree_titanic_survivors.png
- https://www.analyticsvidhya.com/wp-content/uploads/2015/11/bigd.png
- http://www.eric-kim.net/eric-kim-net/posts/1/kernel_trick.html