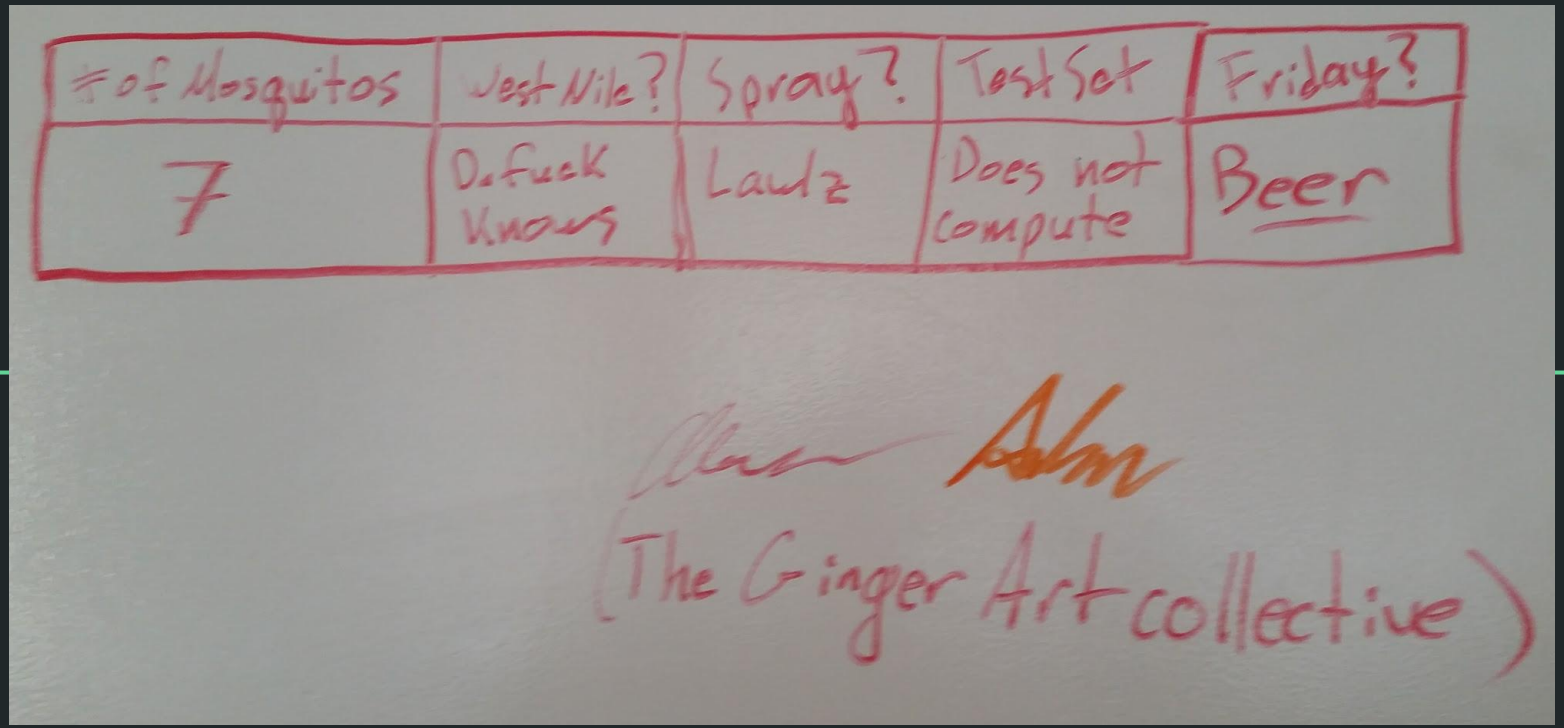
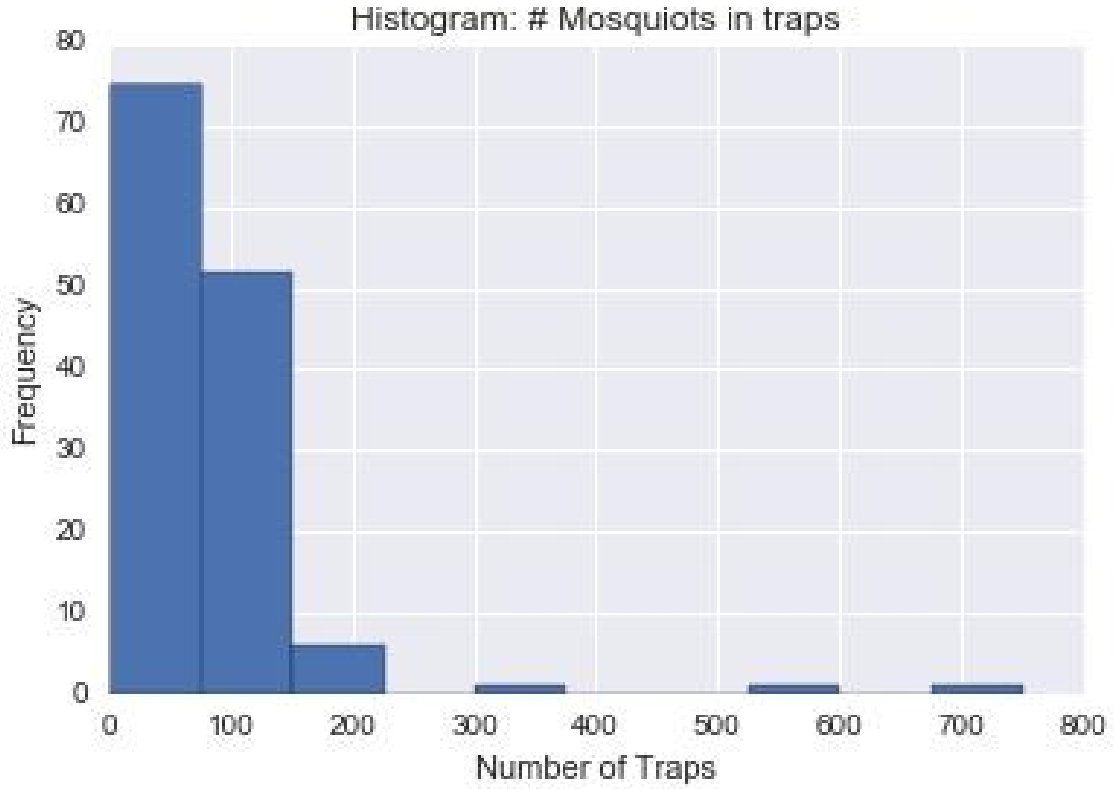
Chicago: West Nile Virus



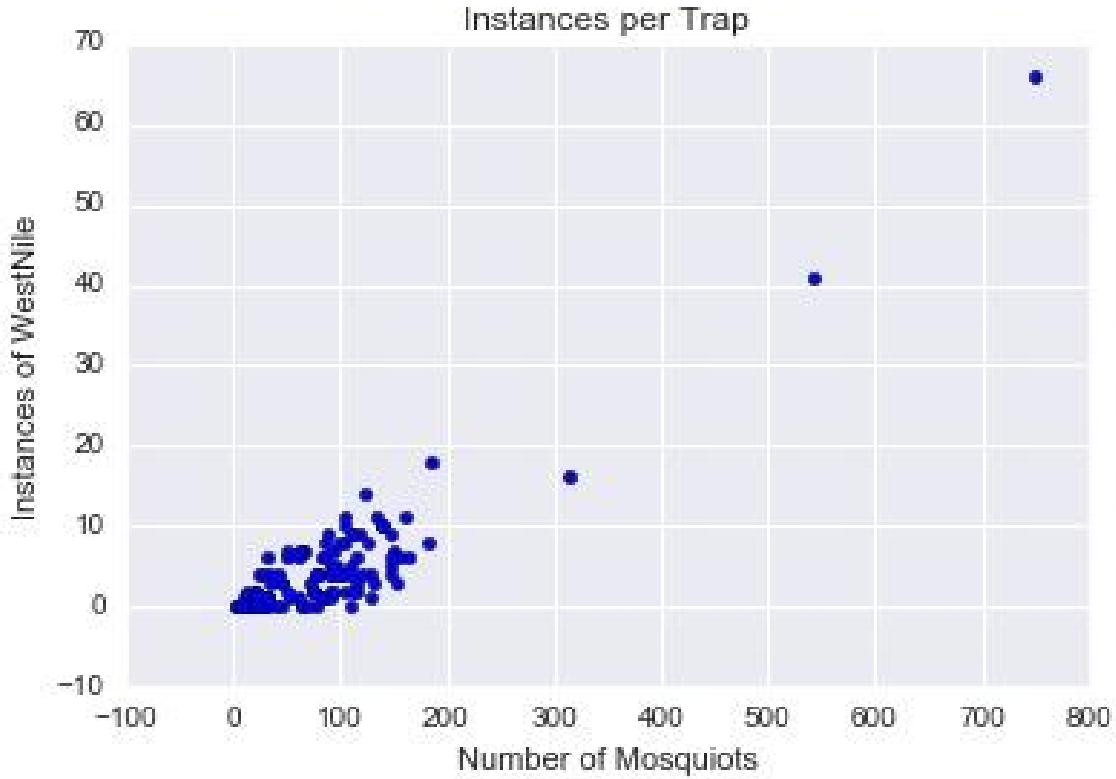
Purpose:

We are attempting to predict a binary value(yes/no) for a test of the West Nile Virus in various traps set around the city of Chicago in odd numbered years between 2007-2013. We will be using a variety of predictive variables including weather, geographical location, and yearly/monthly/weekly trends to the information.

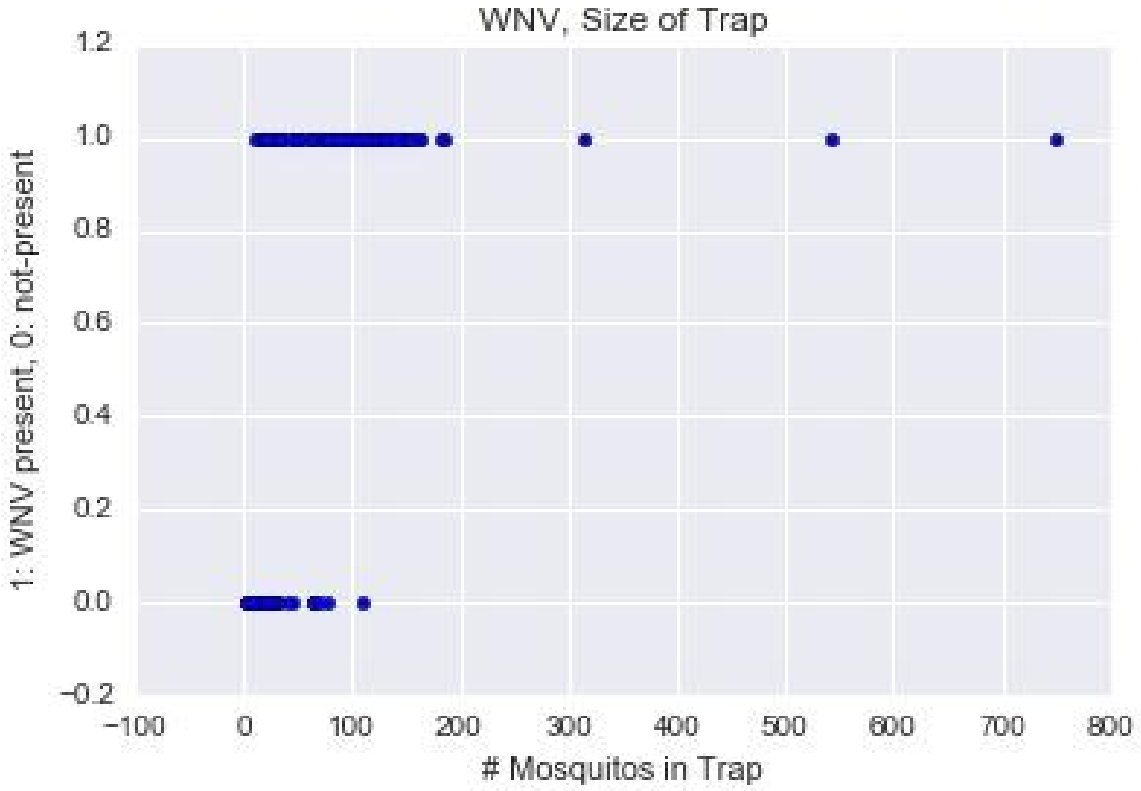
Exploratory Data Analysis



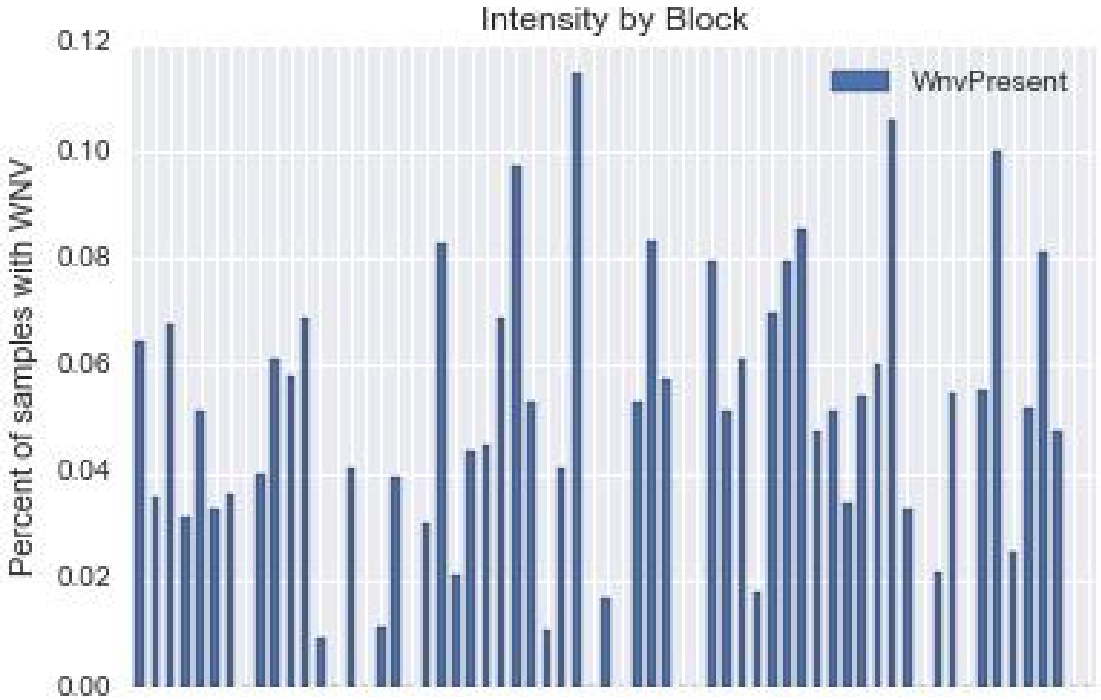
EDA: cont


Instances per Trap
Instances of WestNile
Number of Mosquiots
70
60
50
40
30
20
10
0
−10
−100
0
100
200
300
400
500
600
700
800

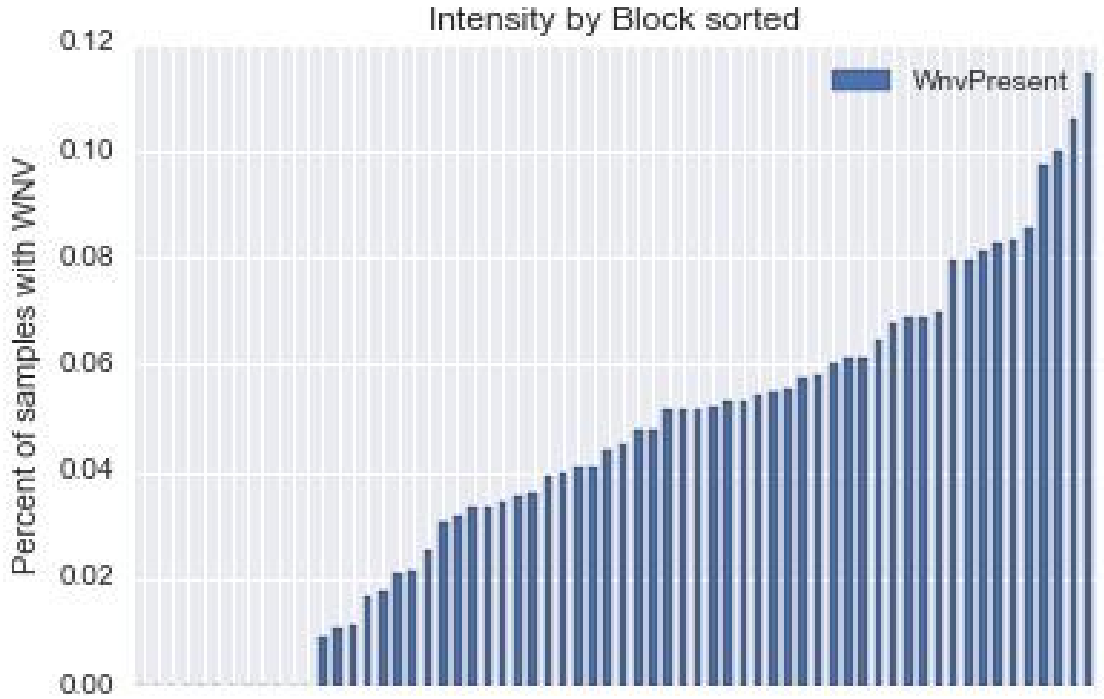
EDA: cont


WNV, Size of Trap
1: WNV present, 0: not-present
Mosquitos in Trap
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−100
0
100
200
300
400
500
600
700
800

Cleaning:


Intensity by Block
WnvPresent
Percent of samples with WNV
0.12
0.10
0.08
0.06
0.04
0.02
0.00

Cleaning: cont



Block

Street

Trap

Week (quarter month)

(Did not utilize test data)

Cleaning: cont

Dummies for Species, finally merged in weather. Did not utilize Spray.

block	street	trap	lat	lon	num_mos	week	s_err	s_pip	s_p/r	...	Sunset	Depth	SnowFall	PrecipTotal
0.034483	0.045802	0.045802	41.682587	-87.707973	2	0.112266	0	1.0	0.0	...	1821.0	0.0	0.0	0.00
0.066414	0.073840	0.073840	41.673408	-87.599862	50	0.044426	0	0.0	0.0	...	1911.0	0.0	0.0	0.00
0.063181	0.100000	0.100000	41.688324	-87.676709	17	0.044426	0	1.0	0.0	...	1911.0	0.0	0.0	0.00
0.077670	0.103261	0.166667	41.999129	-87.795585	42	0.107889	0	0.0	1.0	...	1854.0	0.0	0.0	0.23
0.029412	0.008403	0.008403	41.678618	-87.559308	5	0.044426	0	0.0	1.0	...	1904.0	0.0	0.0	0.33
0.029412	0.098039	0.098039	41.904194	-87.756155	15	0.107889	0	0.0	1.0	...	1854.0	0.0	0.0	0.00
0.066836	0.051282	0.052632	41.726465	-87.585413	50	0.107889	0	1.0	0.0	...	1903.0	0.0	0.0	0.06
0.029091	0.000000	0.000000	41.947227	-87.671457	1	0.107889	0	1.0	0.0	...	1854.0	0.0	0.0	0.00

Feature Selection

Brute Force Algorithm-

Built A decision tree for each individual feature AND every set of two features. Scored each model. Sorted Features by performance.

Features in order:

['Depth', 'SnowFall', 'PrecipTotal', 's_sal', 's_err', ‘s_tar', 'lat', 'AvgSpeed', 'Sunrise', 'ResultSpeed', 'trap', 'Sunset', 'lon', 'street', 'num_mos', 'week', 'StnPressure', 'DewPoint', 'ResultDir', 'block', 'SeaLevel', 'WetBulb', 'Depart', 'Tmin', 'Tmax', 'Tavg', 'Cool', 's_pip', 'Heat', 's_res', 's_ter', 's_p/r']

Model Selection

Brute Force Algorithm II

Fit and scored every model utilizing 10 best features.

Add next best feature to list. Re-fit and re-scored every model.

Repeated until tested with all features.

Models: LDA, Decision Tree, Random Forest, Extra Trees, Bagging, AdaBoost, Gradient Boosting.

Model

Linear Discriminant Analysis with 23 features

'Depth', 'SnowFall', 'PrecipTotal', 's_sal', 's_err', ‘s_tar', 'lat', 'AvgSpeed', 'Sunrise', 'ResultSpeed', 'trap', 'Sunset', 'lon', 'street', 'num_mos', 'week', 'StnPressure', 'DewPoint', 'ResultDir', 'block', 'SeaLevel', 'WetBulb', 'Depart'

Model Details

Accuracy: .949571

Base Confusion Matrix

(991) (2)

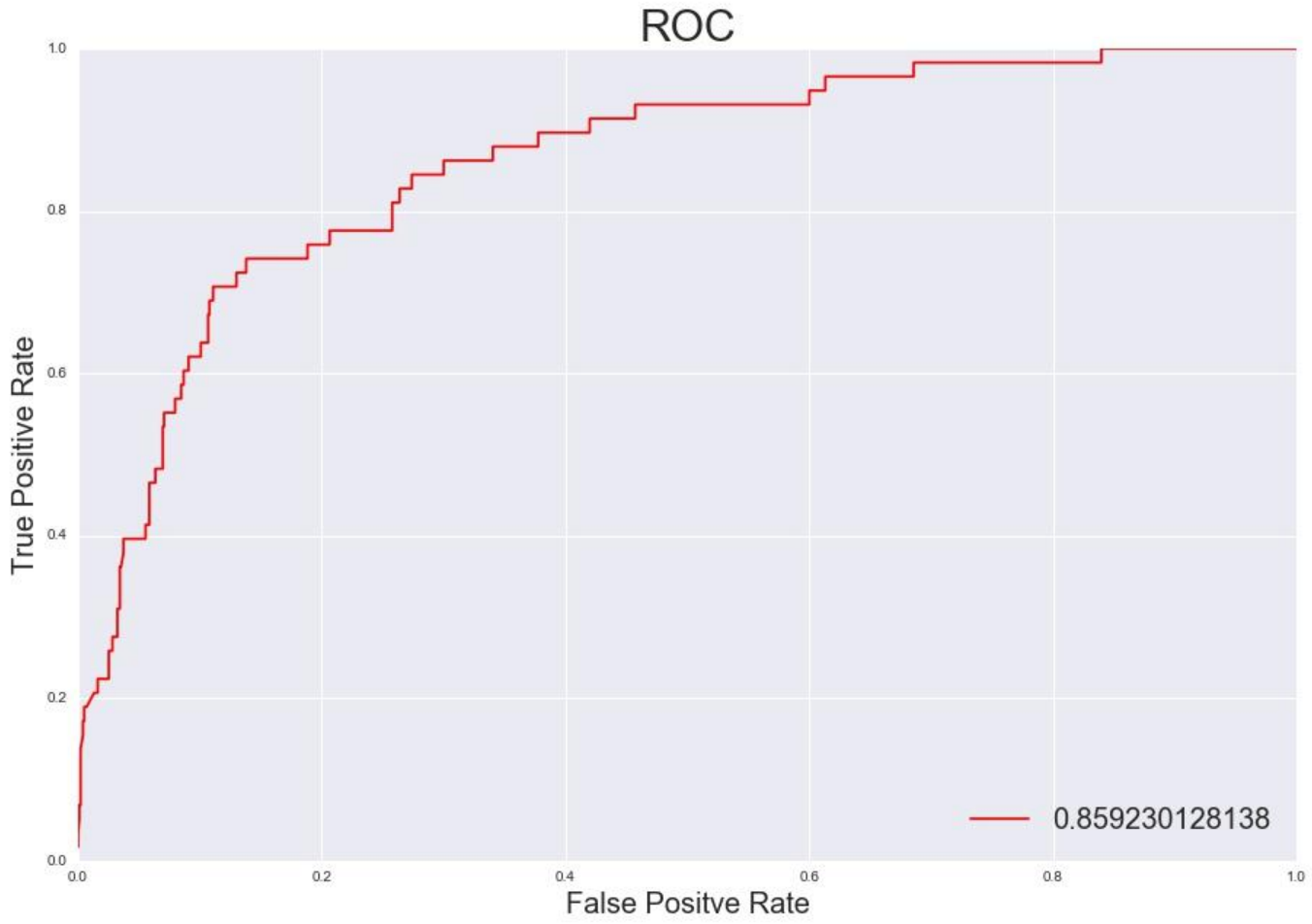
(51) (71)

Adjusted Confusion Matrix

(786) (207)

(13) (45)

*(Adjusted p -> .0553)


ROC
True Positive Rate
False Positve Rate
0.0
0.2
0.4
0.6
0.8
1.0
0.859230128138