Spambase Classification Models

Assignment 3

Computation EDA and Two-Eyed Algorithms in Binary Classification

Kevin Johnson

Northwestern

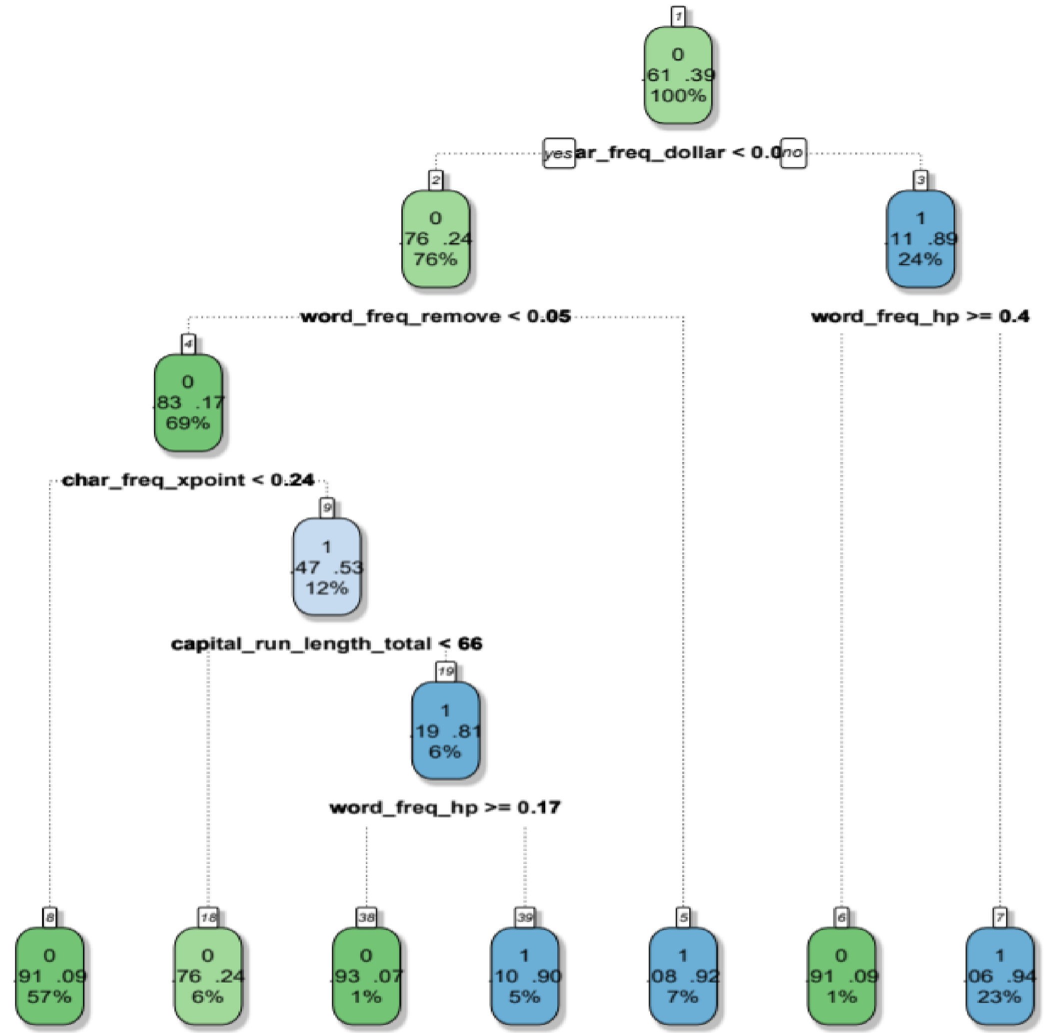
Data Description

- Source: UCI Machine Learning Repository
- Data Structure: 4601 Observations, 58 Variables
- 48 continuous [0,100] variables of type word_freq_WORD
- 6 continuous real [0,100] attributes of type char_freq_CHAR
- 1 continuous real [1,…] variable of type capital_run_length_average
- 1 continuous integer [1,…] variable of type capital_run_length_longest
- 1 continuous integer [1,…] attribute of capital_run_length+ total
- 1 nominal {0,1} class attribute of t

Split into Train and Test Sets

- Train and Test set to be split 50/50
- Train Set Dimensions: (2318,58)
- Test Set Dimensions: (2283, 58)

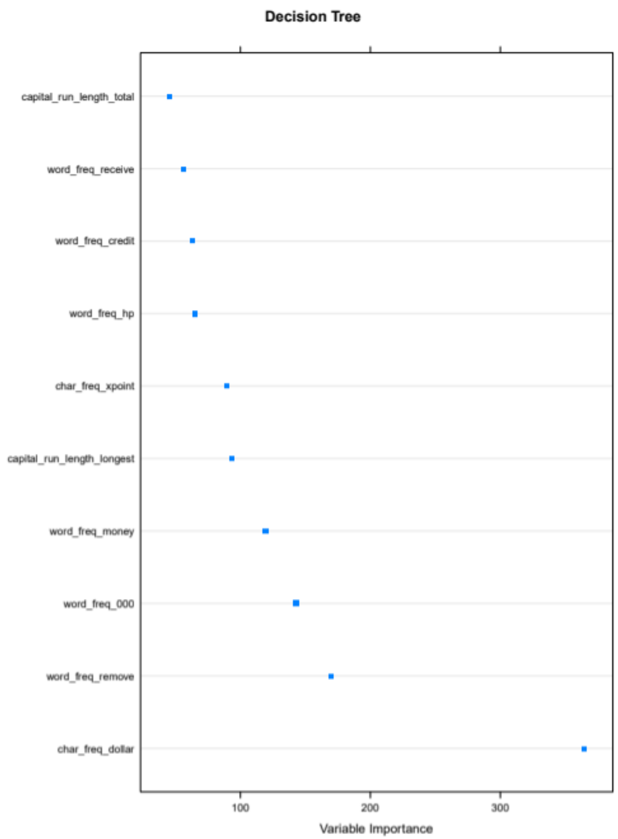
Decision Tree



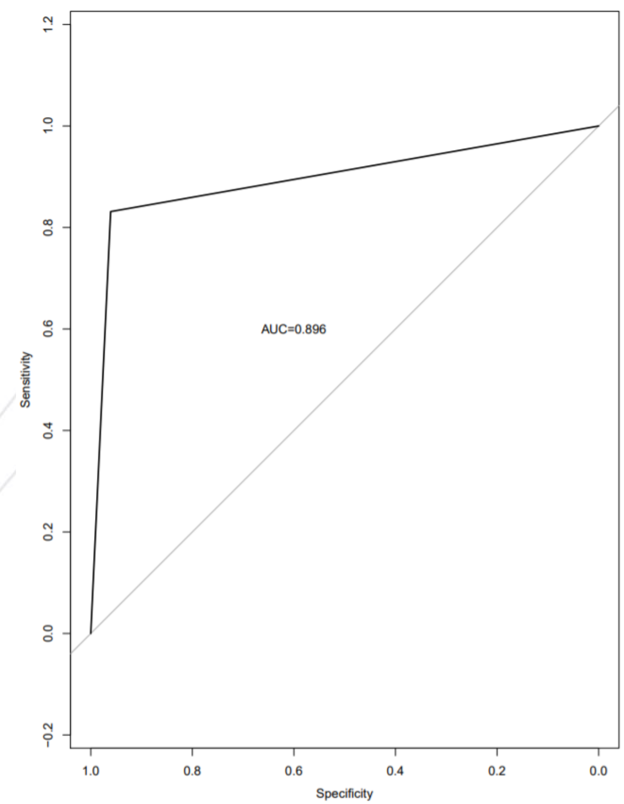
Northwestern

Decision Tree

- Variable Importance Plot



- ROC Curve



Decision Tree

- In-Sample Results

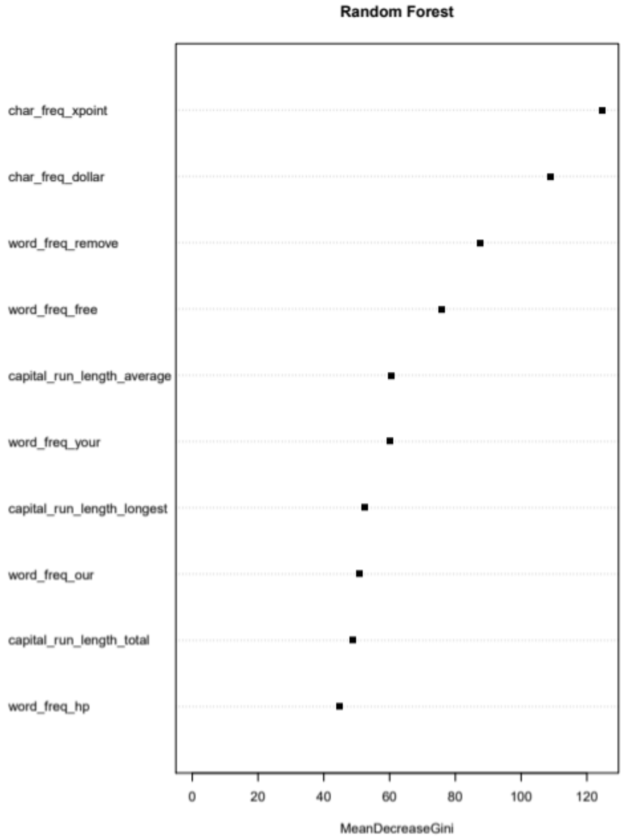
	0	1
0	.961	.039
1	.169	.831

- Out-of-Sample Results

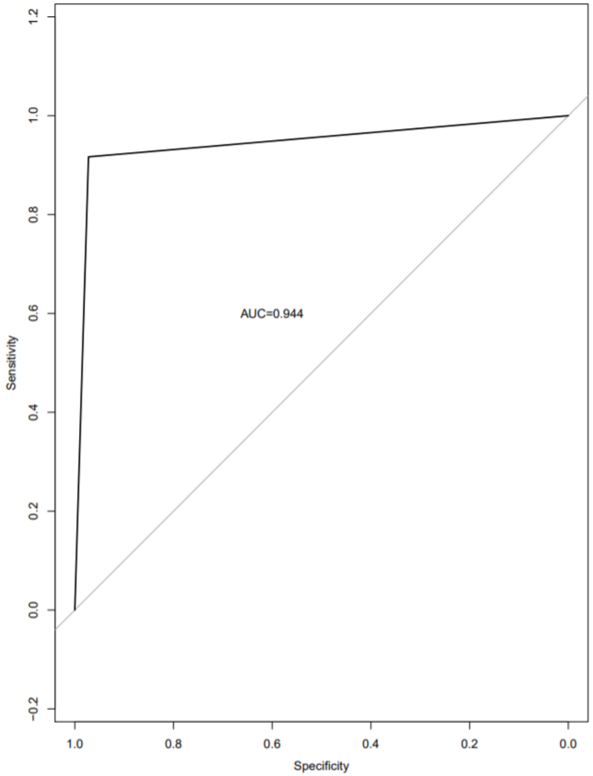
	0	1
0	.940	.060
1	.176	.824

Random Forest

- Variable Importance Plot



- ROC Curve



Random Forest

- In-Sample Results

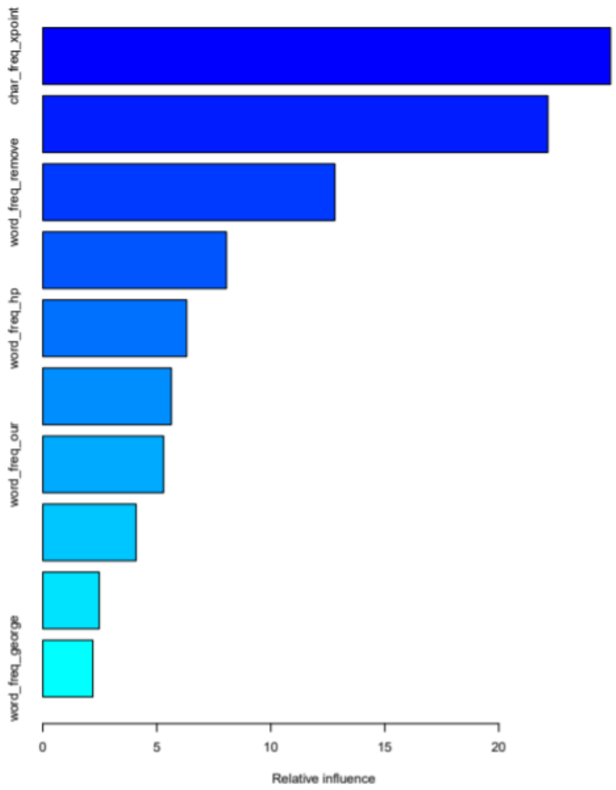
	0	1
0	.948	.052
1	.083	.917

- Out-of-Sample Results

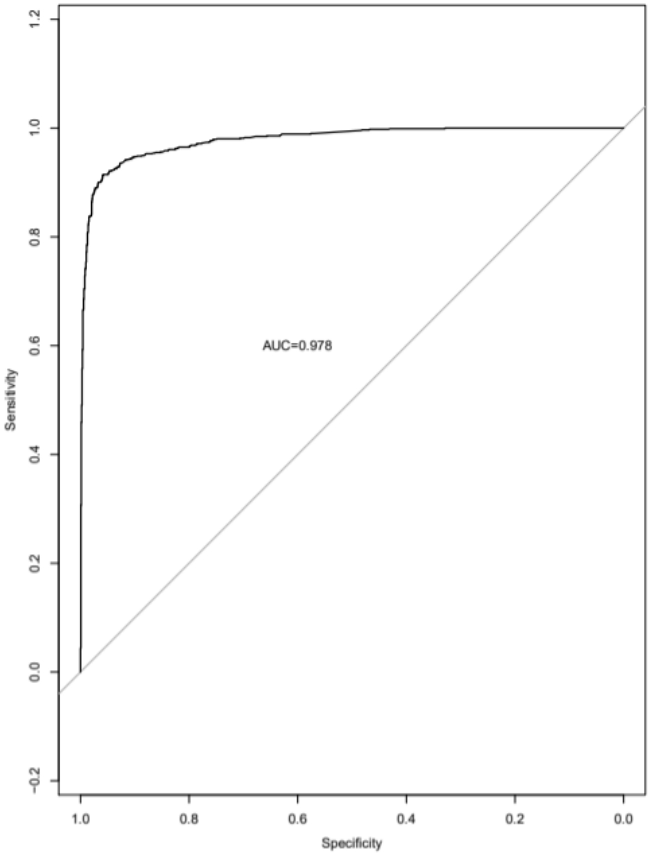
	0	1
0	.963	.037
1	.058	.942

GBM: Bernoulli

- Variable Importance Plot



- ROC Curve



GBM: Bernoulli

- In-Sample Results

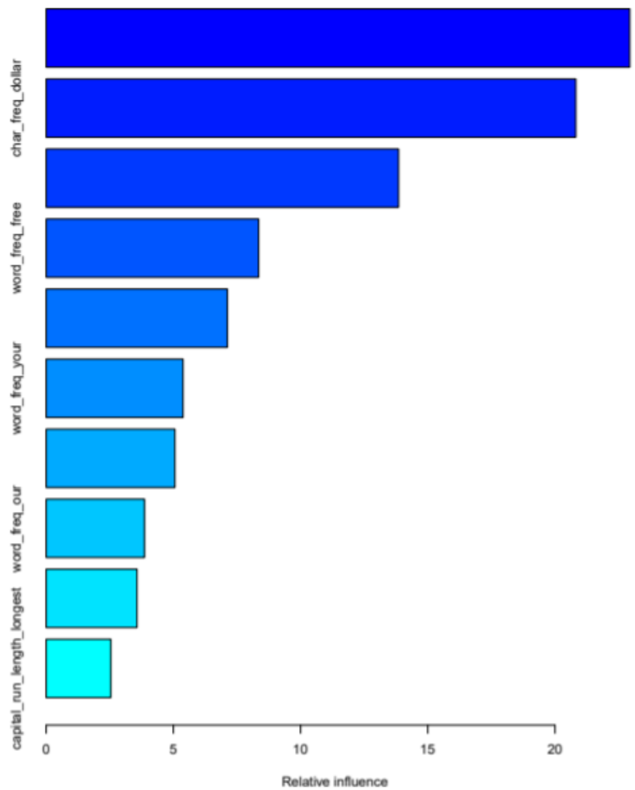
	0	1
0	.967	.033
1	.114	.886

- Out-of-Sample Results

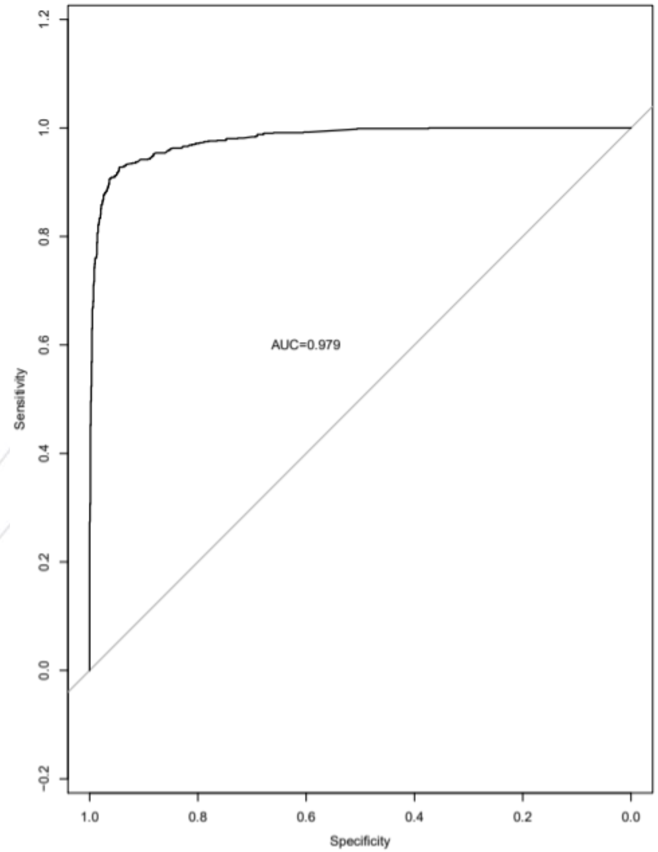
	0	1
0	.938	.062
1	.067	.933

GBM: AdaBoost

- Variable Importance Plot



- ROC Curve



GBM: AdaBoost

- In-Sample Results

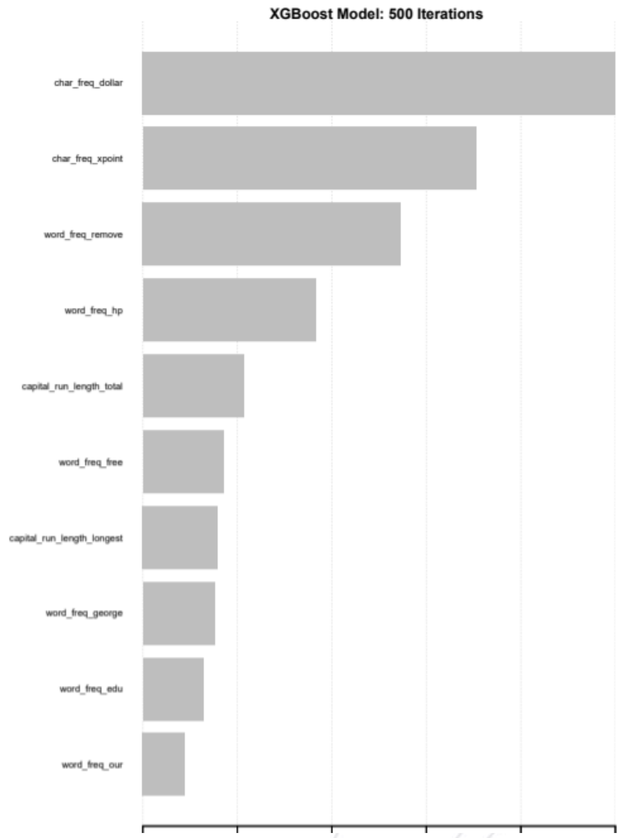
	0	1
0	.958	.042
1	.082	.918

- Out-of-Sample Results

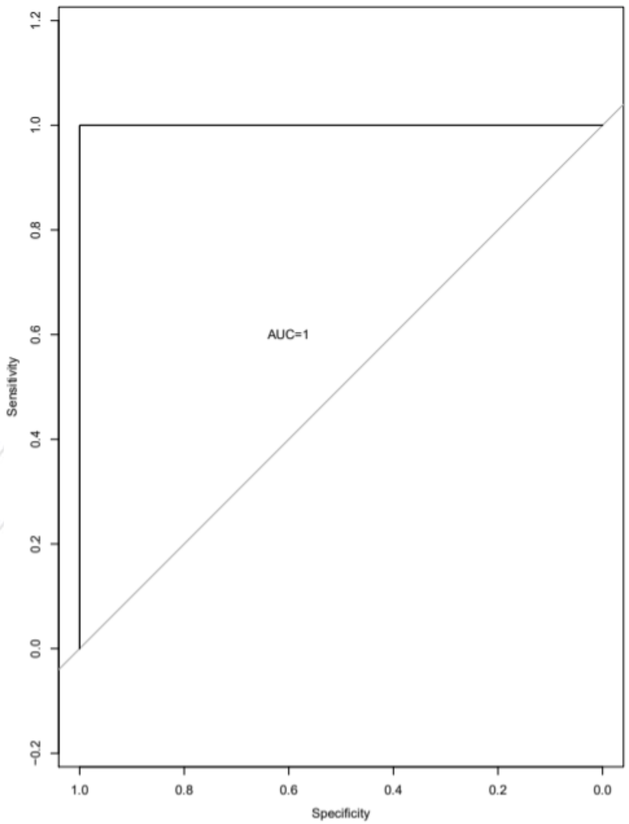
	0	1
0	.944	.056
1	.066	.934

XGBoost: 500 Iterations

- Variable Importance Plot



- ROC Curve



XGBoost: 500 Iterations

- In-Sample Results

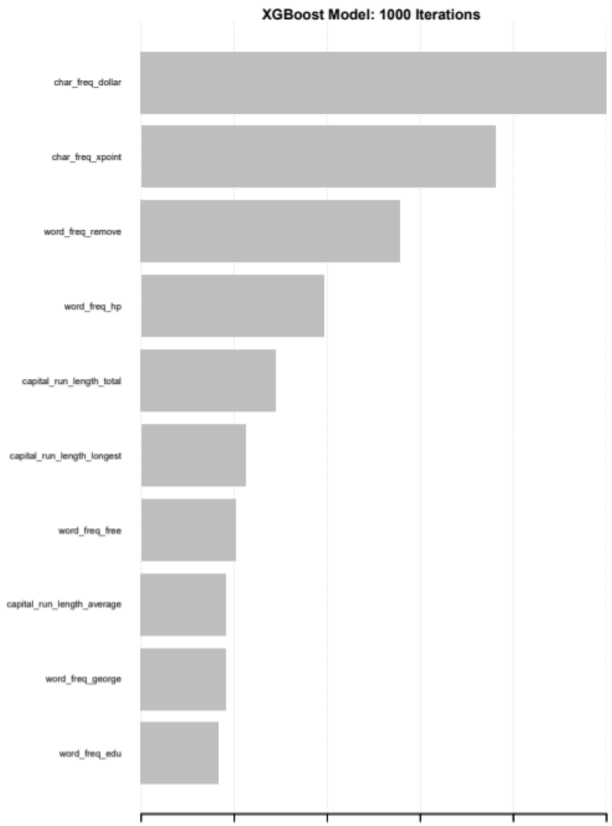
	0	1
0	.999	.001
1	.000	1.000

- Out-of-Sample Results

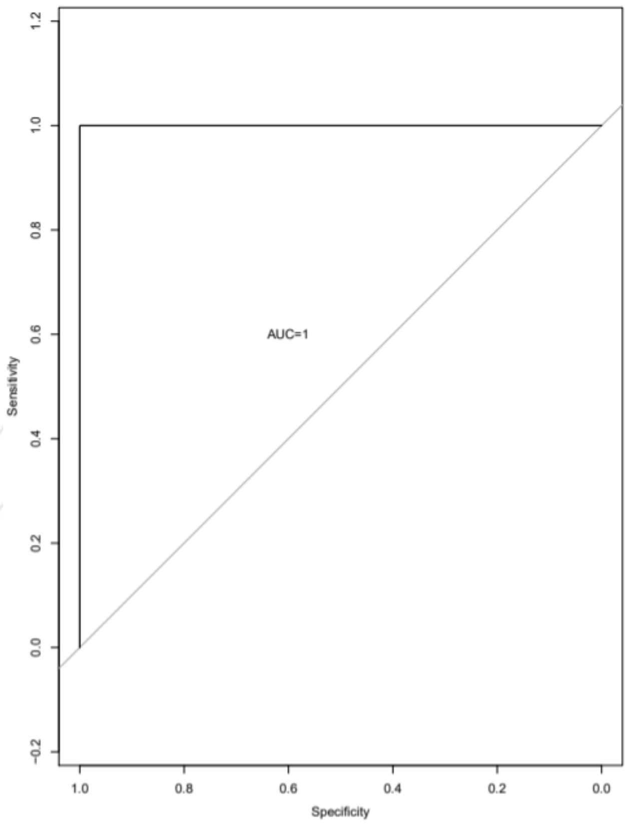
	0	1
0	.958	.042
1	.050	.950

XGBoost: 1000 Iterations

- Variable Importance Plot



- ROC Curve



XGBoost: 1000 Iterations

- In-Sample Results

	0	1
0	.999	.001
1	.000	1.000

- Out-of-Sample Results

	0	1
0	.957	.043
1	.054	.946

Model Comparison

- In-Sample Results

Metric	Decision.Tree	Random.Forest	GBM.Bernoulli	GBM.AdaBoost	XGBoost.500	XGBoost.1000
Accuracy	0.896	0.933	0.927	0.933	1.000	1.000
True Positive	0.831	0.918	0.886	0.918	1.000	1.000
True Negative	0.961	0.948	0.967	0.958	0.999	0.999
False Positive	0.169	0.082	0.114	0.082	0.000	0.000
False Negative	0.039	0.052	0.033	0.052	0.001	0.001
TP+TN	1.792	1.866	1.853	1.876	1.999	1.999
Precision	0.831	0.918	0.886	0.918	1.000	1.000
Recall	0.955	0.946	0.964	0.946	0.999	0.999
Specificity	0.850	0.920	0.895	0.921	1.000	1.000
F1	0.889	0.932	0.923	0.932	1.000	1.000
AUC	0.896	0.944	0.978	0.979	1.000	1.000

- Out-of-Sample Results

Metric	Decision.Tree	Random.Forest	GBM.Bernoulli	GBM.AdaBoost	XGBoost.500	XGBoost.1000
Accuracy	0.891	0.953	0.936	0.939	0.954	0.952
True Positive	0.824	0.942	0.933	0.934	0.950	0.946
True Negative	0.940	0.963	0.938	0.944	0.958	0.957
False Positive	0.176	0.058	0.067	0.066	0.050	0.054
False Negative	0.060	0.037	0.062	0.056	0.042	0.043
TP+TN	1.764	1.905	1.871	1.878	1.908	1.903
Precision	0.824	0.942	0.933	0.934	0.950	0.946
Recall	0.954	0.962	0.938	0.943	0.958	0.957
Specificity	0.842	0.943	0.933	0.935	0.950	0.947
F1	0.884	0.952	0.935	0.939	0.954	0.951
AUC	0.882	0.943	0.977	0.978	0.987	0.986

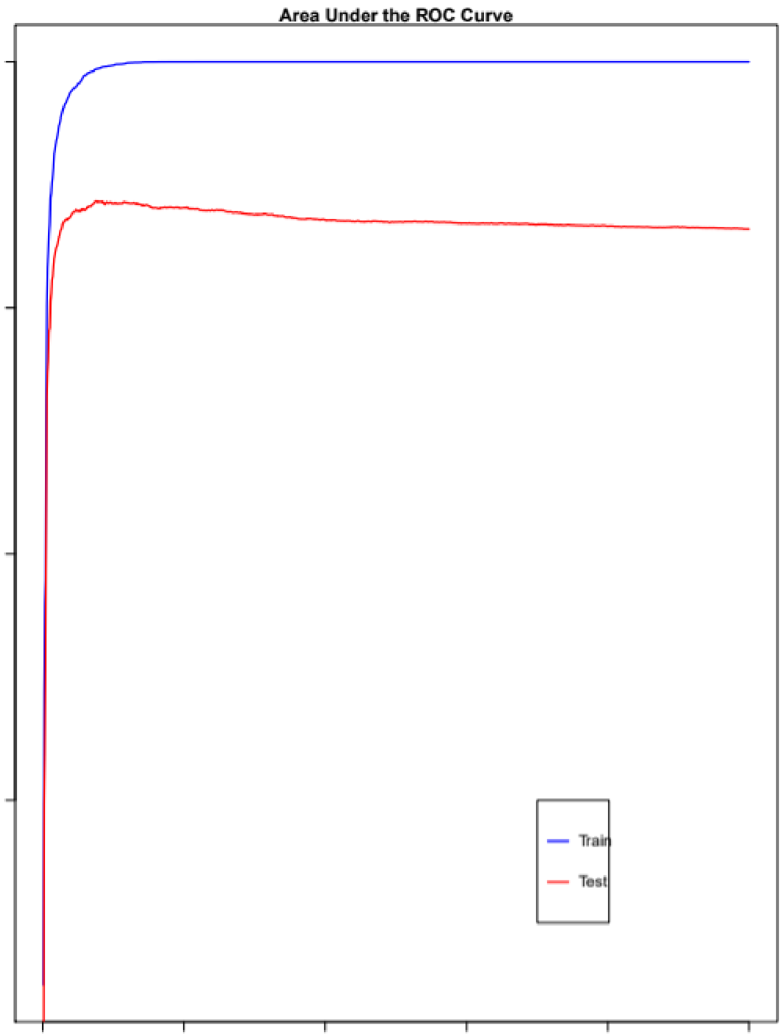
Model Comparison: F1 Score

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

- F1 maintains a balance between Precision and Recall
- Higher Recall important in spam classification
- XGBoost models have highest F1 in both datasets
- 500 Iteration model less overfit than 1000 Iteration model

Model	Train	Test
XGBoost: 500 Iterations	1.000	0.954
XGBoost: 1000 Iterations	1.000	0.951
Random Forest	0.932	0.952
GBM: Bernoulli	0.932	0.939
GBM: AdaBoost	0.923	0.935
Decision Tree	0.889	0.884

XGBoost: Overfitting



- Model begins to overfit at about 50 iterations
- Slippage at 50 iterations ~.010
- Slippage increases to ~.011 at 200 iterations
- Slippage at 1000 iterations ~.0135