

Can Logistic Regression and/or Feature Selection Predict Contaminated Wells? A Case Study

B. Dixon* & Nivedita C
University of South Florida – St. Petersburg

ABSTRACT

Detection of potentially contaminated wells is an important component of environmental protection and management. However, contamination potential mapping is not an easy task due to inherent uncertainties. This study aims at assessing suitability of various techniques in predicting contaminated wells for example logistic regression, feature selection, Neural Networks (NN) and Support Vector Machines (SVM).

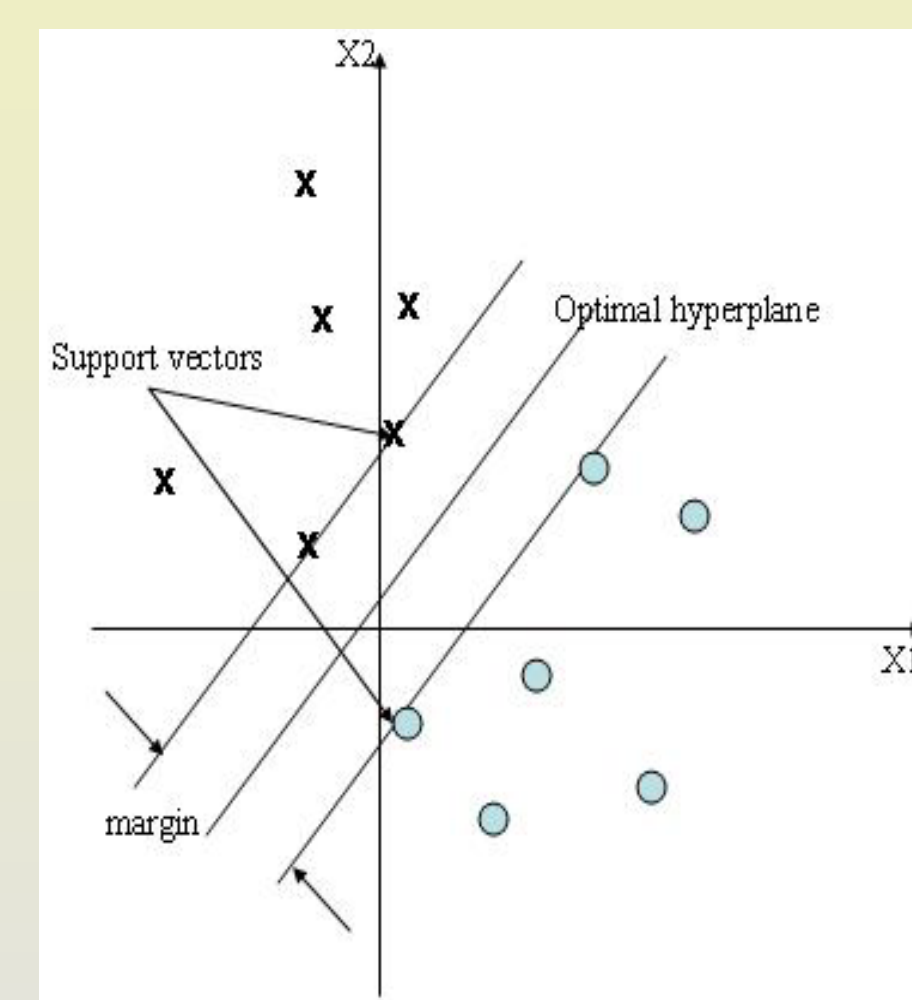
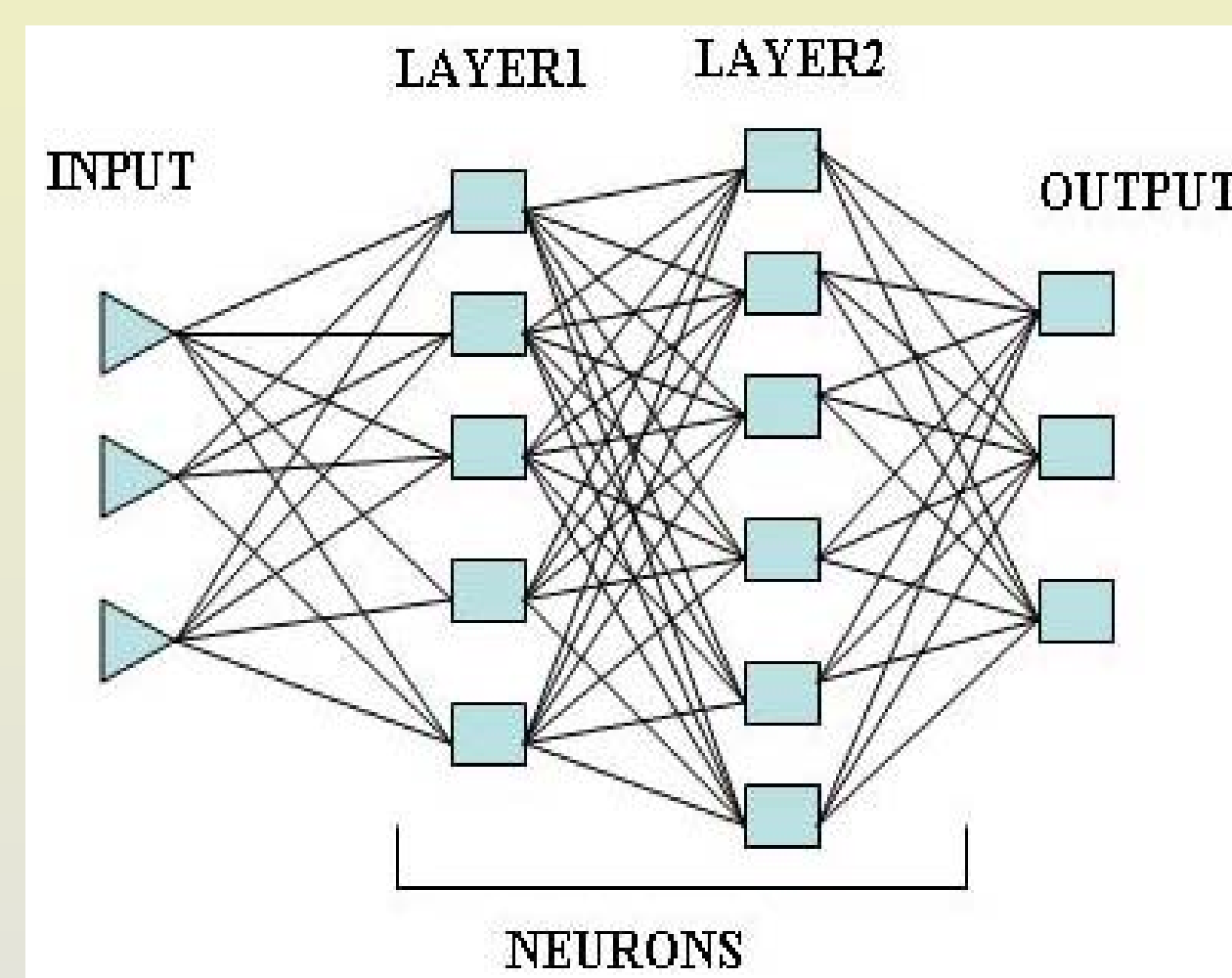
Contamination potential depends on complex interactions of hydro-geological variables. A large number of input variables add to redundancy, cost and time. The logistic regression, feature selection methods were used to identify critical variables in transporting contaminants in and through the soil profile. NN and SVM were used to identify contaminated wells. Variables used in this study included DRASTIC parameters, soil structure (pedality), hydrologic group, landuse, pH, organic matter and bulk density. Well data (nitrate-N) provided by Florida Department of Environmental Protection (FDEP) as part of the Water Supply Restoration Program (WSRP) were used in this study as target class.

The objective of this study was three- fold:

(a) Analyze the input variables and identify the most significant predictors of well contamination by performing feature selection to identify the best subset of variables. (b) Use all the input variables with the NN and SVM to classify wells and compare their performances. (c) Repeat the above (step b) with the variable subset from step (a) and compare results.

Classifiers were compared based on their accuracies and parameters such as sensitivity and specificity. Free Receiver Operating Curves (FROCs) were used for evaluation of classifier performance.

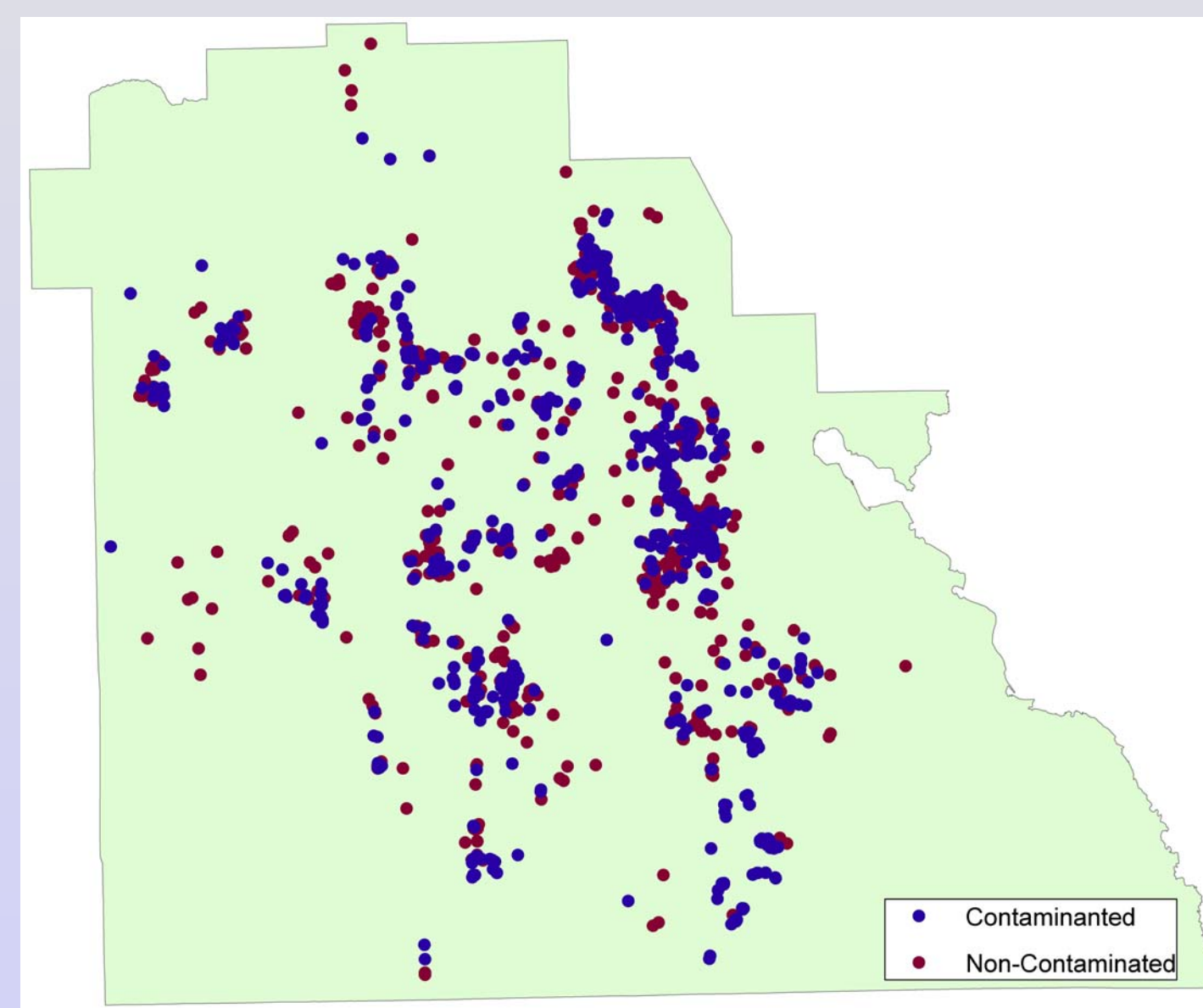
Results show superior performance with the NN as compared to SVM especially on training data. Results on testing data are comparable. Feature selection did not improve accuracy. However, it helped increase the sensitivity or the true positive rate (TPR). Thus, a higher TPR was obtainable with fewer features. In this study, higher TPR is desirable since the cost of detecting a contaminated well incorrectly is far higher than a non-contaminated well going undetected.



MATERIALS AND METHODS

Study Area and Source Data

This work focused on the prediction and classification of well contamination in Polk County, Florida. Well data (nitrate-N) provided by FLDEP as part of the WSRP were used. The wells were tested for nitrate-N concentration and those with a concentration above 3 mg/L were considered contaminated. The figure below shows the location of contaminated and non-contaminated wells in the study area.



Training data sample

| D | R | A | S | T | I | C | lulc_risk | pedality | drainage | hydro_gp | pH | OM | BD | cont/not_cont* |
|----|---|---|---|----|----|---|-----------|----------|----------|----------|----|----|----|----------------|
| 9 | 9 | 8 | 9 | 10 | 9 | 2 | 2 | 1 | 5 | 3 | 9 | 21 | 4 | 0 |
| 9 | 9 | 7 | 8 | 10 | 9 | 2 | 2 | 4 | 4 | 3 | 7 | 14 | 8 | 0 |
| 5 | 9 | 8 | 9 | 9 | 8 | 1 | 2 | 1 | 1 | 1 | 10 | 9 | 10 | 0 |
| 10 | 9 | 7 | 9 | 10 | 9 | 2 | 2 | 1 | 1 | 1 | 10 | 9 | 10 | 0 |
| 9 | 9 | 7 | 9 | 5 | 9 | 2 | 2 | 1 | 1 | 1 | 10 | 9 | 10 | 0 |
| 10 | 9 | 7 | 9 | 10 | 8 | 1 | 2 | 1 | 3 | 3 | 14 | 13 | 10 | 0 |
| 10 | 9 | 7 | 9 | 10 | 9 | 2 | 3 | 1 | 1 | 1 | 10 | 9 | 10 | 0 |
| 10 | 9 | 7 | 9 | 10 | 9 | 4 | 2 | 1 | 1 | 1 | 11 | 9 | 10 | 0 |
| 7 | 9 | 7 | 9 | 5 | 9 | 4 | 2 | 1 | 1 | 1 | 10 | 9 | 10 | 1 |
| 10 | 9 | 7 | 9 | 10 | 9 | 2 | 4 | 1 | 1 | 1 | 11 | 9 | 10 | 1 |
| 7 | 9 | 8 | 9 | 10 | 8 | 2 | 2 | 1 | 1 | 1 | 10 | 9 | 10 | 1 |
| 9 | 9 | 8 | 9 | 10 | 8 | 2 | 2 | 6 | 4 | 4 | 6 | 10 | 9 | 1 |
| 9 | 9 | 8 | 9 | 10 | 8 | 2 | 3 | 1 | 1 | 1 | 10 | 9 | 10 | 1 |
| 10 | 9 | 8 | 9 | 10 | 10 | 2 | 4 | 1 | 1 | 1 | 10 | 9 | 10 | 1 |
| 10 | 9 | 7 | 9 | 10 | 9 | 1 | 2 | 1 | 1 | 1 | 10 | 9 | 10 | 1 |

*0=not contaminated, 1= contaminated

EVALUATION

Evaluation of the classification algorithms was performed using two measures; *Accuracy* and *Confusion Matrix*. ROC curves were plotted by varying the threshold on the predicted output. The following table shows the confusion matrix for a binary classifier:

| | +1 | -1 |
|----|-----------|-----------|
| +1 | TP | FN |
| -1 | FP | TN |

Where: TP = number of correct predictions that an instance is positive
 FP = number of incorrect predictions that an instance is positive
 TN = number of correct predictions that an instance is negative (or zero)
 FN = number of incorrect predictions that an instance is negative (or zero)

The ROC curve gives a graphical representation of these parameters for various thresholds on the output and encapsulates all the information contained in the confusion matrix. Here the FPR is plotted on the x-axis vs. the TPR on the y-axis. Each threshold results in a (TPR, FPR) pair and a series of such pairs are used to plot the ROC curve. In our case, the TPR would be the probability of correctly classifying a contaminated well as contaminated. The FPR is the probability of incorrectly classifying a non-contaminated well as contaminated. These are also known as the *Sensitivity (TPR)* and *Specificity (1-FPR)*. As the sensitivity of the classifier is increased, the specificity is also sacrificed. Thus, an optimum cutoff needs to be chosen, for which these values are acceptable.

Classification and evaluation based on SVM and NN algorithms was carried out and their performances were compared. Also, the performances of these algorithms using all features and the most significant features were compared.

Univariate statistics of input variables- contaminated wells (n=919)

| | Mean | Standard Error | Median | Mode | Standard Deviation | Sample Variance | Minimum | Maximum |
|----------|-------|----------------|--------|-------|--------------------|-----------------|---------|---------|
| D | 8.96 | 0.05 | 9.00 | 10.00 | 1.46 | 2.14 | 3.00 | 10.00 |
| R | 9.00 | 0.00 | 9.00 | 9.00 | 0.00 | 0.00 | 9.00 | 9.00 |
| A | 7.29 | 0.02 | 7.00 | 7.00 | 0.45 | 0.21 | 7.00 | 8.00 |
| S | 8.68 | 0.04 | 9.00 | 9.00 | 1.33 | 1.76 | 2.00 | 10.00 |
| T | 9.78 | 0.02 | 10.00 | 10.00 | 0.55 | 0.30 | 5.00 | 10.00 |
| I | 8.78 | 0.02 | 9.00 | 9.00 | 0.45 | 0.21 | 8.00 | 10.00 |
| C | 1.93 | 0.03 | 2.00 | 2.00 | 0.84 | 0.70 | 1.00 | 4.00 |
| drainage | 2.12 | 0.05 | 1.00 | 1.00 | 1.54 | 2.38 | 1.00 | 6.00 |
| hydro | 1.60 | 0.04 | 1.00 | 1.00 | 1.12 | 1.26 | 1.00 | 5.00 |
| pedality | 1.58 | 0.05 | 1.00 | 1.00 | 1.38 | 1.91 | 1.00 | 6.00 |
| lulc95 | 2.46 | 0.03 | 2.00 | 2.00 | 0.89 | 0.79 | 1.00 | 4.00 |
| pH | 9.02 | 0.09 | 10.00 | 10.00 | 2.62 | 6.88 | 3.00 | 19.00 |
| OM | 9.62 | 0.11 | 9.00 | 9.00 | 3.43 | 11.76 | 2.00 | 25.00 |
| BD | 10.30 | 0.07 | 10.00 | 10.00 | 2.01 | 4.05 | 3.00 | 16.00 |

Univariate statistics of input variables - non contaminated wells (n=5739)

| | Mean | Standard Error | Median | Mode | Standard Deviation | Sample Variance | Minimum | Maximum |
|----------|-------|----------------|--------|-------|--------------------|-----------------|---------|---------|
| D | 8.87 | 0.02 | 9.00 | 10.00 | 1.45 | 2.09 | 1.00 | 10.00 |
| R | 9.00 | 0.00 | 9.00 | 9.00 | 0.00 | 0.00 | 9.00 | 9.00 |
| A | 7.37 | 0.01 | 7.00 | 7.00 | 0.49 | 0.24 | 7.00 | 9.00 |
| S | 8.72 | 0.02 | 9.00 | 9.00 | 1.21 | 1.47 | 2.00 | 10.00 |
| T | 9.77 | 0.01 | 10.00 | 10.00 | 0.72 | 0.52 | 5.00 | 10.00 |
| I | 8.76 | 0.01 | 9.00 | 9.00 | 0.44 | 0.20 | 8.00 | 10.00 |
| C | 1.99 | 0.01 | 2.00 | 2.00 | 0.89 | 0.79 | 1.00 | 4.00 |
| drainage | 2.41 | 0.03 | 2.00 | 1.00 | 2.65 | 7.03 | 1.00 | 6.00 |
| hydro | 1.95 | 0.02 | 1.00 | 1.00 | 1.82 | 3.31 | 1.00 | 6.00 |
| pedality | 1.94 | 0.02 | 1.00 | 1.00 | 1.74 | 3.01 | 1.00 | 6.00 |
| lulc95 | 2.39 | 0.01 | 2.00 | 2.00 | 0.90 | 0.81 | 0.00 | 4.00 |
| pH | 8.62 | 0.04 | 10.00 | 10.00 | 3.06 | 9.34 | 1.00 | 19.00 |
| OM | 9.47 | 0.05 | 9.00 | 9.00 | 4.14 | 17.16 | 1.00 | 25.00 |
| BD | 10.35 | 0.03 | 10.00 | 10.00 | 2.25 | 5.07 | 1.00 | 16.00 |

Logistic fit of outcome by individual predictors

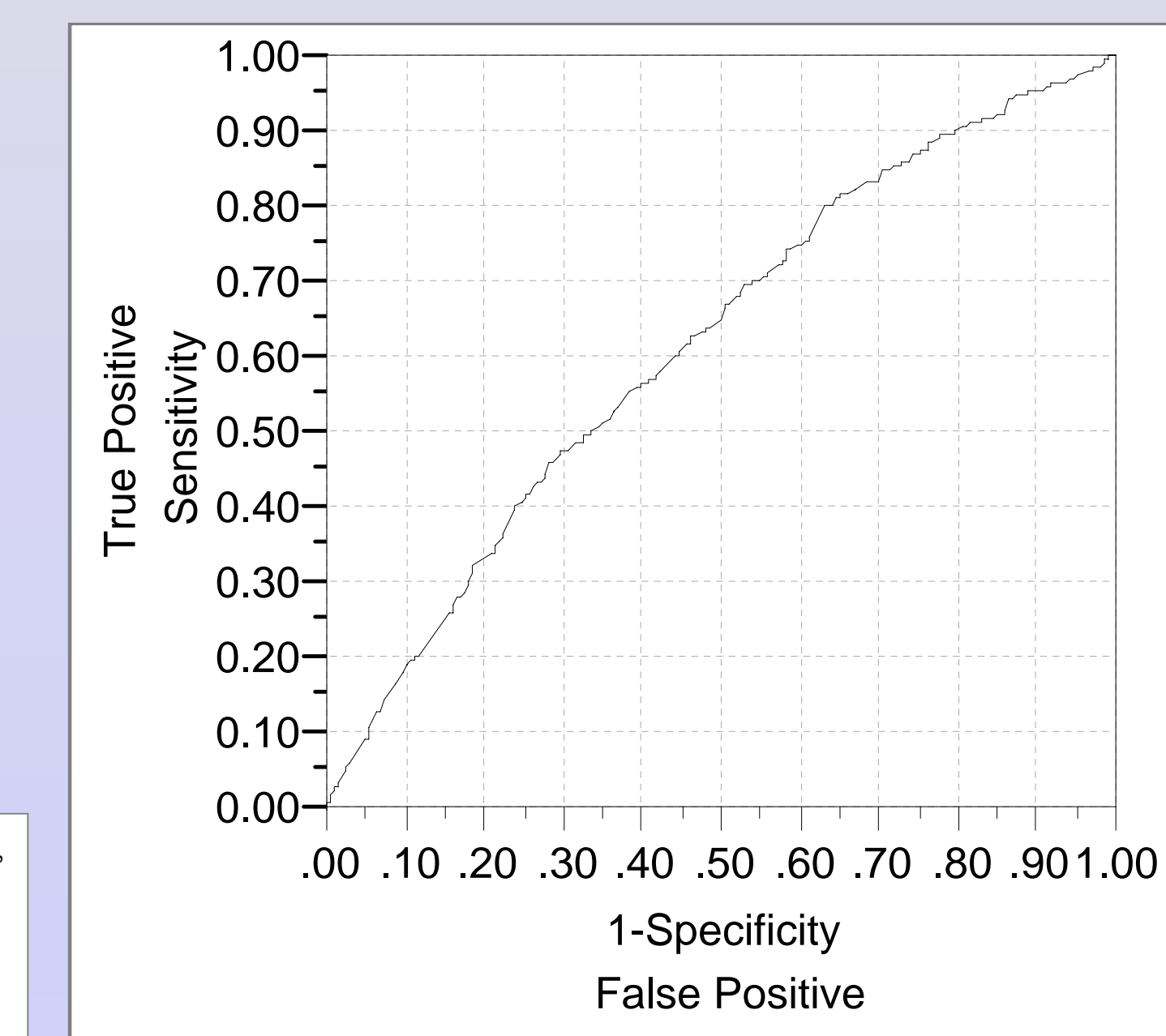
| Term | Parameter Estimate (β) | Std Error | ChiSquare | Prob>ChiSq |
|-----------|------------------------|-------------|--------------|------------------|
| Intercept | 2.11 | 609167.61 | 0.00 | 1.00 |
| D | -0.06 | 0.04 | 2.06 | 0.15 |
| R | -0.01 | 67685.29 | 0.00 | 1.00 |
| A | 0.23 | 0.09 | 7.33 | 0.01 |
| S | 0.03 | 0.03 | 1.05 | 0.31 |
| T | -0.11 | 0.06 | 3.33 | 0.07 |
| I | 0.01 | 0.12 | 0.00 | 0.97 |
| C | 0.11 | 0.04 | 6.83 | 0.01 |
| drainage | -0.01 | 0.03 | 0.17 | 0.68 |
| hydro | 0.24 | 0.05 | 29.31 | <.0001 |
| pedality | 0.05 | 0.03 | 2.51 | 0.11 |
| lulc95 | -0.07 | 0.04 | 3.11 | 0.08 |
| pH | -0.07 | 0.02 | 20.71 | <.0001 |
| OM | -0.05 | 0.02 | 12.21 | 0.00 |
| BD | -0.01 | 0.03 | 0.18 | 0.67 |

Interpretation of β: The parameter estimate (β) gives the increase in log odds of the outcome, for one unit increase in x i.e. e^β represents the change in odds of the outcome, by increasing x by 1 unit. Given below is the interpretation of β:

- If β=0, the odds and probability are the same at all x levels (e^β =1)
- If β>0, the odds and probability increase as x increases (e^β >1)
- If β<0, the odds and probability decrease as x increases (e^β <1)

From the value of β it can be seen that variables A and hydrologic group have the highest strength of association with the outcome. Variables A, S, I, C, hydrologic group and pedality are positively related to the outcome.

The ROC curve for a complete model (consisting of all input variables) without feature selection is shown as:



Area under Curve = 0.61537

For a threshold of 0.15,
 Accuracy= 0.6285
 Confusion matrix =
 0.6457 0.3543
 0.3742 0.6258

Creation of training and testing datasets

The input data layers used were Depth to Ground Water (D), Recharge of aquifer (R), Aquifer media (A), Soil media (S), Topography (T), Impact of vadose zone (I), hydraulic Conductivity (C), Landuse (LULC), pedality, drainage, hydrologic group, pH, Organic Matter (OM) and Bulk Density (BD) i.e. 14 input layers/ variables were used in this study.

Data analysis and Classification

This was performed in three different stages:

Input feature analysis and feature selection using Stepwise Forward Selection (SFS) method
 Use all the input variables with the NN and SVM and compare their performances
 Include the most significant features (from Step 1) with the NN and SVM and compare their performances.

Selection Methods Be Used to Predict Study of Polk County, Florida

edita Candade

ersburg, Geo-Spatial Analytics Lab

Feature selection using SFS

The above univariate analysis gives an interpretation of individual features and their individual relationships with the outcome. The SFS gives the best feature subset. The results from SFS with only the main predictors included are as shown in Table 5. This is called the *main effect* model.

The SFS procedure selected the following input variables as a good subset (and significant when considered together): pedality, Aquifer media (A), soil hydrologic group, Hydraulic Conductivity (C), pH, Organic Matter (OM), Topography (T), landuse and Depth to GW (D).

The Area Under the ROC Curve (AUC) is used to determine the discriminating power of the logistic model, which is close to 1 for a model that discriminates perfectly. The AUC for the main effect model is *0.61537*.

| Kernel | | Training | | Testing | |
|---|--------|-------------|------------------|-------------|------------------|
| | | Accuracy | Confusion matrix | Accuracy | Confusion matrix |
| RBF kernel, g=7 (equal weights for classes) | c=1000 | 0.88 | 0.26 0.74 | 0.86 | 0.12 0.88 |
| | | | 0.01 0.98 | | 0.03 0.97 |
| | c=100 | 0.88 | 0.26 0.74 | 0.86 | 0.12 0.88 |
| | | | 0.01 0.98 | | 0.03 0.97 |
| | c=10 | 0.88 | 0.26 0.74 | 0.86 | 0.12 0.88 |
| | | 0.01 0.98 | | 0.03 0.97 | |
| | c=5 | 0.88 | 0.26 0.74 | 0.86 | 0.12 0.88 |
| | | | 0.01 0.98 | | 0.03 0.97 |
| RBF kernel (with c=50 for class 1 and c=10 for class -1) | g=9 | 0.75 | 0.84 0.16 | 0.62 | 0.65 0.35 |
| | | | 0.26 0.74 | | 0.39 0.61 |
| | g=7 | 0.75 | 0.84 0.16 | 0.63 | 0.64 0.36 |
| | | | 0.26 0.74 | | 0.37 0.63 |
| | g=5 | 0.14 | 1 0 | 0.14 | 1 0 |
| | | | 1 0 | | 1 0 |
| Polynomial kernel (with c=50 for class 1 and c=10 for class -1) | d=7 | 0.62 | 0.64 0.36 | 0.6 | 0.6 0.4 |
| | | | 0.38 0.62 | | 0.4 0.6 |
| | d=3 | 0.6 | 0.72 0.28 | 0.57 | 0.66 0.34 |
| | | | 0.42 0.58 | | 0.44 0.56 |

* All numerical values rounded to two decimal places

The confusion matrix is interpreted as follows:

| | | |
|---|-----|-----|
| | 1 | 0 |
| 1 | TPR | FNR |
| 0 | FPR | TNR |

1=contaminated well, 0= non-contaminated well

TPR= proportion of contaminated wells correctly detected

TNR= proportion of non-contaminated wells correctly detected

FPR= proportion of non-contaminated wells wrongly classified as contaminated

FNR= proportion of contaminated wells wrongly classified as non-contaminated

With feature selection, the TPR on training data increased from 0.7168 to 0.7625. Same was observed with testing data where TPR increased from 0.6652 to 0.7065. Thus, feature selection helped improve the sensitivity or the true positive rate of the models.

Thus, with feature selection, the TPR on the training data increased from 0.8148 to 0.8671. This implies that the sensitivity of the model or the rate of detecting a 'positive' or a contaminated well increased to 0.8671. However, this was at the cost of reduction in the TNR or the specificity from 0.6863 to 0.6368. Since detection of contaminated wells is more crucial in our study, feature selection turned out to be an important step. With feature selection, the TPR on testing data increased from 0.6457 to 0.6783. However, the TNR reduced from 0.6258 to 0.5763. The AUC dropped from 0.6810 to 0.6525.

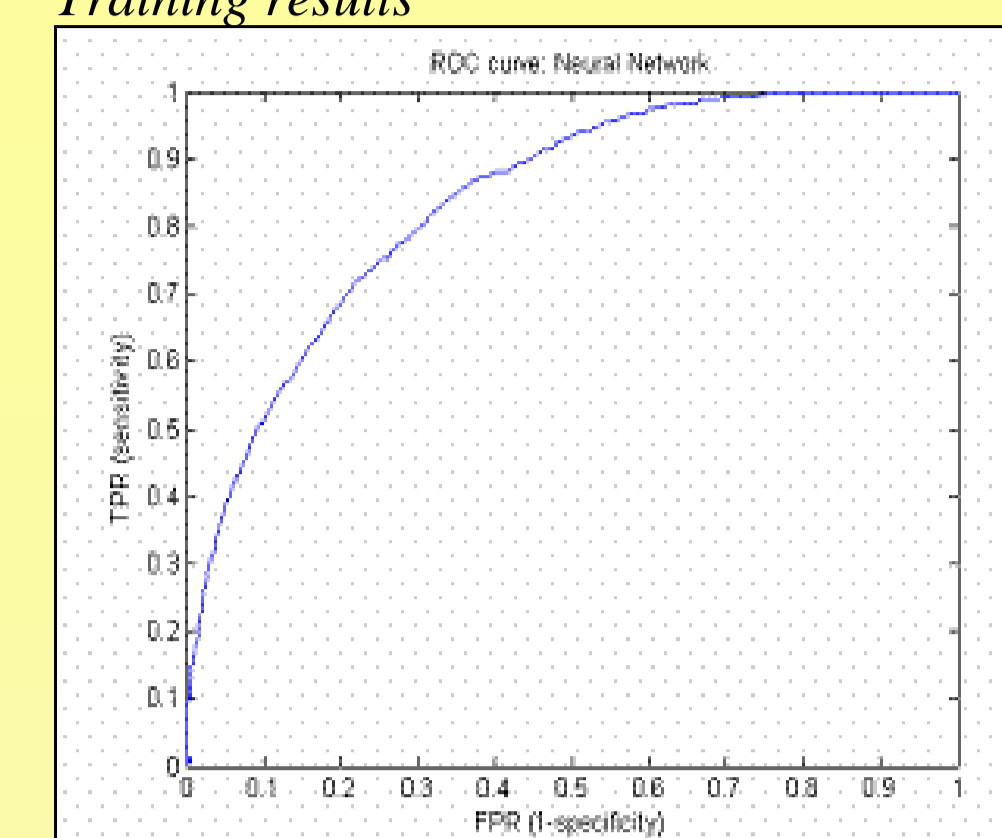
Summary of results from NN and SVM

| Input | | | NN | | SVM | |
|-------------------|----------|------------------|-------------|-------------|-------------|-------------|
| All 14 features | Training | AUC | 0.84 | | 0.72 | |
| | | Accuracy | 0.70 | | 0.6 | |
| | | confusion matrix | 0.81 | 0.19 | 0.72 | 0.28 |
| | Testing | AUC | 0.68 | | 0.66 | |
| | | Accuracy | 0.63 | | 0.58 | |
| | | confusion matrix | 0.65 | 0.35 | 0.67 | 0.33 |
| Feature selection | Training | AUC | 0.84 | | 0.71 | |
| | | Accuracy | 0.67 | | 0.58 | |
| | | confusion matrix | 0.87 | 0.13 | 0.76 | 0.24 |
| | Testing | AUC | 0.65 | | 0.65 | |
| | | Accuracy | 0.59 | | 0.55 | |
| | | confusion matrix | 0.68 | 0.32 | 0.71 | 0.29 |

Neural Network

1. Results using all 14 variables

Training results



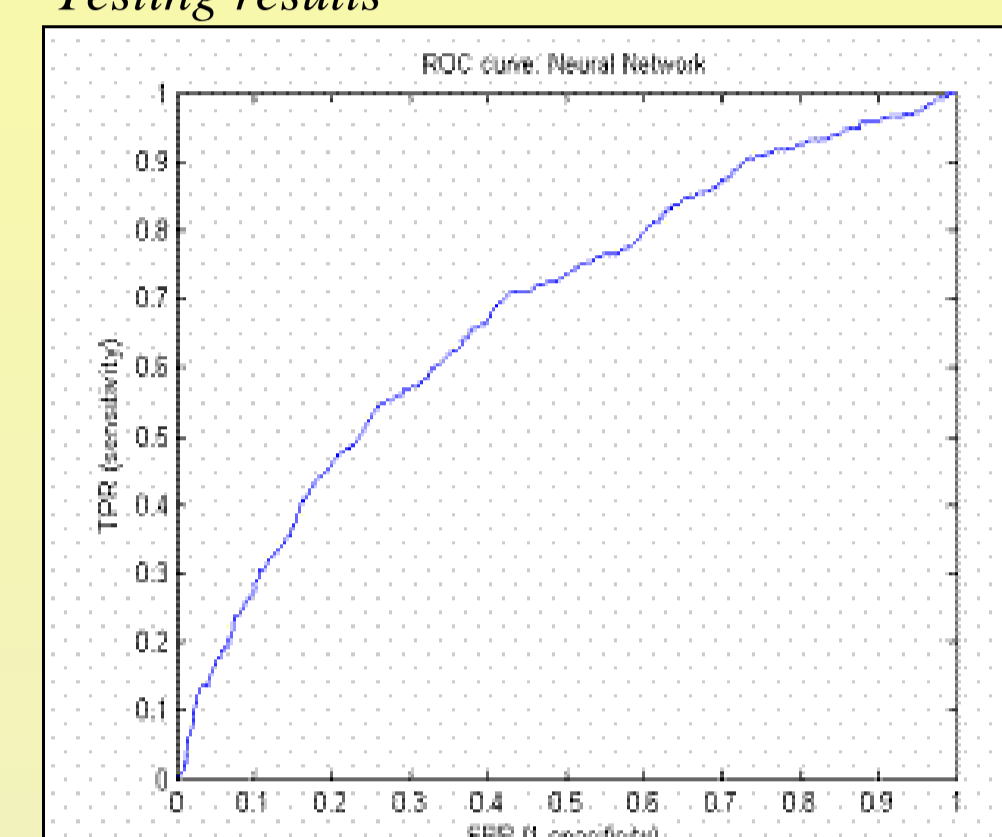
For a threshold of 0.15,

Accuracy = 0.7040

Confusion matrix =
0.8148 0.1852
0.3137 0.6863

Area under Curve (AUC) = 0.8385

Testing results



For a threshold of 0.15,

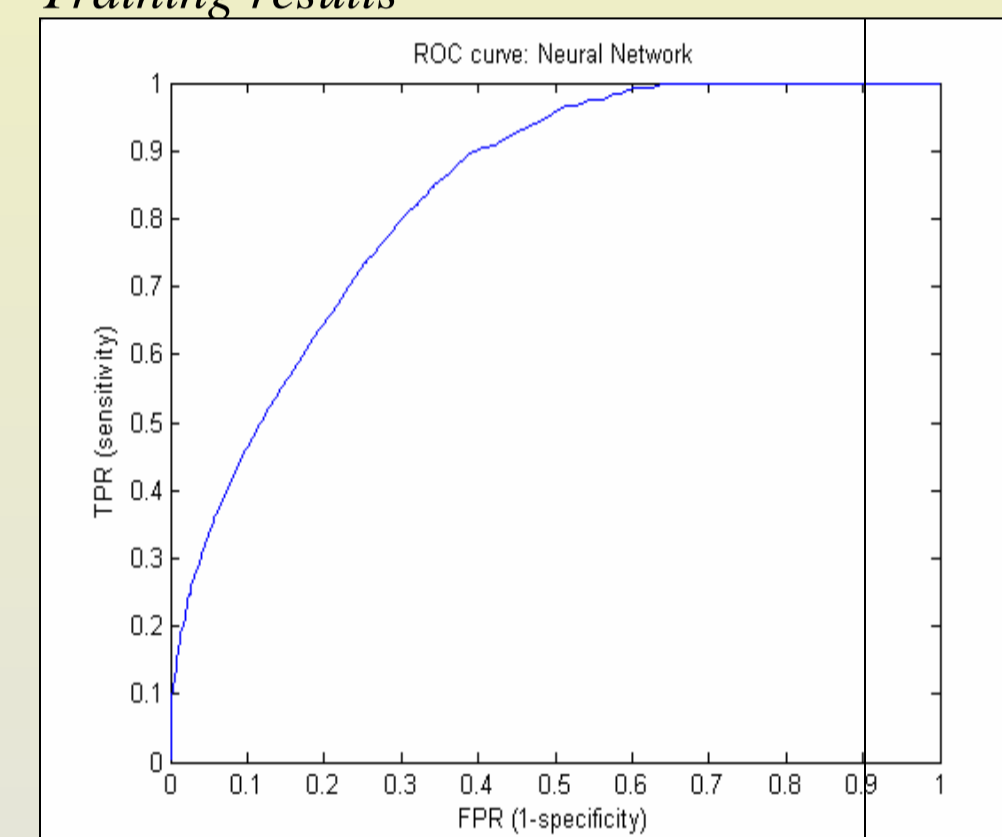
Accuracy = 0.6285

Confusion matrix =
0.6457 0.3543
0.3742 0.6258

Area under Curve (AUC) = 0.6810

2. Results using variables from feature selection

Training results



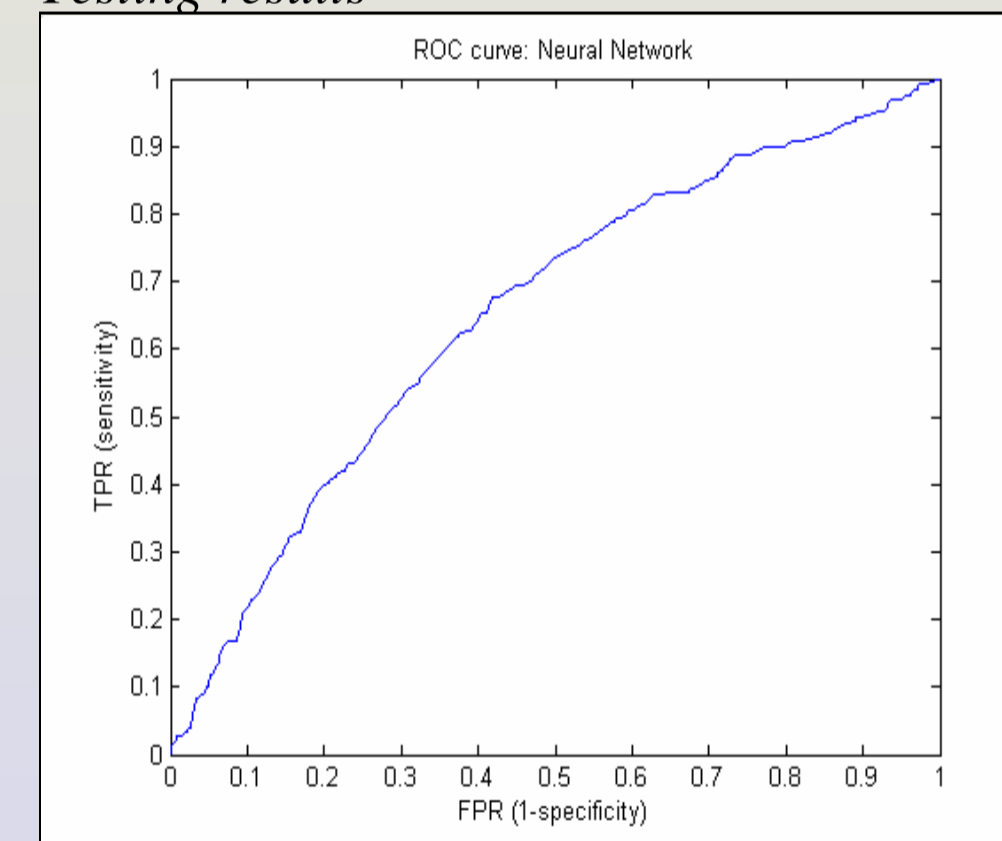
For a threshold of 0.15,

Accuracy = 0.6686

Confusion matrix =
0.8671 0.1329
0.3632 0.6368

Area under Curve = 0.8350

Testing results



For a threshold of 0.15,

Accuracy = 0.5904

Confusion matrix =
0.6783 0.3217
0.4237 0.5763

Area under Curve = 0.6525

CONCLUSIONS AND FUTURE WORK

- Univariate analysis did not show discrimination between the contaminated and non-contaminated classes.
- NN and SVM and the selection of features using logistic regression and SFS showed good classification accuracy and improved TPR.
- NN outperformed the SVM in our study. (This goes to show that SVM is sensitive to various parameters such as choice of kernel, careful selection of kernel parameters etc.)
- SVM showed good performance on unseen data thus proving its generalization performance.
- Accuracy was not considered a good measure of performance in our study as the dataset was highly unbalanced.
- Considering the TPR as a key parameter, it is seen that feature selection was a very essential step.
- On training data, the highest observed TPR was 0.8671 with the NN while on testing the highest was 0.7065 with the SVM. (Both these results are for the dataset using variables from the feature selection process.)

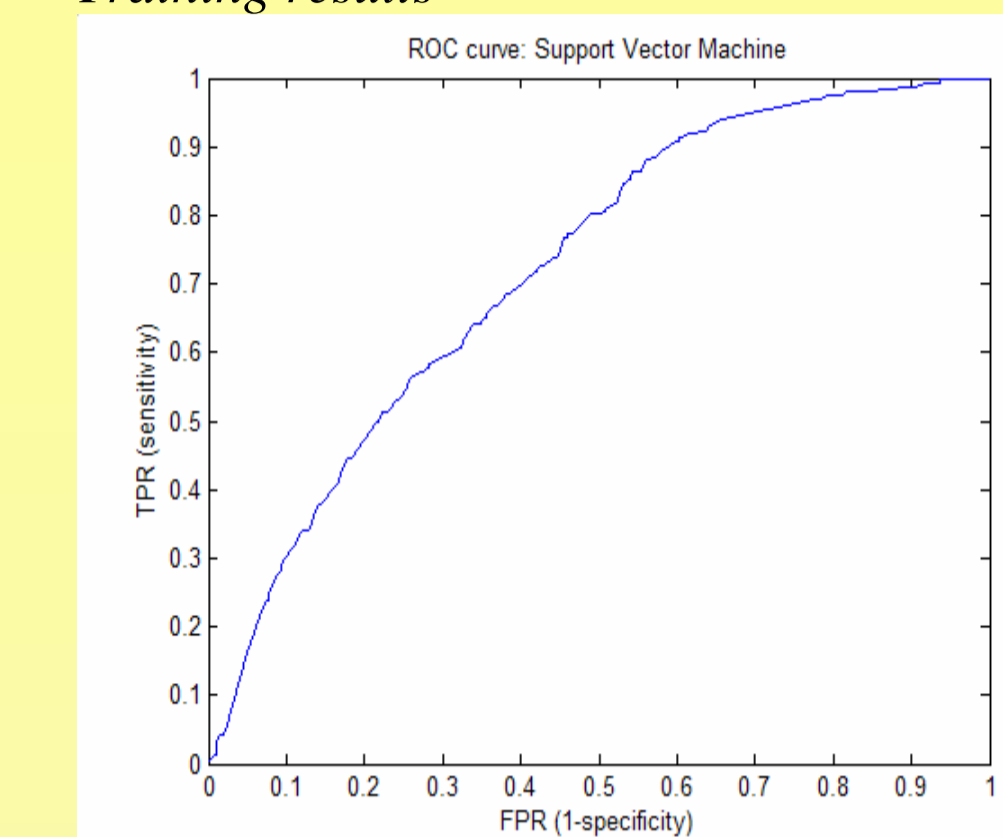
Future work would include the addition of more wells since a large, representative dataset is essential for a good prediction model. The SVM needs to be explored further with other parameters and kernels. The use of other feature selection methods and classification algorithms could help improve performance of the above models.

Contact: Dr. Barnali Dixon – Asst. Professor
University of South Florida – St. Petersburg
140 7th Ave. S. – DAV 210
St. Petersburg, FL 33701

Support Vector Machine (RBF kernel, radius=7)

1. Results using all 14 variables

Training results



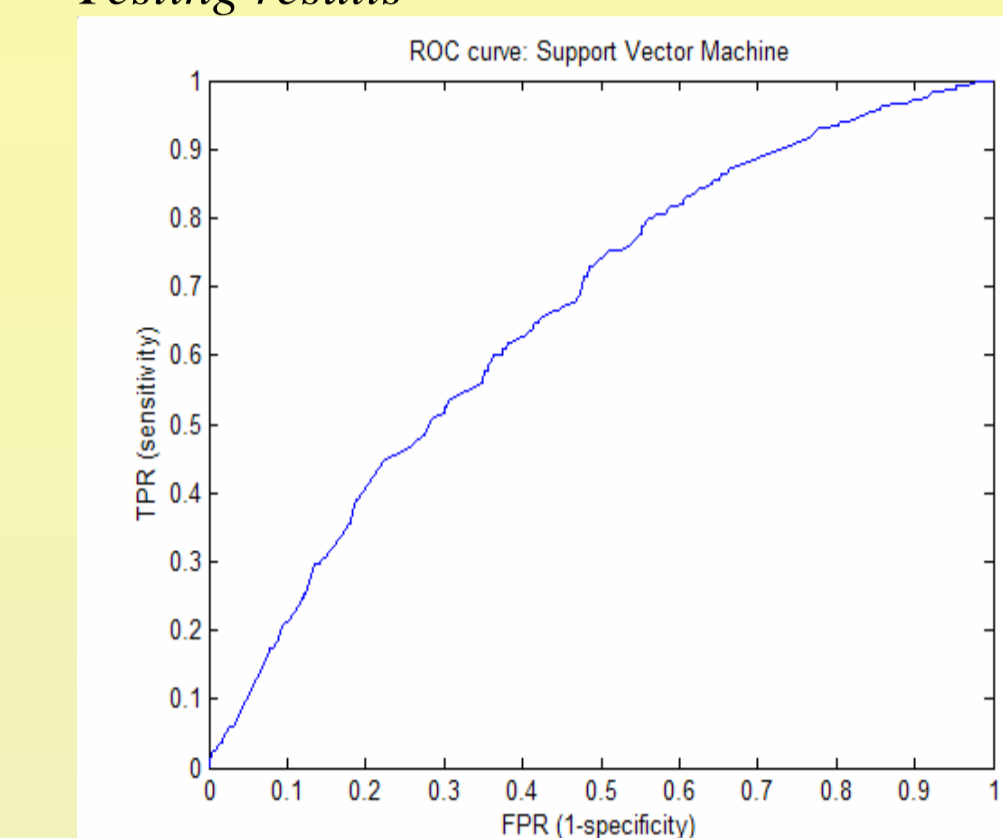
For a threshold of 0.13,

Accuracy = 0.6

Confusion matrix =
0.7168 0.2832
0.4162 0.5838

Area under Curve (AUC) = 0.7214

Testing results



For a threshold of 0.13,

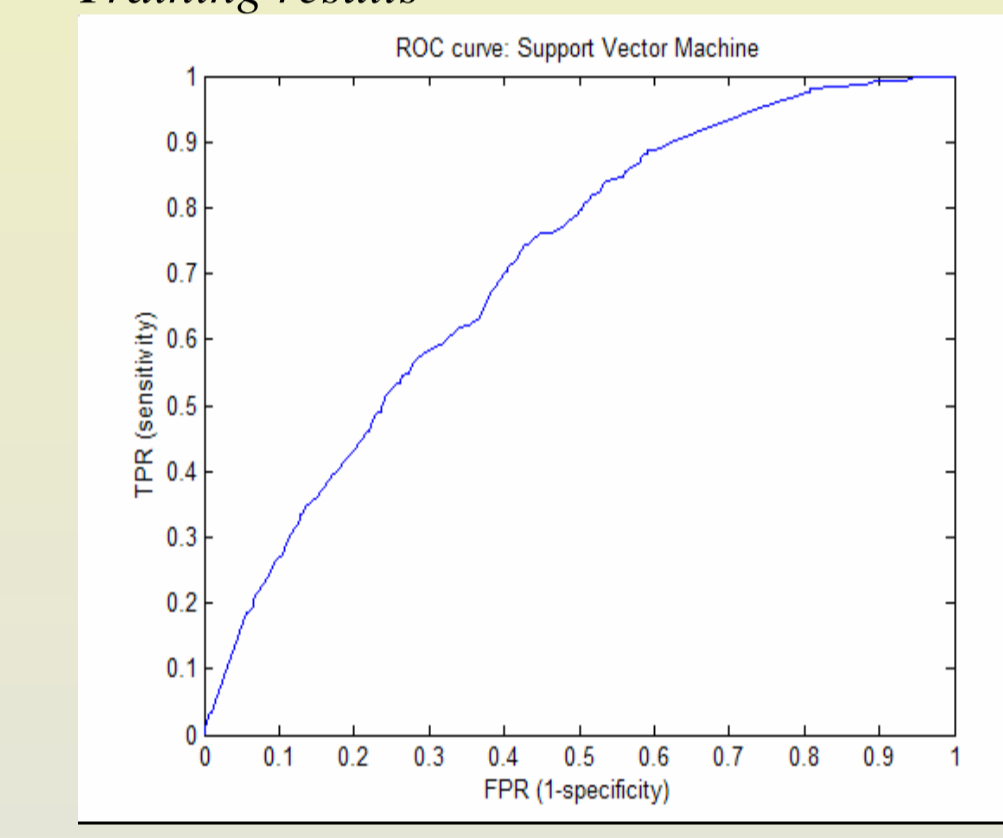
Accuracy = 0.5754

Confusion matrix=
a. 0.3348
a. 0.5610

Area under Curve (AUC) = 0.6626

2. Results using variables from feature selection

Training results



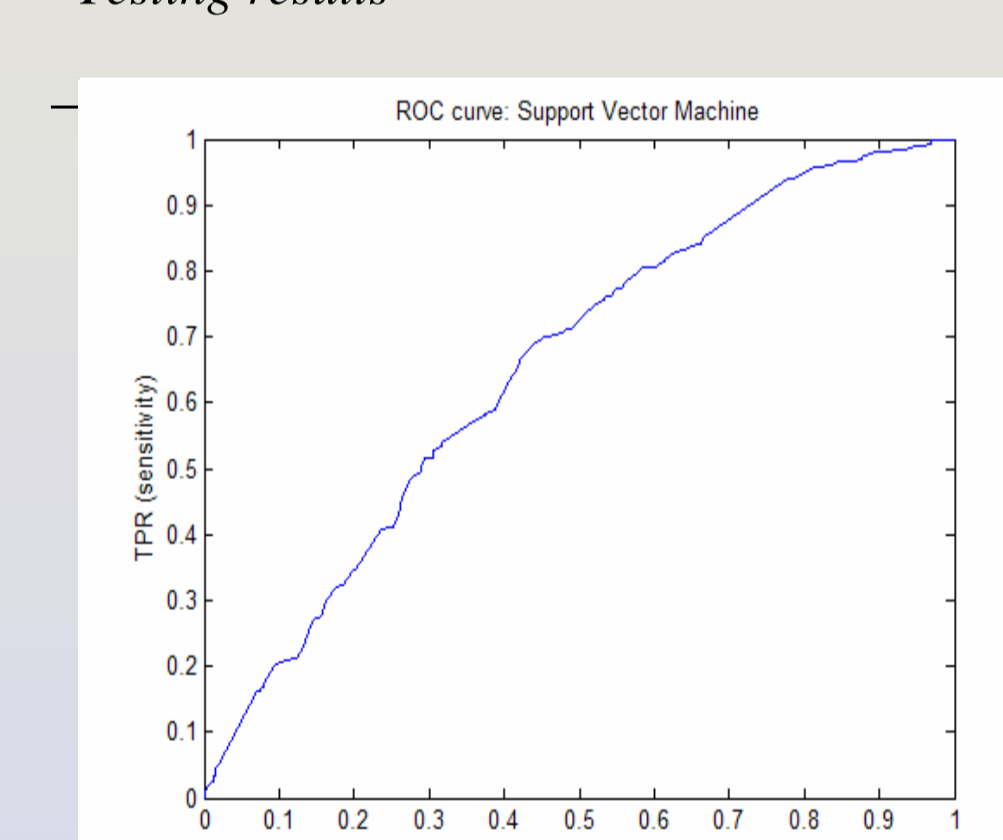
For a threshold of 0.13,

Accuracy = 0.5784

Confusion matrix=
a. 0.2375
a. 0.5490

Area under Curve (AUC) = 0.7097

Testing results



For a threshold of 0.13,

Accuracy = 0.5483

Confusion matrix =
0.7065 0.2935
0.4770 0.5230

Area under Curve (AUC) = 0.6542