PREDICTION OF CROP DAMAGE ON ST. KITTS

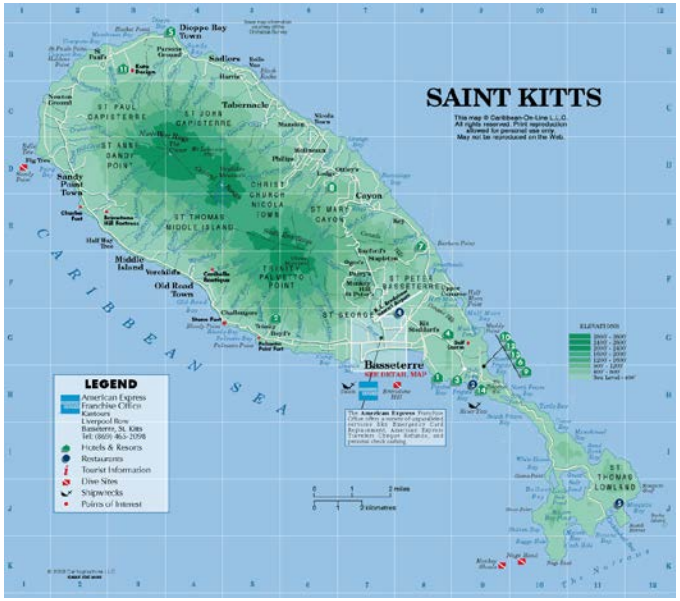
Joey Alamo
Senan Agblonon
Mentor: Dr. Daniel Sewell

Purpose

- To develop a predictive model to evaluate the probability of crop damage occurrence due to vervet monkeys on the island of St. Kitts

Background – Vervet Monkeys on St. Kitts

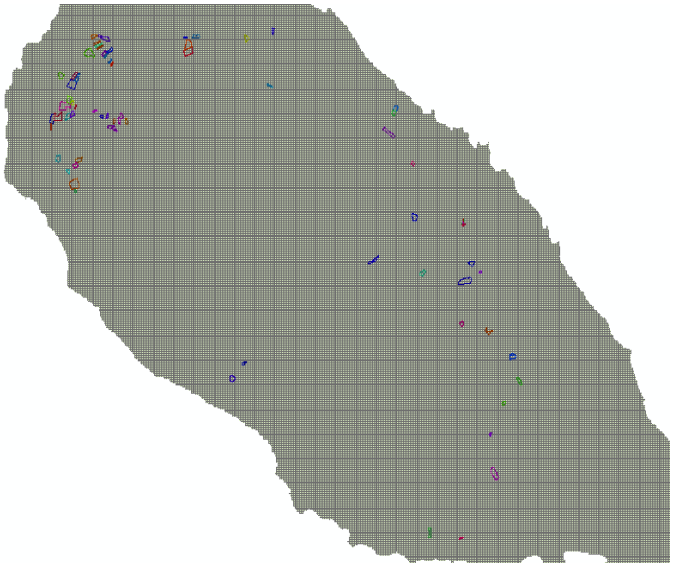
Vervet Monkeys
Source: africadreamsafaris.com



St. Kitts
Source: www.caribbean-on-line.com

- Vervet Monkeys introduced to St. Kitts with the introduction of African slave labor forces
- Crop Damage due to vervet monkeys has existed for over 350 years

Background – Data Collection



Using GPS and GIS technology, a half-acre grid system has been implemented.

- Data collected by anthropologist Kerry Dore
- 65 farms sampled from 9 parishes
- 6115 observations of half-acre cells

Background – Covariates Influencing Crop Damage

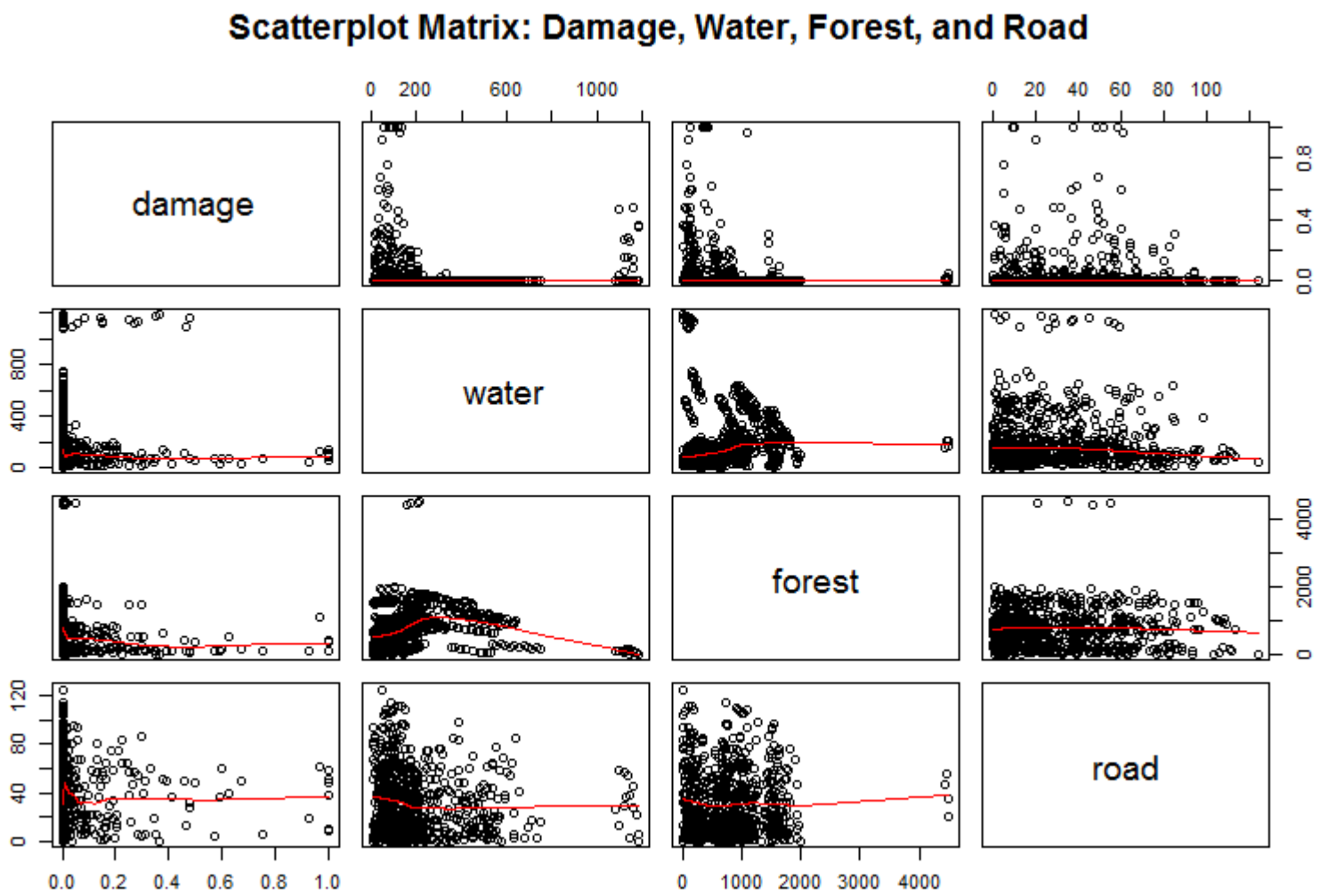
- Independent Variable: Damage (proportion between 0 and 1)
- Continuous Covariates:
  - Distance to Water (m)
  - Distance to Road (m)
  - Distance to Forest (m)
- Discrete Covariates:
  - Season: Mango Season (May-August)
    Non-Mango Season (September-April)
  - Guarding: Extent Farms Guarded For Crop Damage (Scale: 1-8)
  - Preference: Most Preferred Crop for Crop Raiding (Scale: 0-10)
  - Neighbors: Number of Neighboring Farms

Data

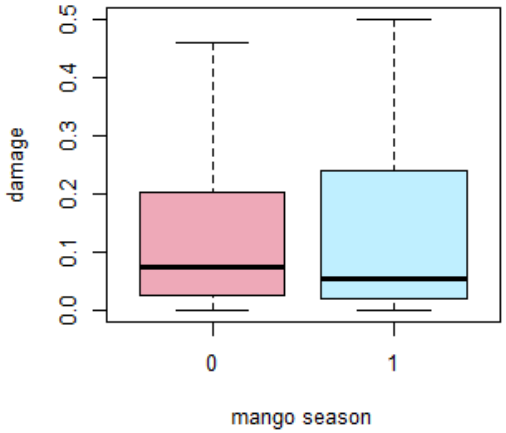
Farm Cells with Damage	174
Farm Cells without Damage	5941
Total	6115

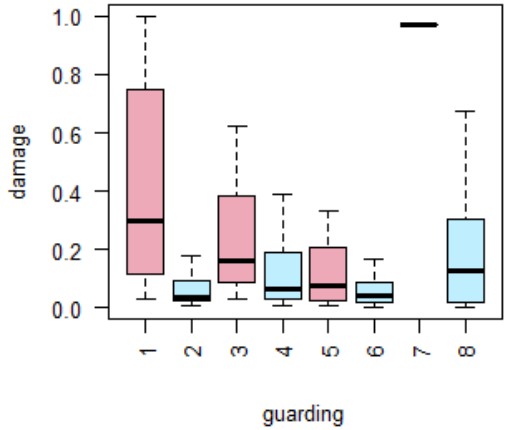
$$Prevalence \approx 0.028$$

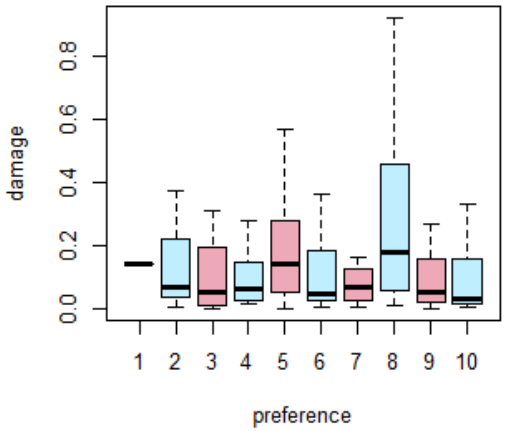
Data

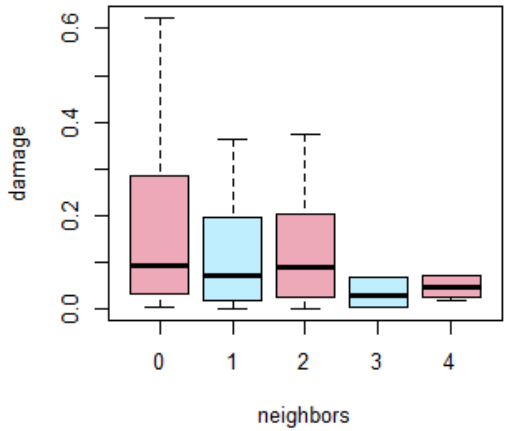

Scatterplot Matrix: Damage, Water, Forest, and Road
damage
water
forest
road
0 200 600 1000
0 20 40 60 80 100
0.0 0.4 0.8
0 400 800
0 2000 4000
0 40 80 120
0.0 0.2 0.4 0.6 0.8 1.0
0 1000 2000 3000 4000

Data


damage
0.0
0.1
0.2
0.3
0.4
0.5
0
1
mango season


damage
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
6
7
8
guarding


damage
0.0
0.2
0.4
0.6
0.8
1
2
3
4
5
6
7
8
9
10
preference


damage
0.0
0.2
0.4
0.6
0
1
2
3
4
neighbors

Methods – Data Splitting

- Data Split into Two Mutually Disjoint Sets:
  - Training Set: 80% of data,
    regression models built from this set
  - Testing Set: remaining 20% of data,
    tests the predictive strength of models built from training set

Generalized Linear Models

- Models utilize Logistic Regression, where:

$$\log\left(\frac{P(damage)}{1-P(damage)}\right) = X\beta$$

- Saturated Model:

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -5.943e-01  6.726e-01  -0.884  0.37686
water             -1.677e-02  3.740e-03  -4.484 7.33e-06 ***
road               7.758e-04  1.051e-02   0.074  0.94115
forest            -2.331e-03  5.454e-04  -4.275 1.91e-05 ***
guarding          -4.696e-01  1.161e-01  -4.046 5.22e-05 ***
pref               2.643e-01  3.338e-02   7.917 2.43e-15 ***
neighbors         -8.182e-01  1.973e-01  -4.147 3.37e-05 ***
mango              9.308e-01  7.769e-01   1.198  0.23088
forest:guarding    2.903e-04  9.897e-05   2.933  0.00336 **
road:forest        8.362e-06  7.815e-06   1.070  0.28463
guarding:mango    -3.705e-01  1.309e-01  -2.829  0.00467 **
water:guarding     2.168e-03  4.795e-04   4.521 6.17e-06 ***
road:guarding      2.820e-04  1.665e-03   0.169  0.86551
pref:mango        -7.450e-02  8.428e-02  -0.884  0.37672
guarding:neighbors 7.101e-02  3.481e-02   2.040  0.04139 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Generalized Linear Models

- Reduced Model:

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -6.118e-01  6.287e-01  -0.973 0.330534
water             -1.742e-02  3.726e-03  -4.675 2.94e-06 ***
road               6.993e-03  3.458e-03   2.022 0.043177 *
forest            -2.141e-03  5.161e-04  -4.148 3.35e-05 ***
guarding          -4.862e-01  1.036e-01  -4.692 2.71e-06 ***
pref               2.531e-01  3.098e-02   8.168 3.14e-16 ***
neighbors         -8.346e-01  1.973e-01  -4.231 2.32e-05 ***
mango              4.813e-01  5.917e-01   0.813 0.415985
forest:guarding    3.328e-04  9.109e-05   3.653 0.000259 ***
guarding:mango    -3.815e-01  1.322e-01  -2.885 0.003911 **
water:guarding     2.258e-03  4.776e-04   4.728 2.26e-06 ***
guarding:neighbors 7.112e-02  3.469e-02   2.051 0.040311 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(All p-values constrained to be below 0.05)

Calculating Estimated Probability of Damages

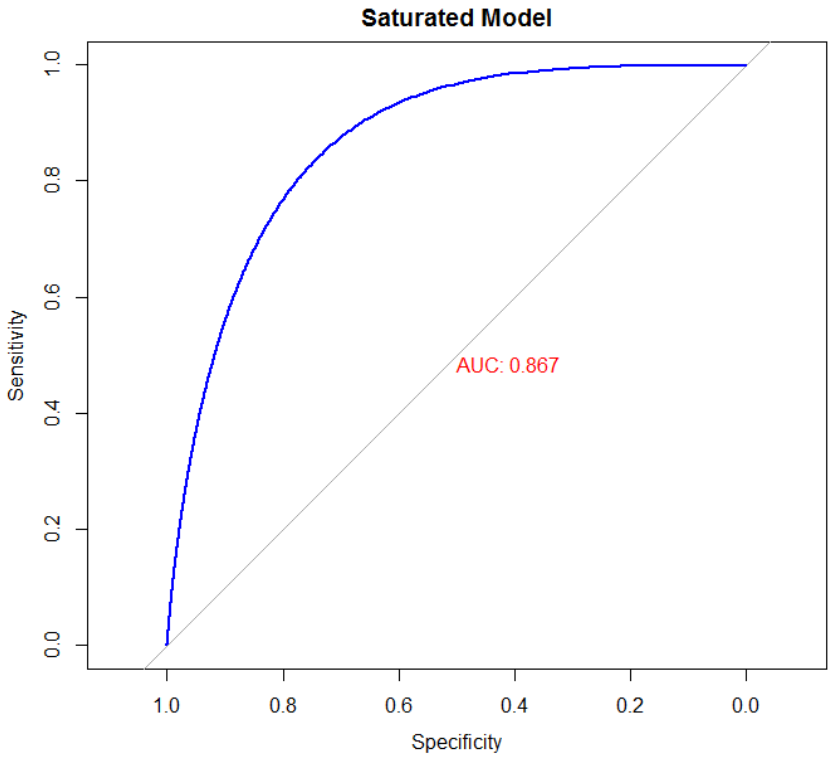
- $\hat{\beta}$ = matrix of coefficients taken from R
- X = matrix of corresponding dependent variables from testing set

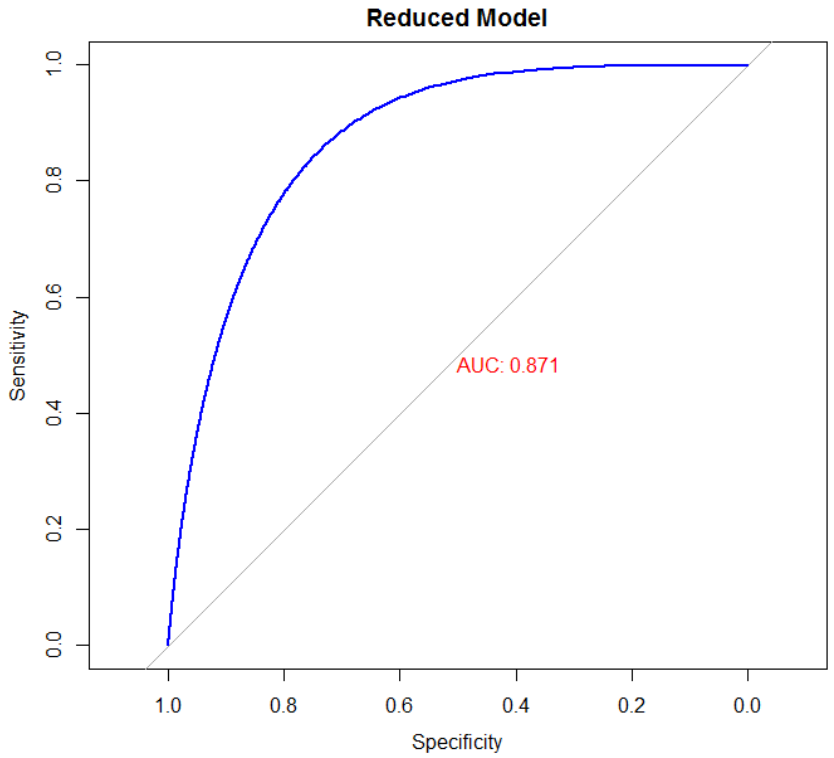
- Inverse Logit Transformation

$$\hat{p} = \frac{e^{X\hat{\beta}}}{1 + e^{X\hat{\beta}}}$$

Where $\hat{p}$ is the estimated probabilities of incurring damage (bounded between 0 and 1)

ROC plots





- Two Methods for Finding Optimal Sensitivity/Specificity:
  - Closest Top Left
  - Youden: Maximizes: $J = Specificity + Sensitivity - 1$

Maximizing Specificity and Sensitivity

	Saturated Model – Youden	Saturated Model – Closest Top Left	Reduced Model – Youden	Reduced Model – Closest Top Left
Threshold	0.03077822	0.03077822	0.02329385	0.0267694
Sensitivity	0.84848485	0.84848485	0.87878788	0.8484848
Specificity	0.79177162	0.79177162	0.75230898	0.7707809

Reduced Model – Youden Method:

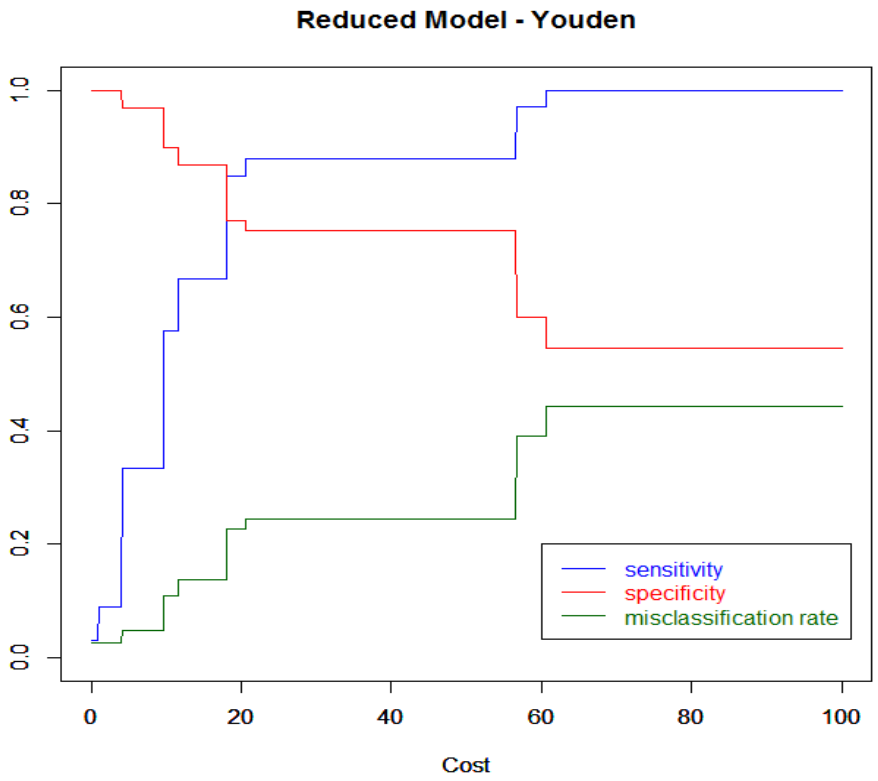
	Damage +	**Damage -**	**Total**
Prediction +	29	295	324
Prediction -	4	896	900
Total	33	1191	1224

Misclassification Rates

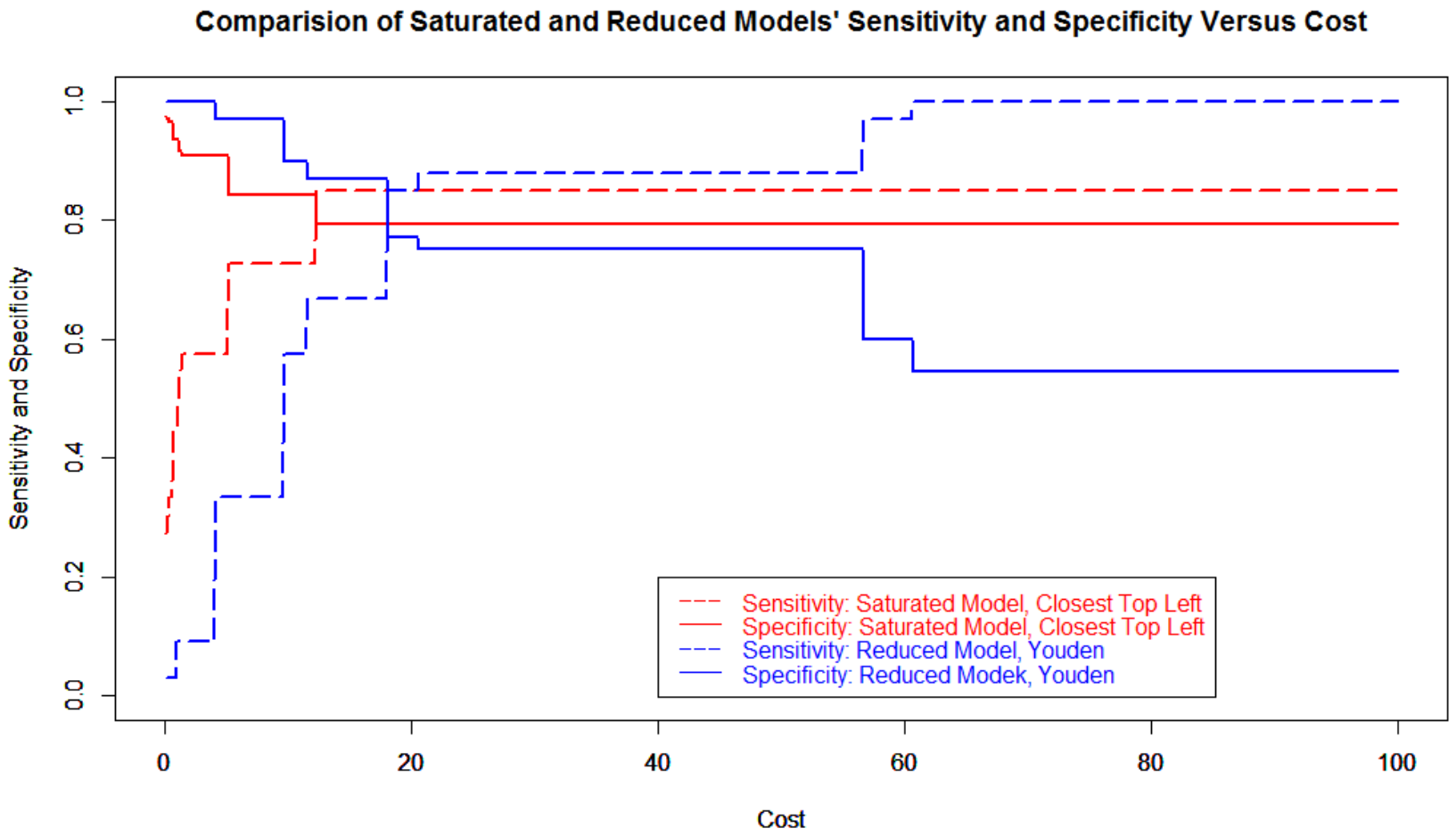
Model	Misclassification Rate
Saturated - Youden	0.2066993
Saturated – Closest Top Left	0.2066993
Reduced – Youden	0.244281
Reduced – Closest Top Left	0.2271242

Cost and Prevalence

- Calculations of Optimal Specificity/Sensitivity Can Account for Cost and Low Prevalence (Prevalence=0.02845)



Cost and Prevalence



Conclusions

- Saturated Model provides the lowest misclassification rate
- Reduced Model with Youden optimization provides the highest sensitivity but has the highest misclassification rate

Conclusions

- Statistically Significant Covariates:
  - Water
  - Road
  - Forest
  - Guarding
  - Preference
  - Neighbors
- Statistically Significant Interaction Terms:
  - Forest and Guarding
  - Mango Season and Guarding
  - Water and Guarding
  - Neighbors and Guarding

Future Work

- Account for Outliers in Water, Road, and Forest Covariates
- Account for Hierarchical Variation between Farms and Individual Cells

Acknowledgements

We would like to thank and acknowledge:

- Mentor: Dr. Daniel Sewell
- Teaching Assistant: Lauren Sager
- ISIB Classroom Teacher: Gideon Zamba

ISIB Program sponsored by the National Heart Lung and Blood Institute (NHLBI) Grant#: HL131467