West Nile Virus Prediction

Yen-Lin Lin

Jun/12/2015

- Background

- Data Exploration
- Data Processing and Preparation
- Model Building
- Result Analysis and Conclusion
- Recommendation

Background

West Nile virus (WNV)

1. Most people infected with WNV will have no symptoms.
2. About **1 in 5** people who are infected will develop a fever with other symptoms.
3. Less than **1%** of infected people develop a serious, sometimes fatal, neurologic illness.

Background

- In 2002, the first human cases of West Nile virus were reported in Chicago.
- By 2004 the **City of Chicago and the Chicago Department of Public Health** (CDPH) had established a comprehensive surveillance and control program that is still in effect today.

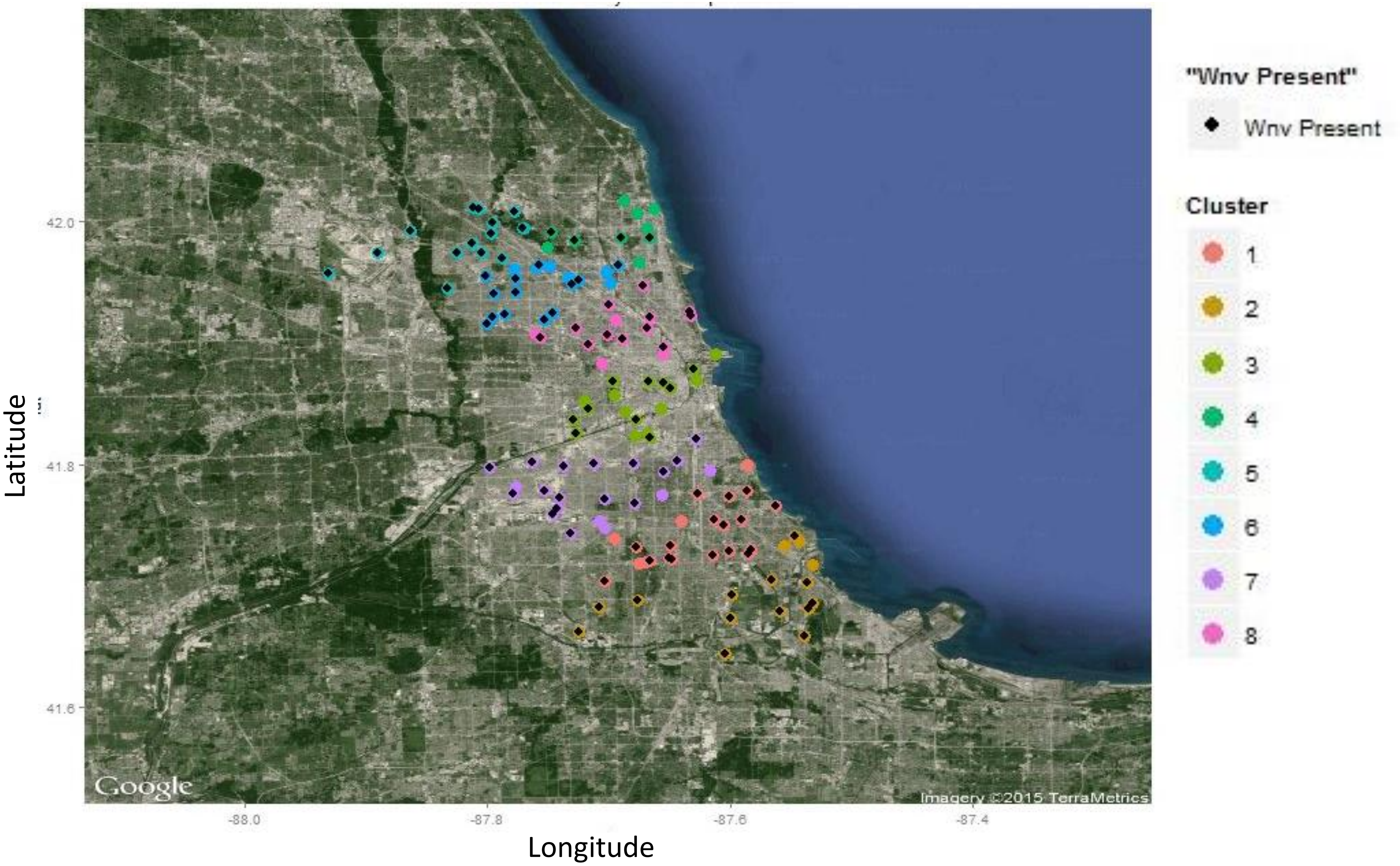
Late Spring→Fall
Mosquitos in traps

Tested for WNV

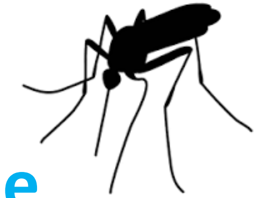
When and Where?
Spay Airborne Pesticides

- Background
- Data Exploration 
- Data Processing and Preparation
- Model Building
- Result Analysis and Conclusion
- Recommendation

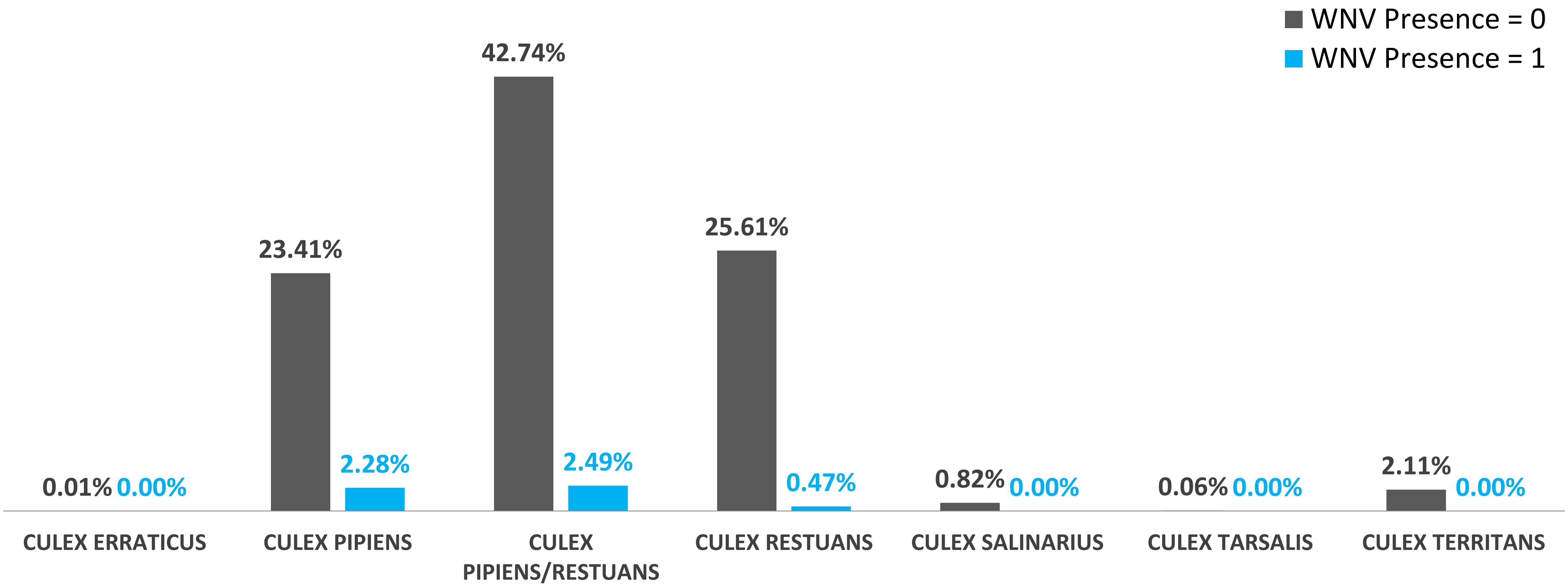
WNV Present Overlayed on Trap Clusters



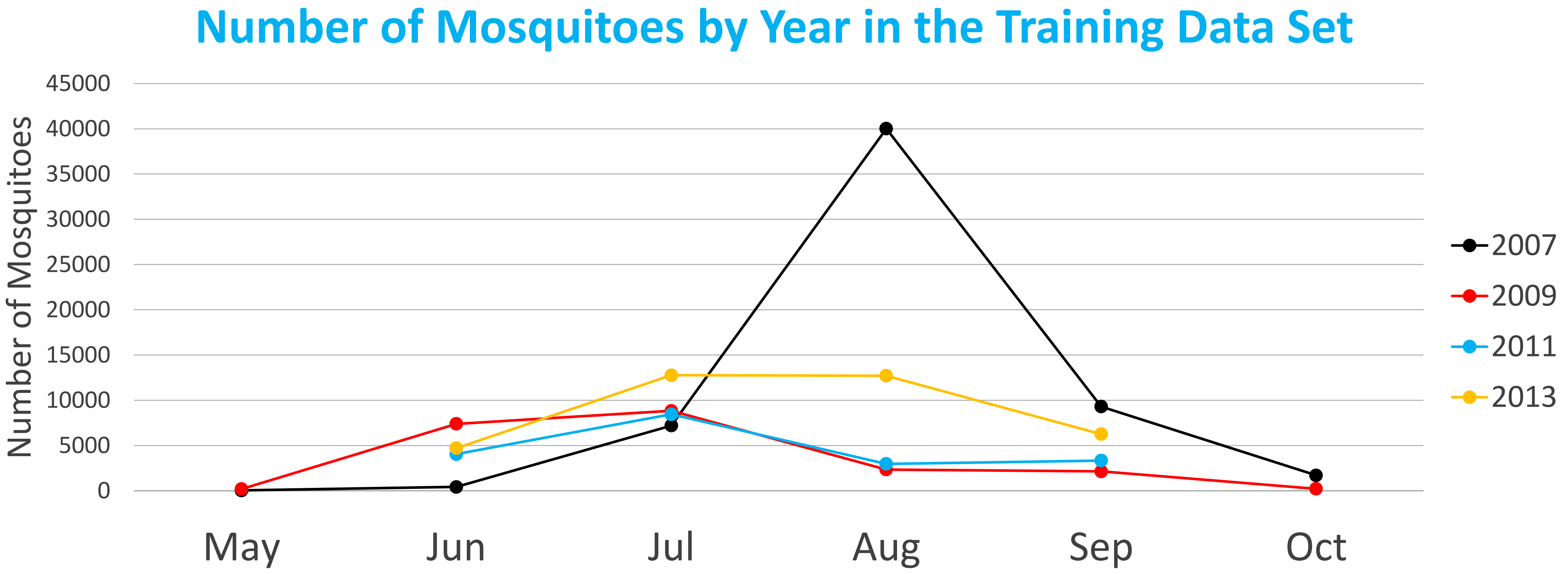
Data Exploration



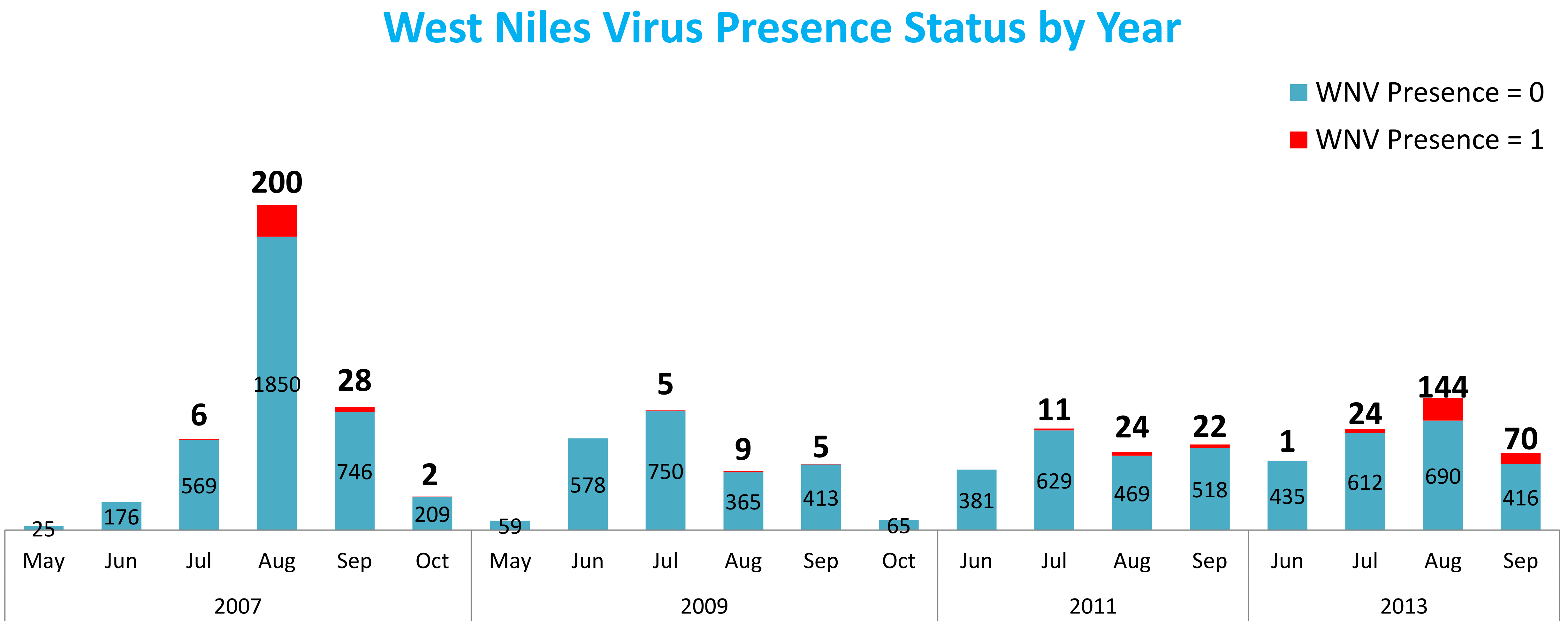
West Niles Virus Presence Status by Mosquito Type



Data Exploration



Data Exploration



- Background
- Data Exploration
- Data Processing and Preparation

- Model Building
- Result Analysis and Conclusion
- Recommendation

Data Processing

Date
Address
Species
Block
Street
Trap
AddressNumberAndStreet
Latitude
Longitude
AddressAccuracy
NumMosquitos
WnvPresent

1. Weather

2. Main Data Set

Station
Date
Tmax
Tmin
Tavg
DewPoint
WetBulb
Heat
Cool
Sunrise
Sunset
CodeSum
Depth
Water1
SnowFall
PrecipTotal
StnPressure
SeaLevel
ResultSpeed
ResultDir
AvgSpeed

Weather

Field Name	Specification and Preparation
Average Temperature (Tavg)	1. Weekly Moving Average (Tavg.ma1—1 wk, Tavg.ma2—2 wks) 2. Transform to ordinal variable based on quantile (Tavg.ordinal)
Precipitation Level (PrecipTotal)	1. Weekly Moving Average (PrecipTotal.ma2—2 wks, PrecipTotal.ma3—3 wks) 2. Heavy rain flag, threshold = 2.165 inch
Wind Speed (AvgSpeed)	1. Low Wind Flag, Threshold = 3.72m/s LowWind.byMean, LowWind.byLow 2. Weekly Moving Average (AvgSpeed.ma3—3 wks)

Weather

Field Name	Specification and Preparation
Dew Point (DewPoint)	1. Weekly Moving Average (DewPoint.ma1—1 wk)
Relative Humidity (RH)	1.100*(exp((17.625*DewPoint[°C])/(243.04+DewPoint[°C]))/exp((17.625*temperature[°C])/(243.04+temperature[°C]))) 2. Weekly Moving Average (relHum.ma4—4 wks)
Day Time Length (daytime)	1. Calculate day time length using Sunrise and Sunset 2. Weekly Moving Average (daytime.ma4—4 wks)

Main Data Set

Field Name	Specification and Preparation
Date	Convert to Month and Year Variable
Species	Eg. CULEX PIPIENS, CULEX PIPIENS/RESTUANS
Longitude, Latitude	1. Define Location 2. Hot Spot (Frequency of Positive test) (HotSpot, log.HotSpot)
Number of Mosquitoes (NumMosquitoes)	Numeric Variable
West Nile Virus Present Target Variable (WnvPresent)	Binary Variable

- Background
- Data Exploration
- Data Processing and Preparation
- Model Building

- Result Analysis and Conclusion
- Recommendation

Model Building

1. Feature Selection
2. Data Partition and Resampling Methods
3. Models

Feature Selection

Feature Selection

Month6
Month7
Month8
Month9
Month10
Species1
Tavg
Tavg.ma1
Tavg.ma2
Tavg.ordinal2
Tavg.ordinal3
Tavg.ordinal4
RH
RH.ma4
DewPoint
DewPoint.ma1
PrecipTotal
PrecipTotal.ma2
PrecipTotal.ma3
HeavyRain1
AvgSpeed
AvgSpeed.ma3
LowWind.byMean1
LowWind.byLow1
daytime
daytime.ma4
HotSpot
log.HotSpot
NumMosquitos

Variable Selection Using **Lasso Regression**:

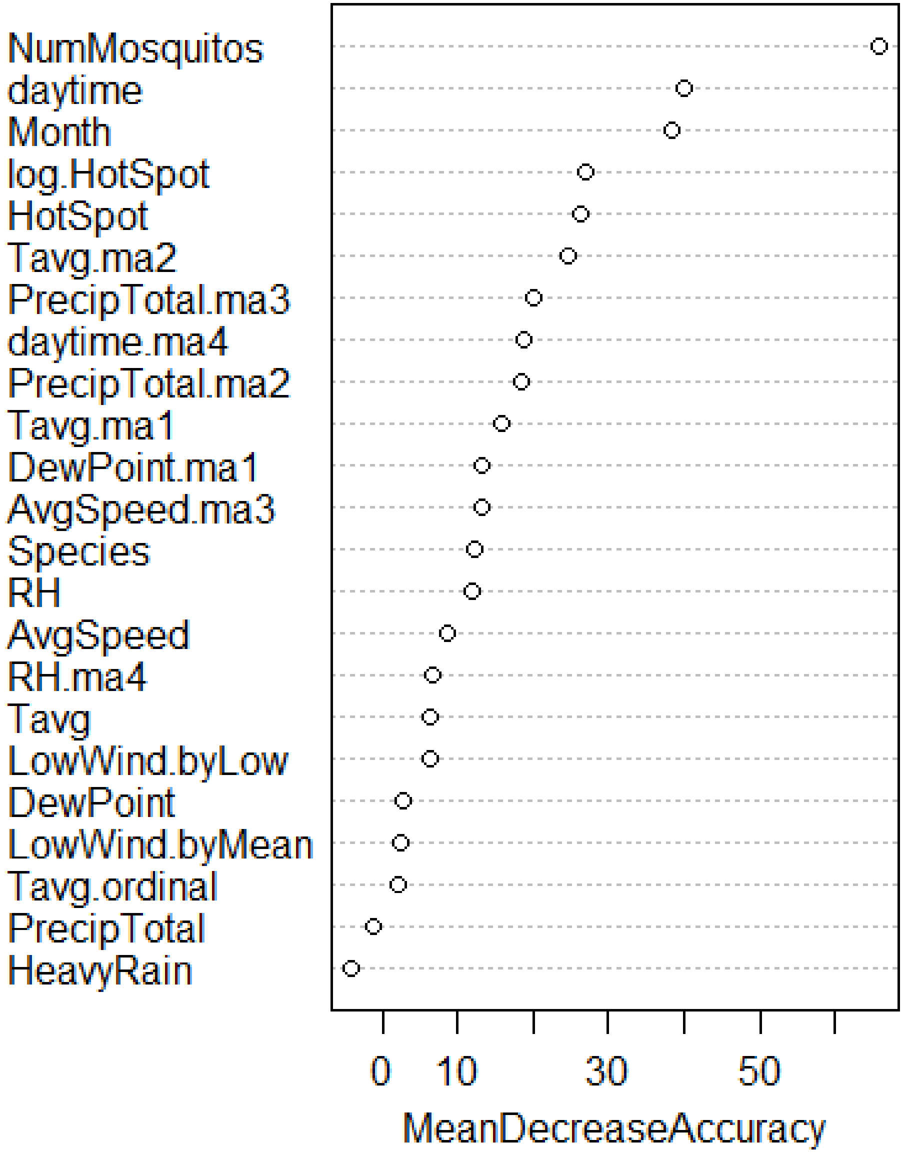
Feature Selection → Lasso →

(Intercept)	-9.09943
Month6	-0.56489
Month7	.
Month8	1.337419
Month9	0.363079
Month10	.
Species1	0.923034
Tavg	0.003948
Tavg.ma1	.
Tavg.ma2	0.066717
Tavg.ordinal2	-0.27215
Tavg.ordinal3	0.107812
Tavg.ordinal4	.
RH	.
RH.ma4	0.051223
DewPoint	.
DewPoint.ma1	0.017863
PrecipTotal	.
PrecipTotal.ma2	.
PrecipTotal.ma3	.
HeavyRain1	.
AvgSpeed	.
AvgSpeed.ma3	.
LowWind.byMean1	-0.8671
LowWind.byLow1	.
daytime	-0.39435
daytime.ma4	.
HotSpot	.
log.HotSpot	1.291344
NumMosquitos	0.006268

Formula will be adopted for the prediction of WnvPresent using **lasso** and **glm** regression:

WnvPresent ~ Month + Species + Tavg.ordinal
+RH.ma4+DewPoint.ma1+LowWind.byMean
+ daytime + log.HotSpot + NumMosquitos

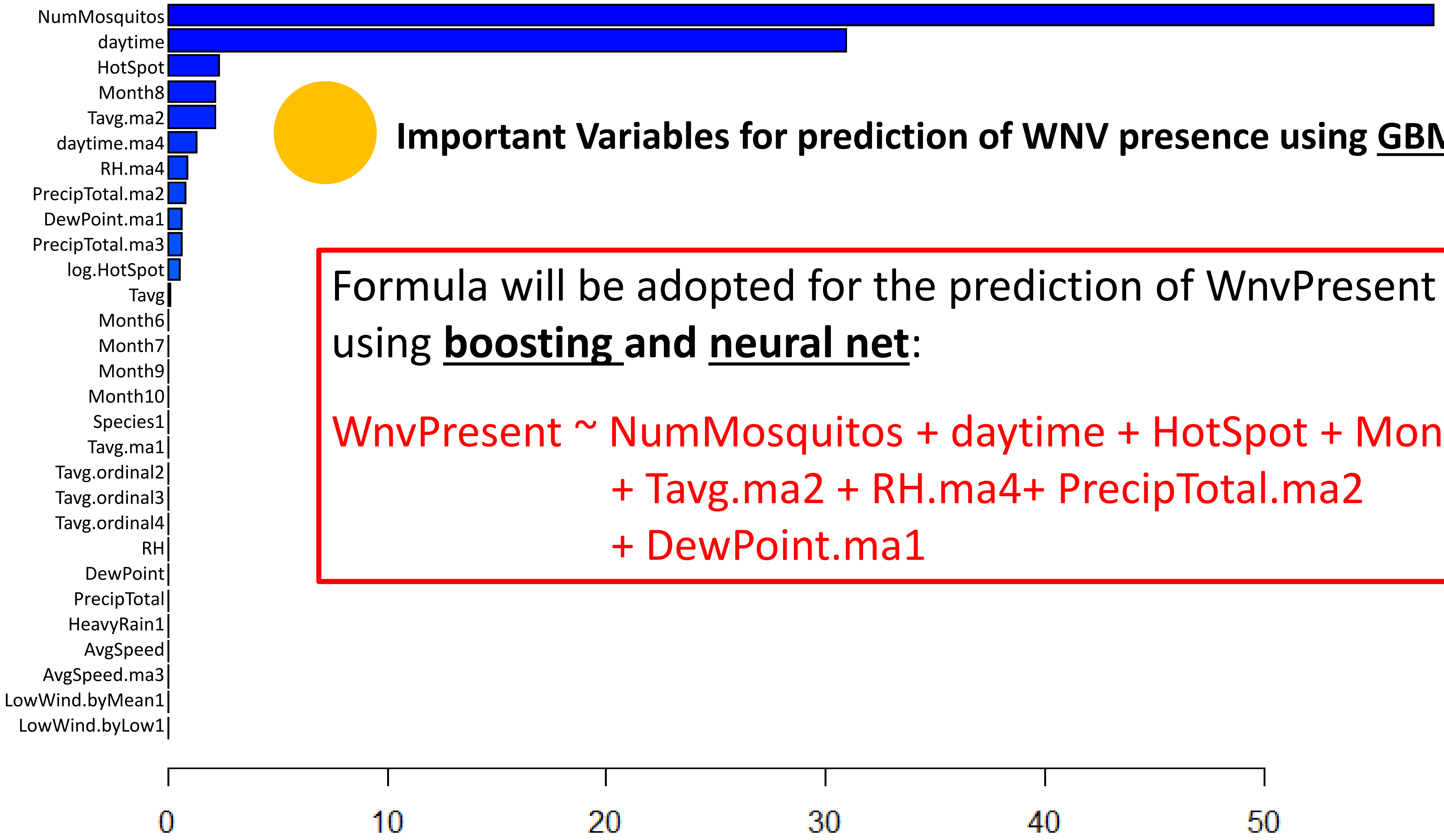
Important Variables for prediction of WNV presence using <u>Random Forest</u>:



Formula will be adopted for the prediction of WnvPresent using **bagging and random forest**:

WnvPresent ~ NumMosquitos + daytime + Month
+ log.HotSpot + Tavg.ma2
+ PrecipTotal.ma3 + DewPoint.ma1
+ AvgSpeed.ma3 + Species + RH

Important Variables for prediction of WNV presence using <u>GBM</u>:



Formula will be adopted for the prediction of WnvPresent using **boosting and neural net**:

WnvPresent ~ NumMosquitos + daytime + HotSpot + Month
+ Tavg.ma2 + RH.ma4+ PrecipTotal.ma2
+ DewPoint.ma1

Data Partition and Resampling Methods

Imbalanced Data: WnvPresent : 95% {NO} 5% {YES}

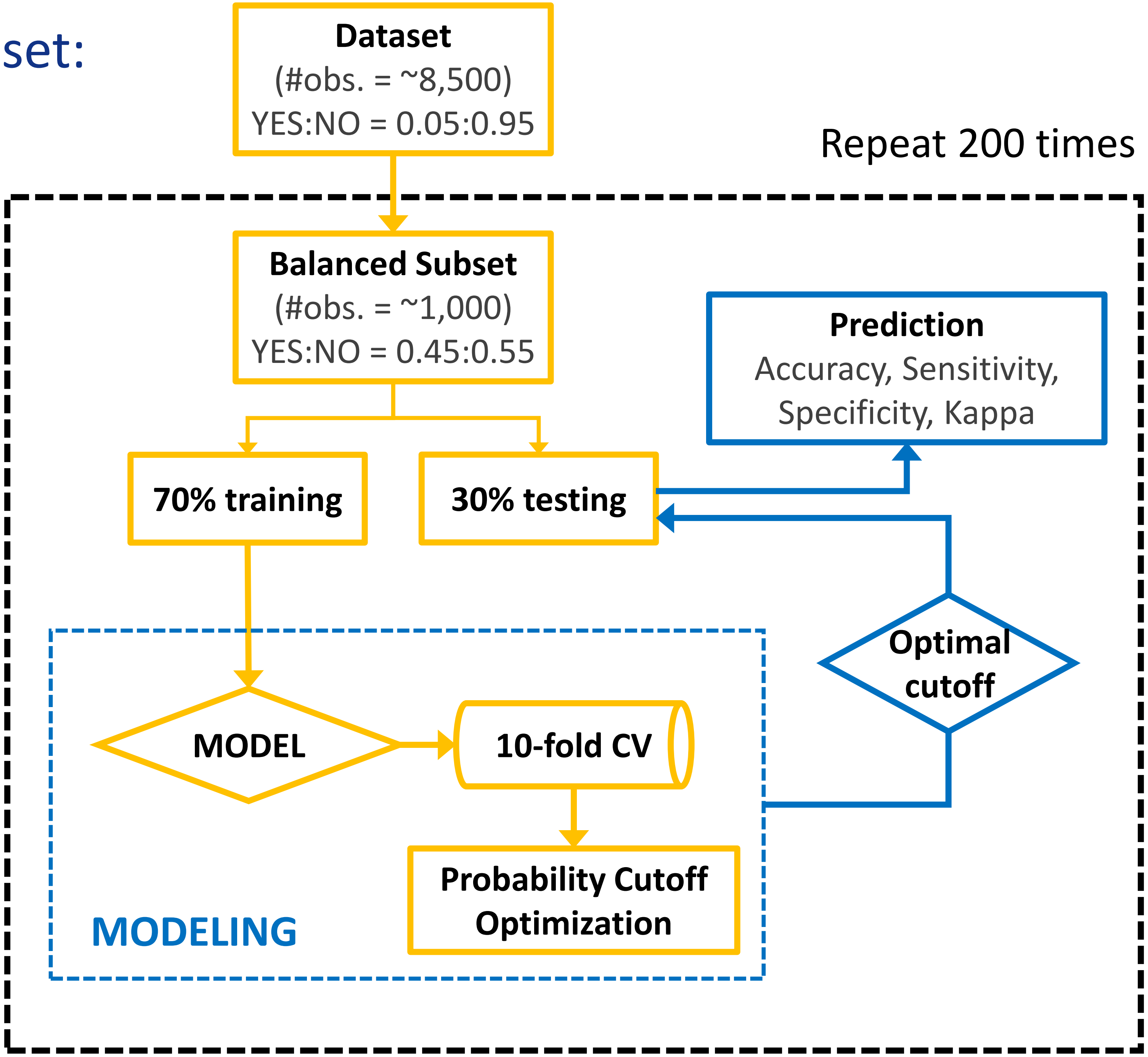
1. Bootstrapping

VS.

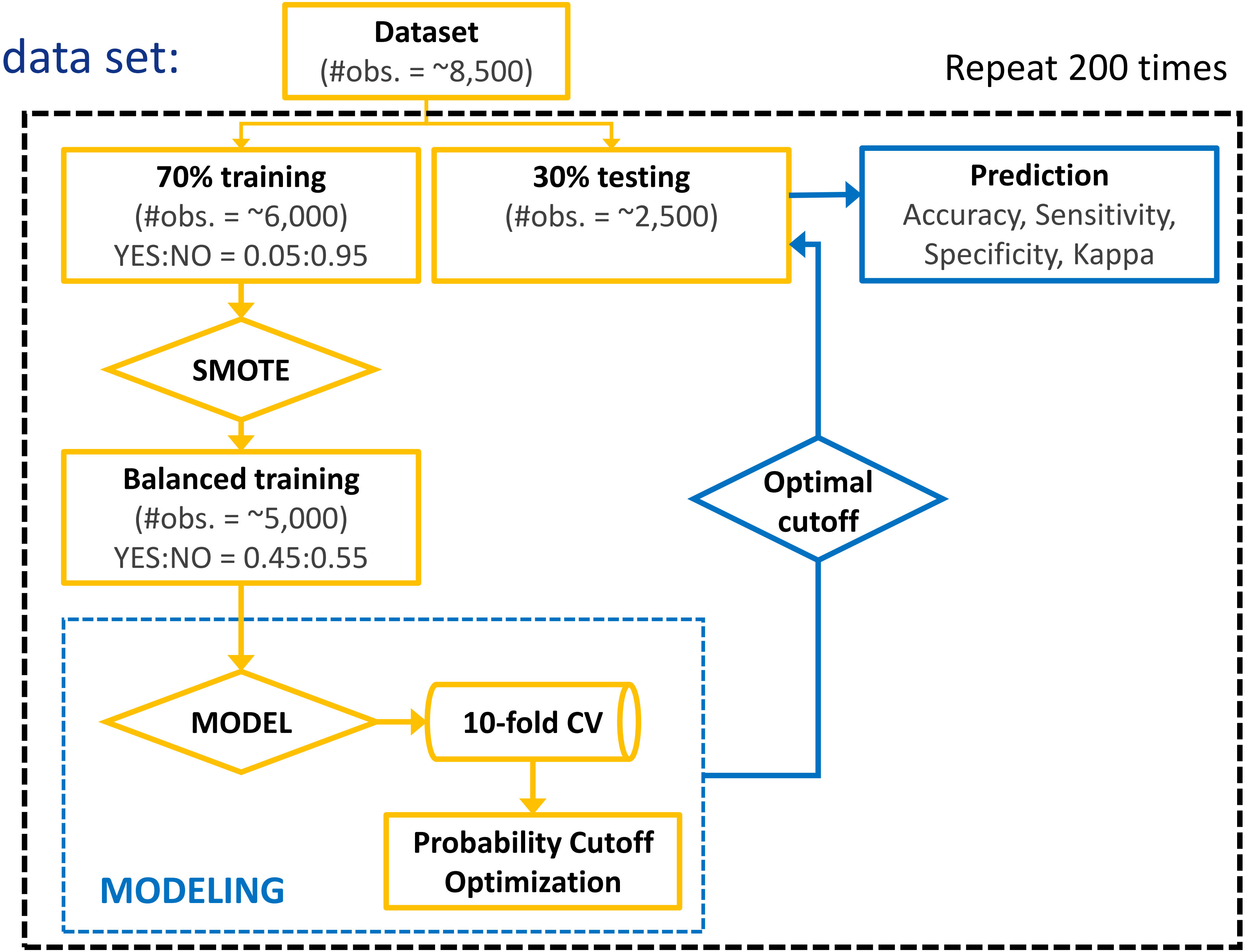
2. SMOTE Package

Synthetic Minority Over-sampling Technique (SMOTE)

Deal with imbalanced data set:
Approach 1: **Bootstrapping**



Deal with imbalanced data set:
2: **SMOTE** package



GLM

Lasso

Naïve Bayesian

Bagging

Random Forest

Boosting

SVM

Neural Network

Models

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537

Models

- GLM
- **Lasso**
- Naïve Bayesian
- Bagging
- Random Forest
- Boosting
- SVM
- Neural Network

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548

Models

GLM

Lasso

Naïve Bayesian

Bagging

Random Forest

Boosting

SVM

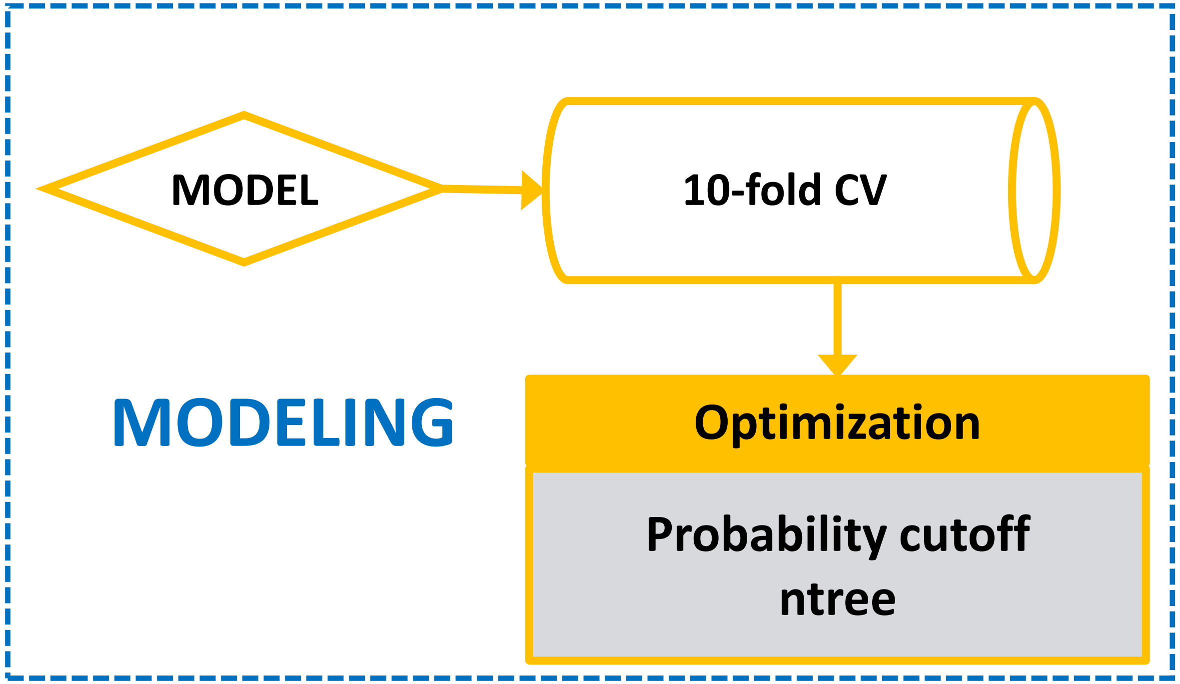
Neural Network

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146

Models

- GLM
- Lasso
- Naïve Bayesian
- **Bagging**
- Random Forest
- Boosting
- SVM
- Neural Network

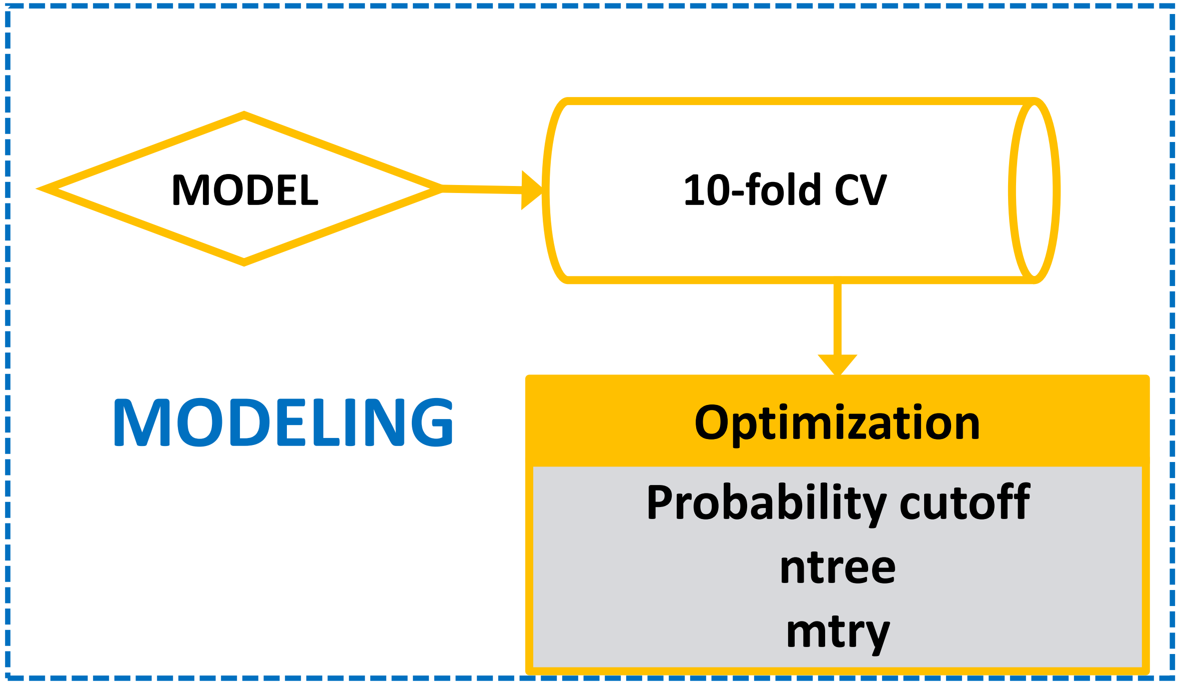
	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146
bagging	0.780	0.810	0.755	0.560



Models

- GLM
- Lasso
- Naïve Bayesian
- Bagging
- Random Forest
- Boosting
- SVM
- Neural Network

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	0.919	0.513



Models

GLM

Lasso

Naïve Bayesian

Bagging

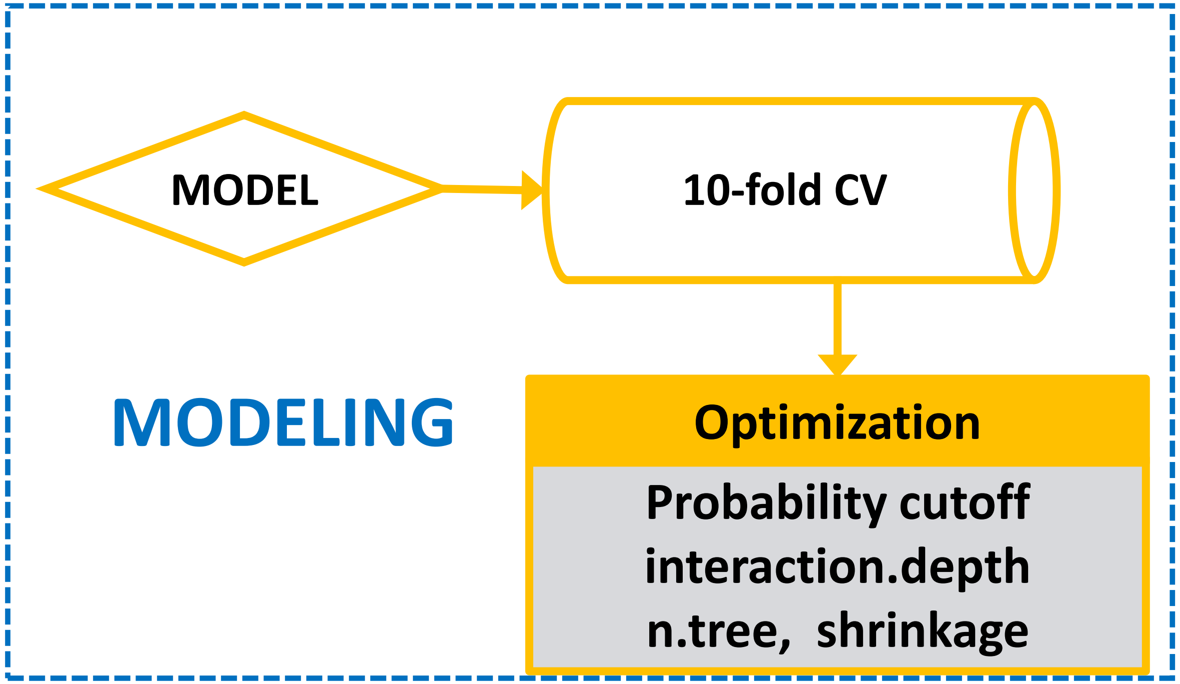
Random Forest

Boosting

SVM

Neural Network

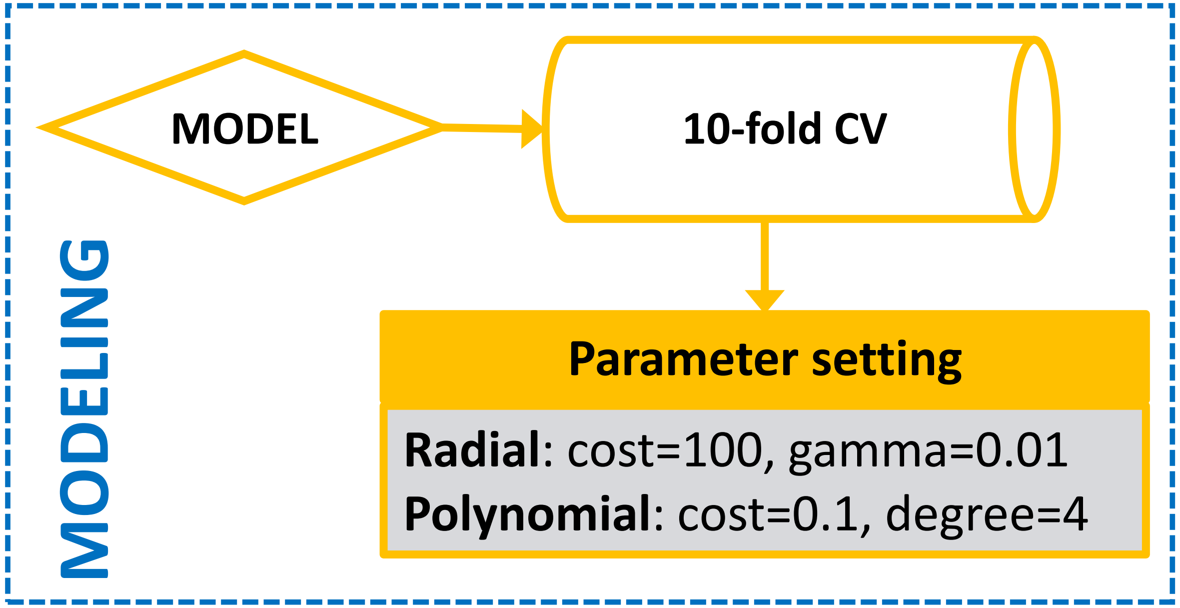
	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	0.919	0.513
boosting	0.897	0.945	0.859	0.796



Models

- GLM
- Lasso
- Naïve Bayesian
- Bagging
- Random Forest
- Boosting
- SVM
- Neural Network

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	0.919	0.513
boosting	0.897	0.945	0.859	0.796
SVM - radial	0.763	0.811	0.724	0.528
SVM - polynomial	0.716	0.815	0.635	0.438



GLM

Lasso

Naïve Bayesian

Bagging

Random Forest

Boosting

SVM

Neural Network

Models

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	0.919	0.513
boosting	0.897	0.945	0.859	0.796
SVM - radial	0.763	0.811	0.724	0.528
SVM - polynomial	0.716	0.815	0.635	0.438
neural network	0.756	0.760	0.753	0.509

Used 2 hidden layers: nodes = c (29,14)

The sufficient number of hidden nodes in the first layer: $\sqrt{(m+2)N} + 2\sqrt{N/(m+2)}$

For second layer: $m\sqrt{N/(m+2)}$

where N = # obs.; m = # predictors

Reference: How many hidden layers and nodes? By D. STATHAKIS. International Journal of Remote Sensing , Vol. 30, No. 8, 20 April 2009, 2133–2147

Models

GLM

Lasso

Naïve Bayesian

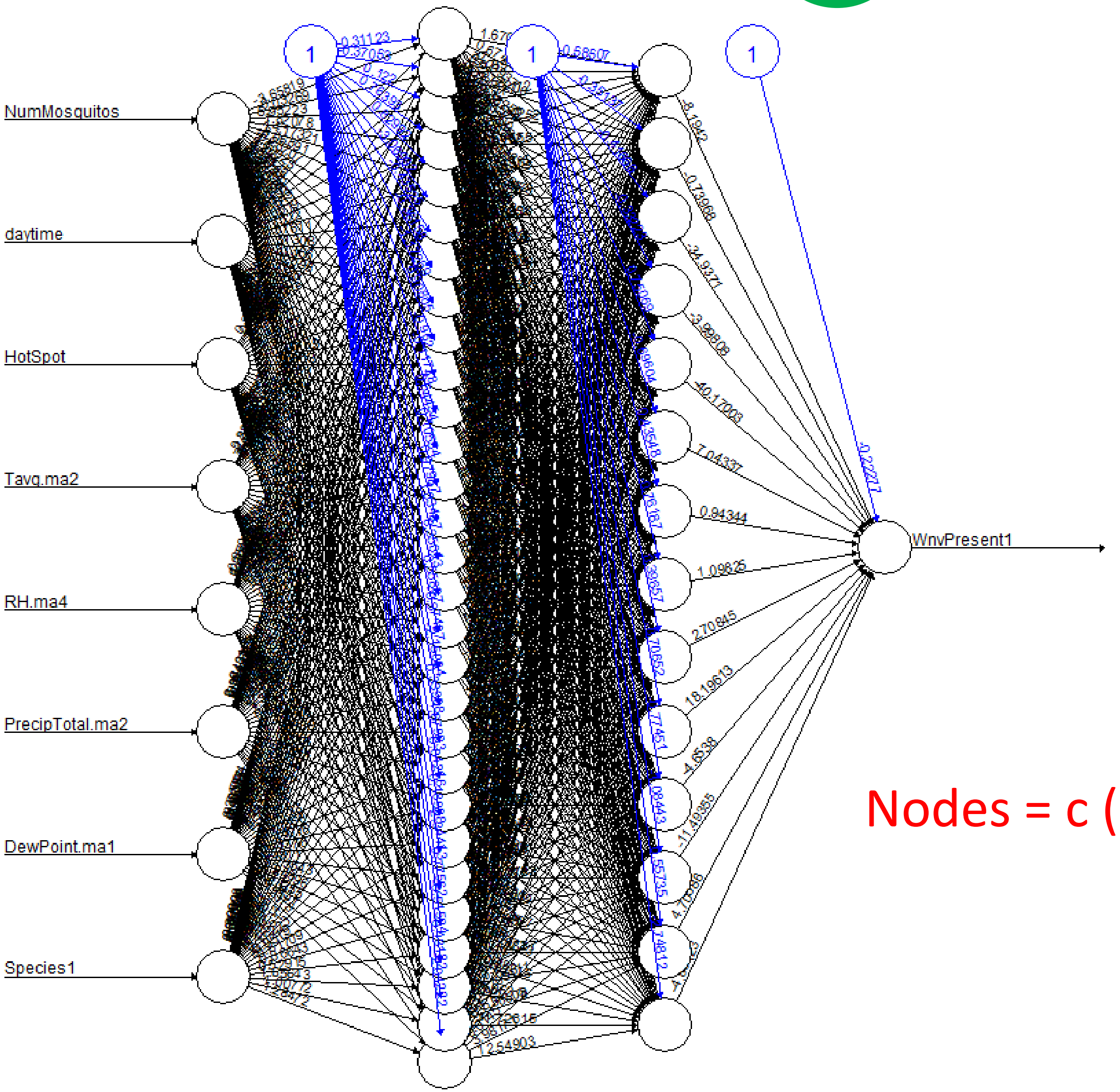
Bagging

Random Forest

Boosting

SVM

Neural Network



Nodes = c (29,14)

Models

- GLM
- Lasso
- Naïve Bayesian
- Bagging
- Random Forest
- Boosting
- SVM
- Neural Network

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	0.919	0.513
boosting	0.897	0.945	0.859	0.796
SVM - radial	0.763	0.811	0.724	0.528
SVM - polynomial	0.716	0.815	0.635	0.438
neural network	0.756	0.760	0.753	0.509

Models

- GLM
- Lasso
- Naïve Bayesian
- Bagging
- Random Forest
- Boosting
- SVM
- Neural Network

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
Naïve Bayesian	0.380	0.815	0.026	-0.146
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	0.919	0.513
boosting	0.897	0.945	0.859	0.796
SVM - radial	0.763	0.811	0.724	0.528
SVM - polynomial	0.716	0.815	0.635	0.438
neural network	0.756	0.760	0.753	0.509

Comparison of Bootstrapping and SMOTE

Approach 1: Bootstrapping

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	0.919	0.513
boosting	0.897	0.945	0.859	0.796
SVM - radial	0.763	0.811	0.724	0.528
SVM - polynomial	0.716	0.815	0.635	0.438
neural network	0.756	0.760	0.753	0.509
Naïve Bayesian	0.380	0.815	0.026	-0.146

Approach 2: SMOTE package

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.842	0.600	0.856	0.234
lasso	0.839	0.593	0.853	0.228
bagging	0.893	0.503	0.916	0.287
random forest	0.937	0.253	0.976	0.267
boosting	0.969	0.454	0.999	0.570
SVM - radial	0.467	0.928	0.441	0.069
SVM - polynomial	0.633	0.820	0.622	0.114

- **Bootstrapping** has higher sensitivity
- **Bootstrapping** has higher kappa

Conclusion

Approach 1: Bootstrapping

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.768	0.804	0.738	0.537
lasso	0.773	0.824	0.731	0.548
bagging	0.780	0.810	0.755	0.560
random forest	0.767	0.579	**0.919**	0.513
boosting	**0.897**	**0.945**	0.859	**0.796**
SVM - radial	0.763	0.811	0.724	0.528
SVM - polynomial	0.716	0.815	0.635	0.438
neural network	0.756	0.760	0.753	0.509
Naïve Bayesian	0.380	0.815	0.026	-0.146

Approach 2: SMOTE package

	Accuracy	Sensitivity	Specificity	Kappa
GLM	0.842	0.600	0.856	0.234
lasso	0.839	0.593	0.853	0.228
bagging	0.893	0.503	0.916	0.287
random forest	0.937	0.253	0.976	0.267
boosting	**0.969**	0.454	**0.999**	**0.570**
SVM - radial	0.467	**0.928**	0.441	0.069
SVM - polynomial	0.633	0.820	0.622	0.114

1. When compared with SMOTE, bootstrapping produces more robust results in terms of sensitivity and kappa.
2. Boosting performs the best among all of the above methods in terms of accuracy and kappa.
3. Overall, bootstrapping on boosting performs the best overall!