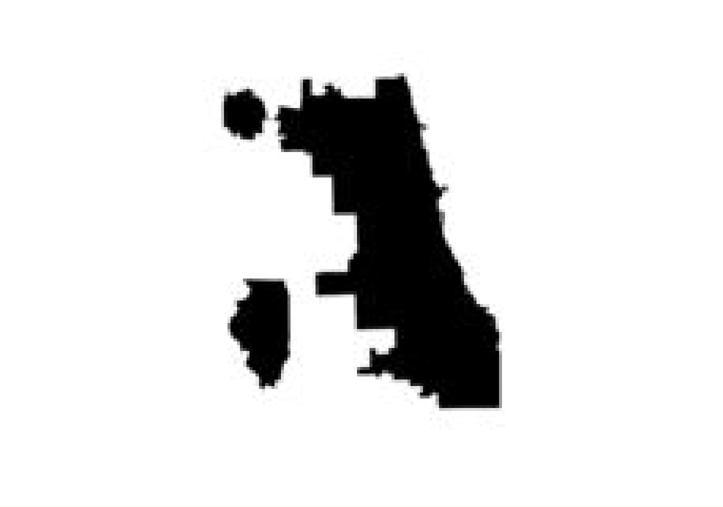
WEST NILE VIRUS

Pesticides - To spray or not to spray?

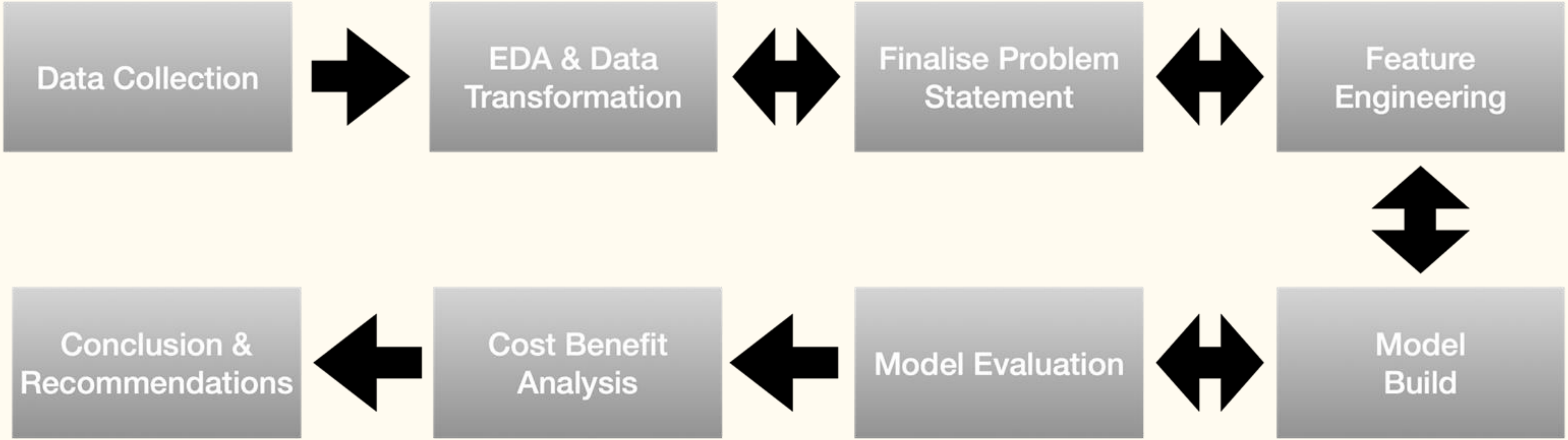
DSI 16 Project 4 : Dominic Ong / Vikaskalia / Peter Wong / Jeriel Wong / Cheyanne Wong

Problem Statement

- Make predictions where West Nile Virus is present in the city of Chicago
- Predictions will be used to decide where to spray
- Conduct cost-benefit analysis



Data Science WorkFlow



Data Collection

EDA & Data Transformation

Finalise Problem Statement

Feature Engineering

Model Build

Model Evaluation

Cost Benefit Analysis

Conclusion & Recommendations

Data Description

	Period									
Dataset	**2007**	**2008**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**Rows**	**Columns**
Train	✓		✓		✓		✓		10506	12
Test		✓		✓		✓		✓	116293	11
Weather	✓	✓	✓	✓	✓	✓	✓	✓	14835	4
Spray					✓	✓			2944	22

Data Cleaning & Transformation

Data Cleaning

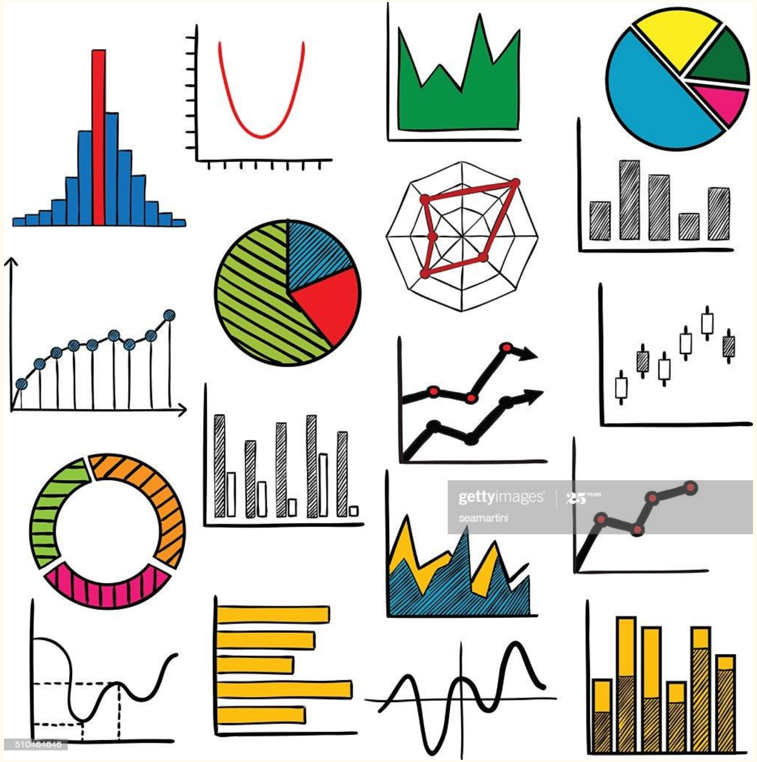
Merging Rows > 50 Mosquitos

Merge Train & Weather dataset

Data Imputation

Weather Station 1 & 2 Ffill

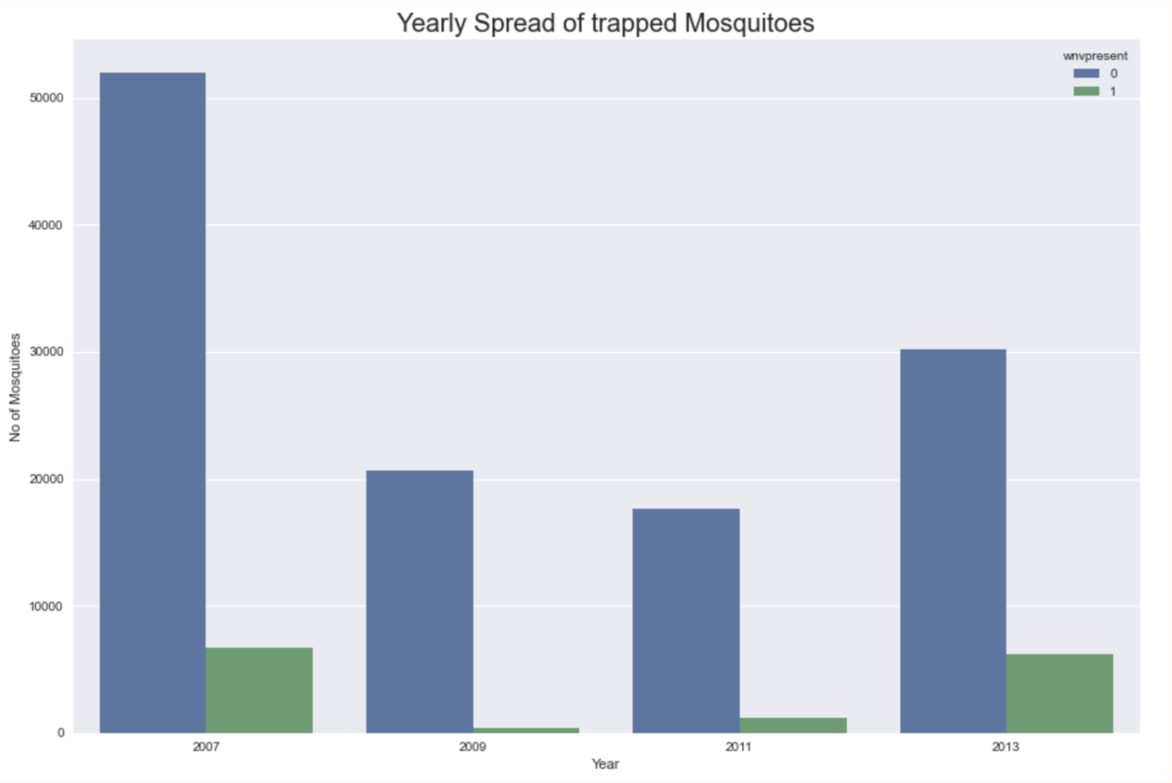
EDA



EDA

Spread of Trapped Mosquitoes By Year

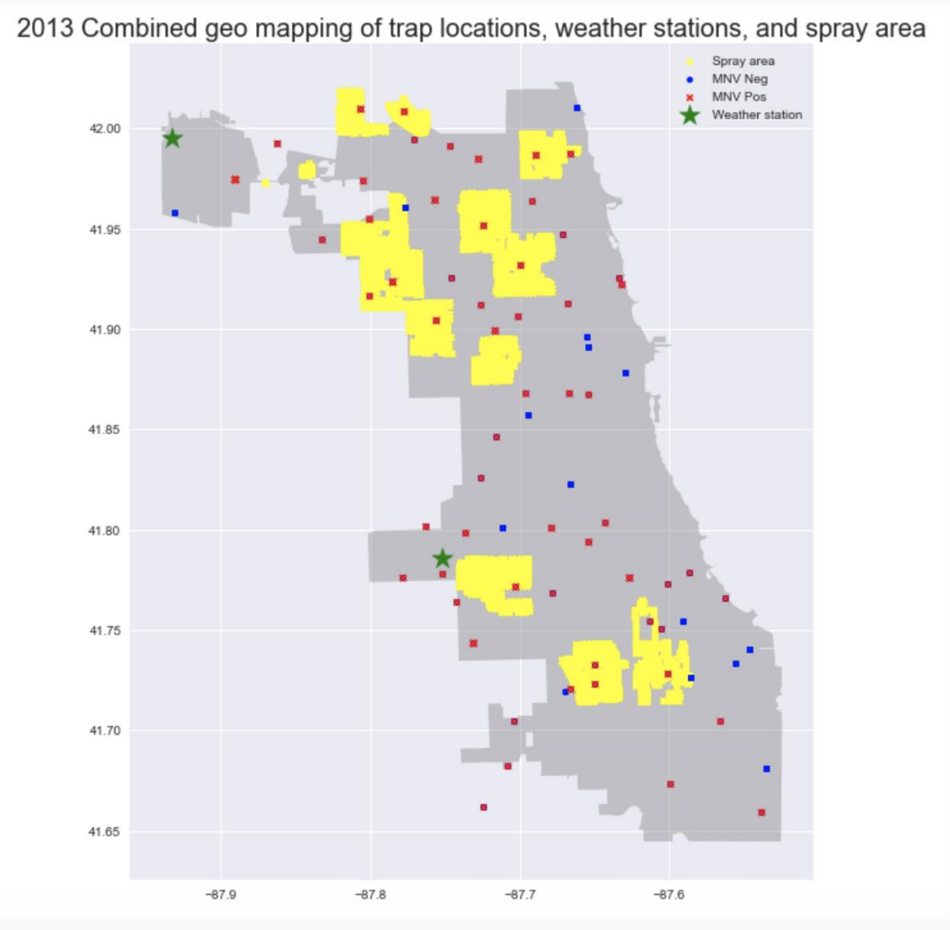
Even though the total number of mosquitoes caught in 2013 was lower than that of 2013, the percentage of WNV presence went up in 2013.



EDA

2013 Trap locations and Spray Area

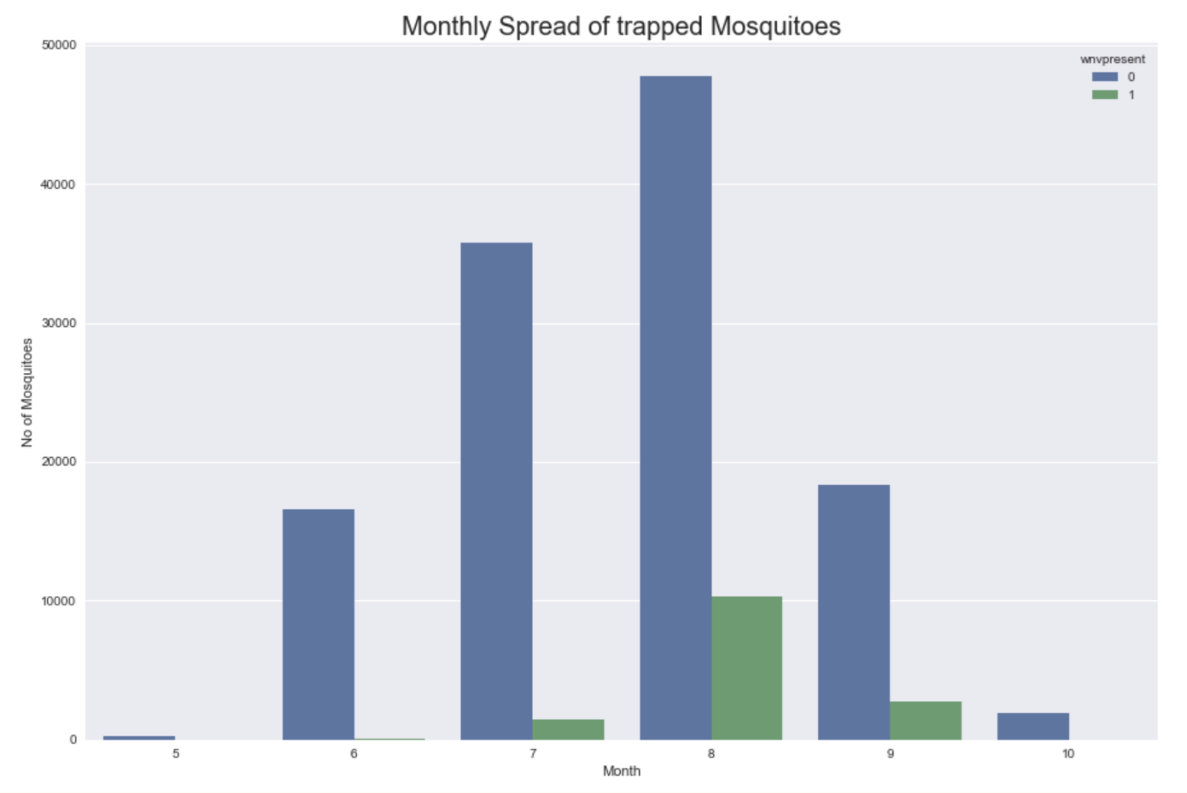
In 2013, WNV presence was found in most traps across the city. The area near Station 1 in the northern region seems to be a hotspot for WNV presence. The spraying of pesticide is concentrated in this region.



EDA

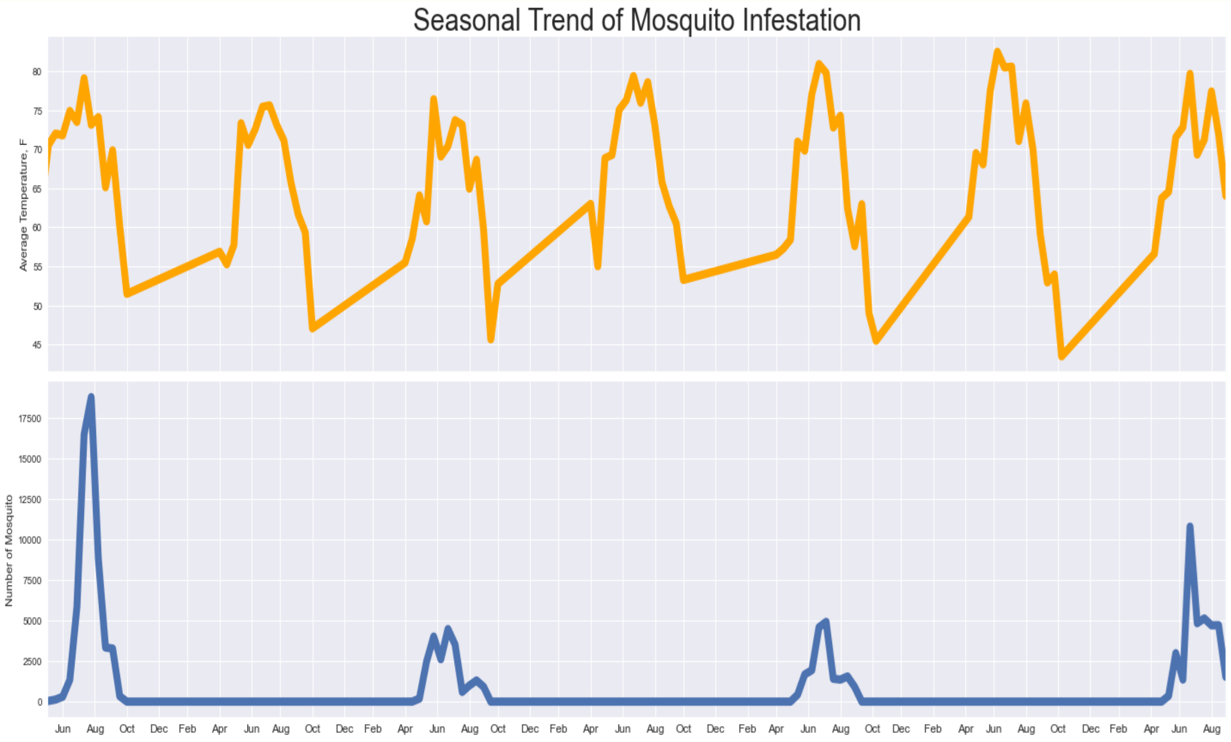
Spread of Trapped Mosquitoes By Month

Number of mosquitoes trapped was the highest in the month of August where the weather is hot and humid. The presence of WNV was also higher in this month.



EDA

Seasonal Trend of Mosquitoes Infestation With Ave Temperature

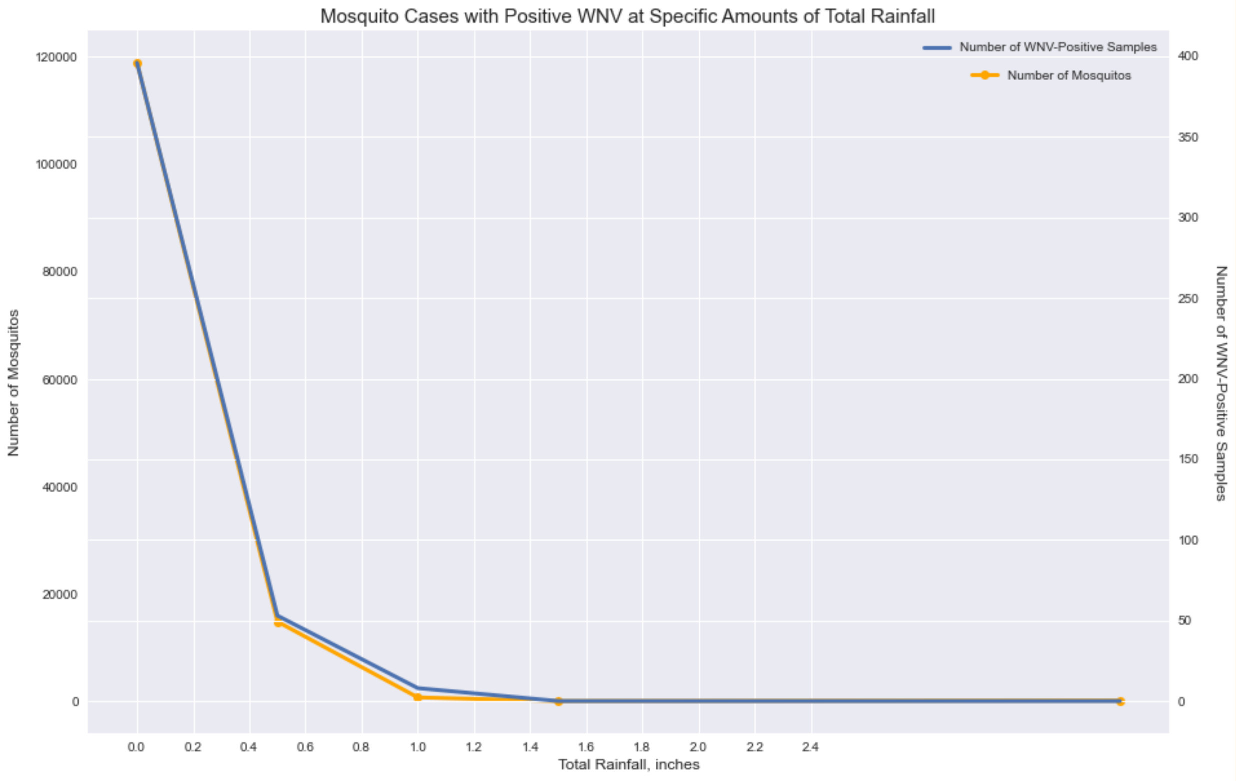


The above graph shows that mosquitoes prefer the higher temperatures as when temperature increase so does the number of mosquitoes.

EDA

No. of Mosquitoes Cases with Total Rainfall

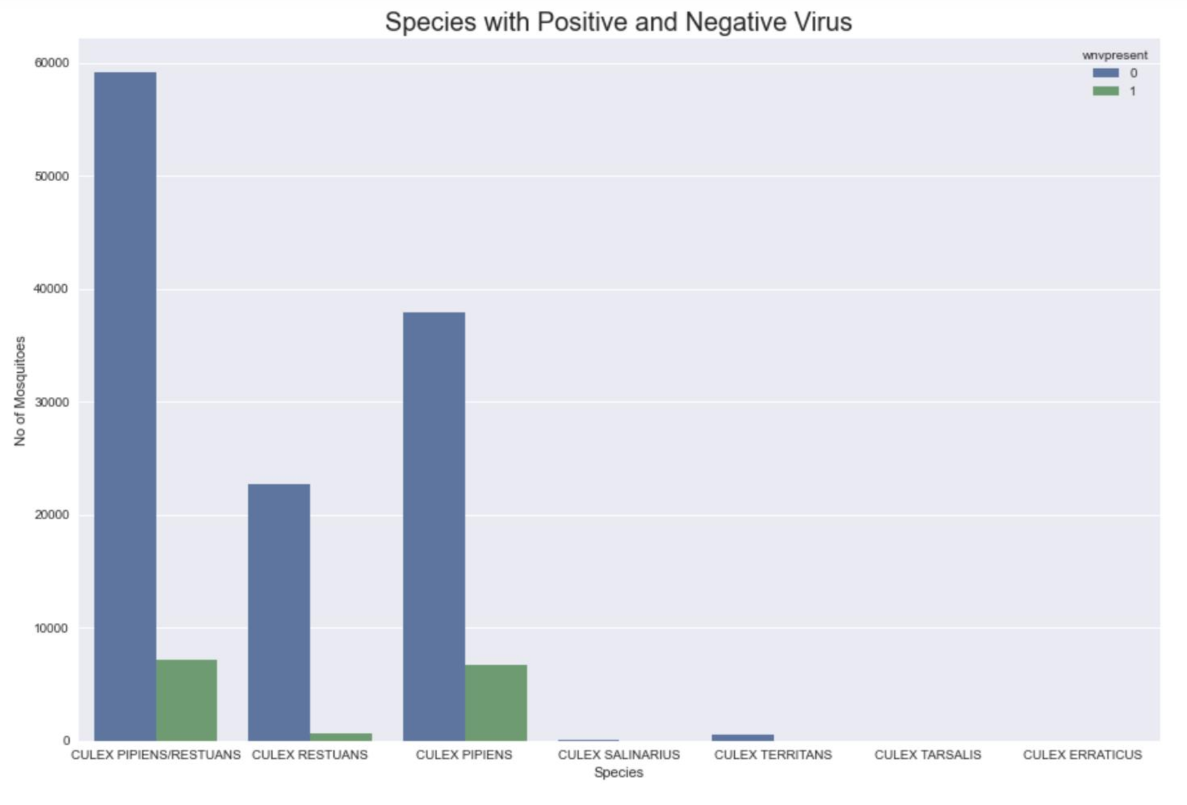
Total rainfall (precipitation) is inversely proportional to both the number of mosquitos and number of WNV-positive traps.



EDA

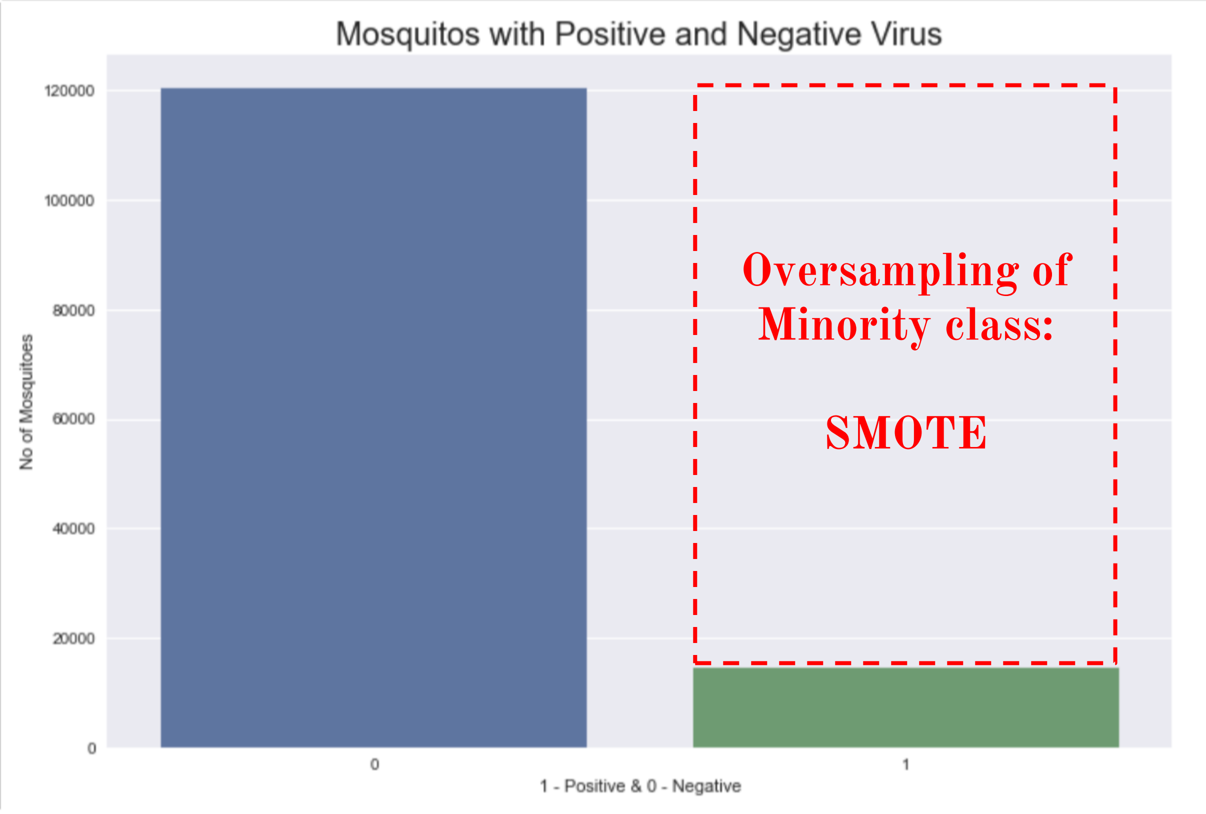
Mosquito Species with Positive & Negative Virus

The types of mosquitoes carrying the WNV virus are Culex Restuans and Culex Pipiens. Traps with presence of these mosquitoes have a higher probability of testing positive for the virus as compared to other types of mosquitoes.



EDA

Imbalanced Class



Oversampling of Minority class:

SMOTE

Feature Engineering

One-Hot Encoding

Time-lagged Weather Conditions

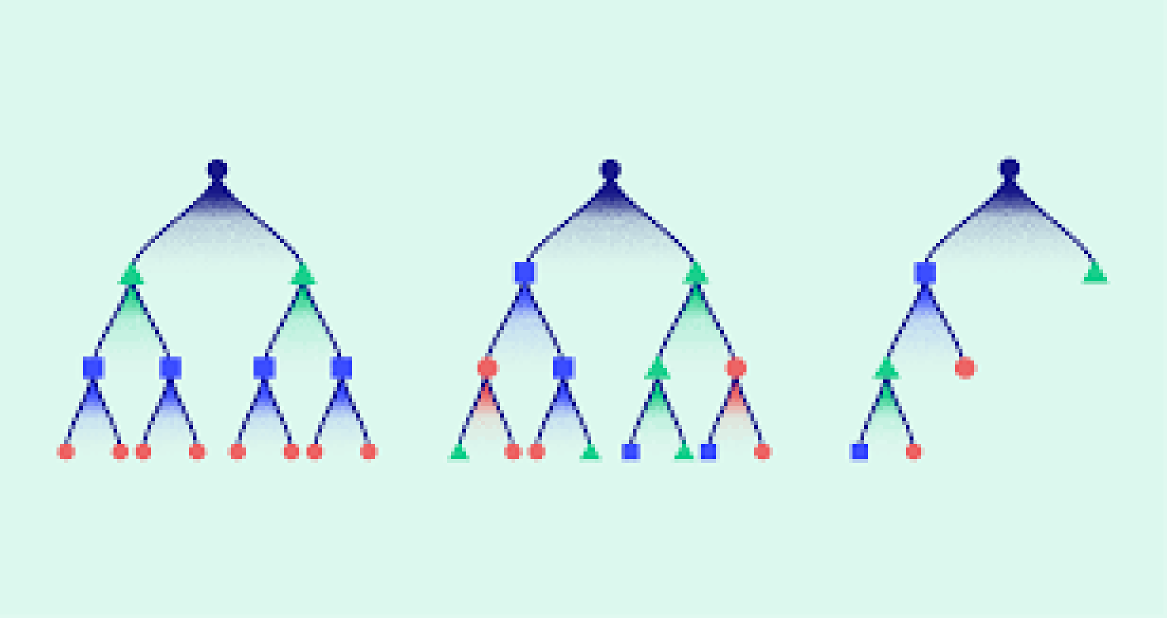
Interaction Terms

Principal Component Analysis

SMOTE

Multicollinearity Reduction

Modeling



Modeling Approach

Model Types

- Logistic Regression
- Random Forest
- XGBoost

Tuning Techniques

- Pipeline
- GridSearch
- PCA

CRITERIA
ANALYSIS
PERFORMANCE
APPROACH
100 %
EVALUATION
IMPROVEMENT
RESULTS
FEEDBACK
ASSESSMENT

Evaluation Approach

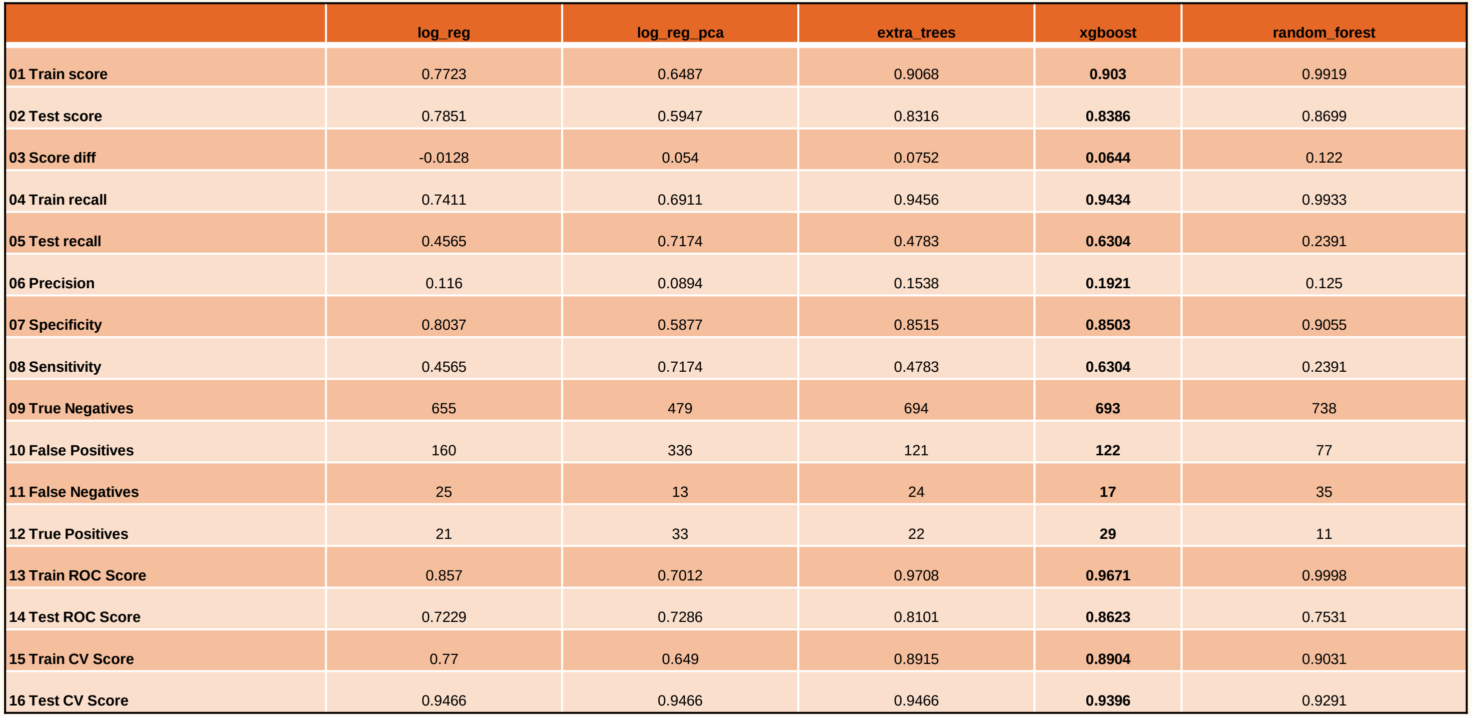
Metrics

- Accuracy
- Precision
- ROC_AUC

Methods

- Cross Validation (Kfold)
- Confusion Matrix
- Feature Importance Analysis
- ROC_AUC Curve
- Misclassification Analysis

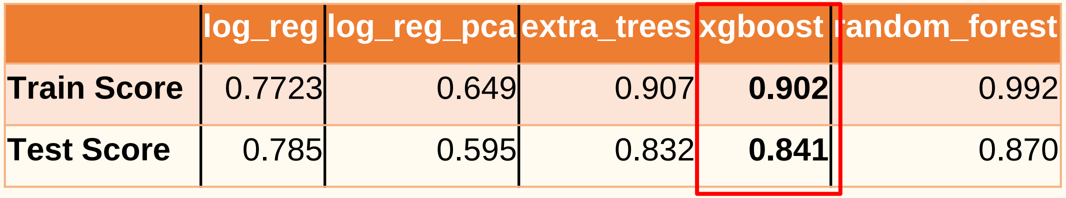
Model



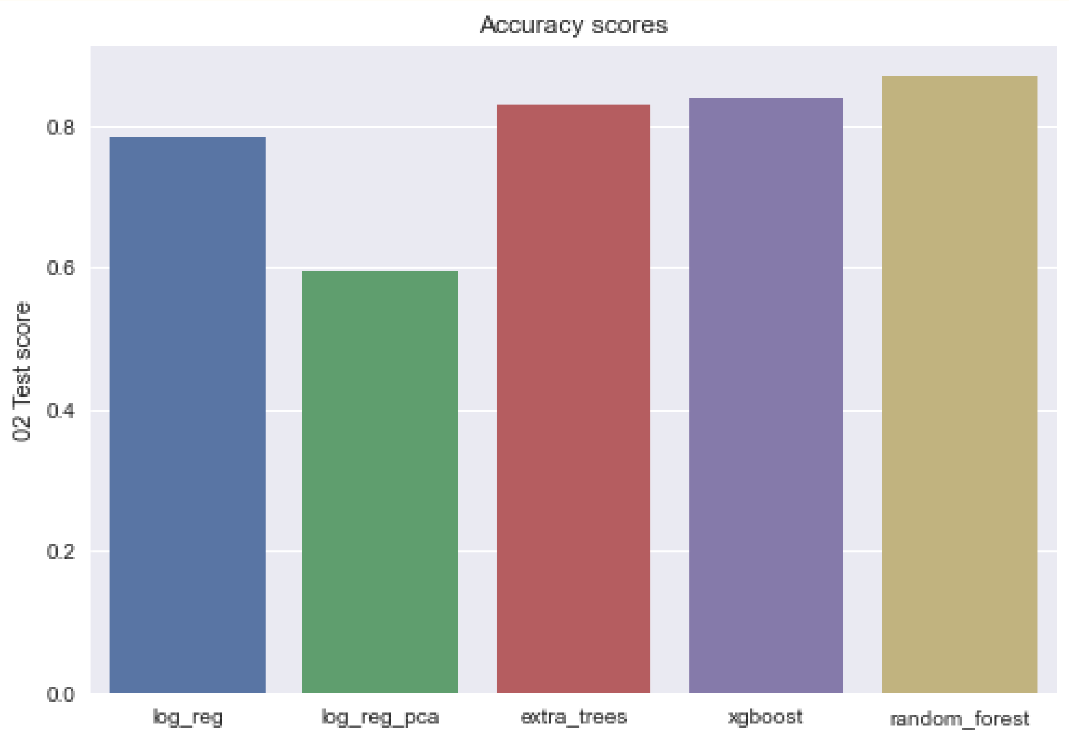
	log_reg	log_reg_pca	extra_trees	xgboost	random_forest
01 Train score	0.7723	0.6487	0.9068	**0.903**	0.9919
02 Test score	0.7851	0.5947	0.8316	**0.8386**	0.8699
03 Score diff	-0.0128	0.054	0.0752	**0.0644**	0.122
04 Train recall	0.7411	0.6911	0.9456	**0.9434**	0.9933
05 Test recall	0.4565	0.7174	0.4783	**0.6304**	0.2391
06 Precision	0.116	0.0894	0.1538	**0.1921**	0.125
07 Specificity	0.8037	0.5877	0.8515	**0.8503**	0.9055
08 Sensitivity	0.4565	0.7174	0.4783	**0.6304**	0.2391
09 True Negatives	655	479	694	**693**	738
10 False Positives	160	336	121	**122**	77
11 False Negatives	25	13	24	**17**	35
12 True Positives	21	33	22	**29**	11
13 Train ROC Score	0.857	0.7012	0.9708	**0.9671**	0.9998
14 Test ROC Score	0.7229	0.7286	0.8101	**0.8623**	0.7531
15 Train CV Score	0.77	0.649	0.8915	**0.8904**	0.9031
16 Test CV Score	0.9466	0.9466	0.9466	**0.9396**	0.9291

Evaluation

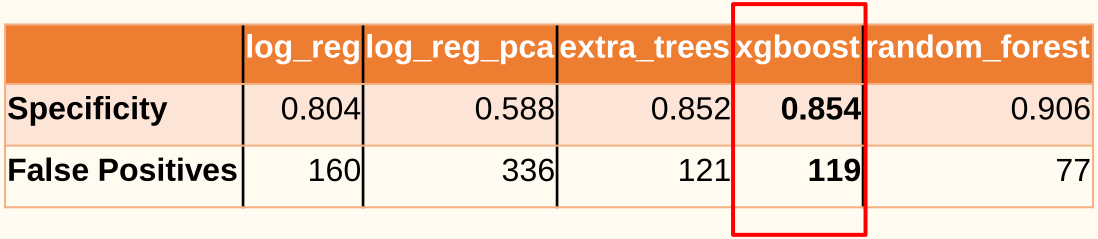
Accuracy



	log_reg	log_reg_pca	extra_trees	xgboost	random_forest
Train Score	0.7723	0.649	0.907	**0.902**	0.992
Test Score	0.785	0.595	0.832	**0.841**	0.870



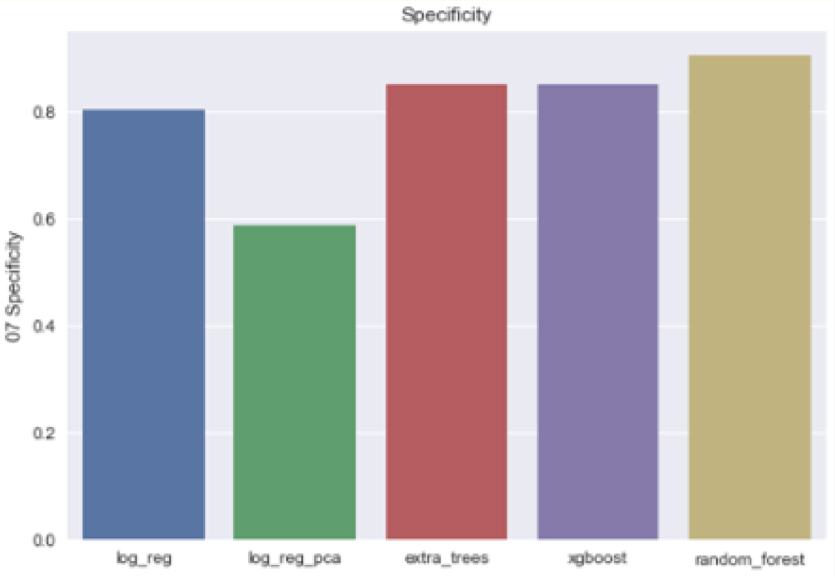
Evaluation

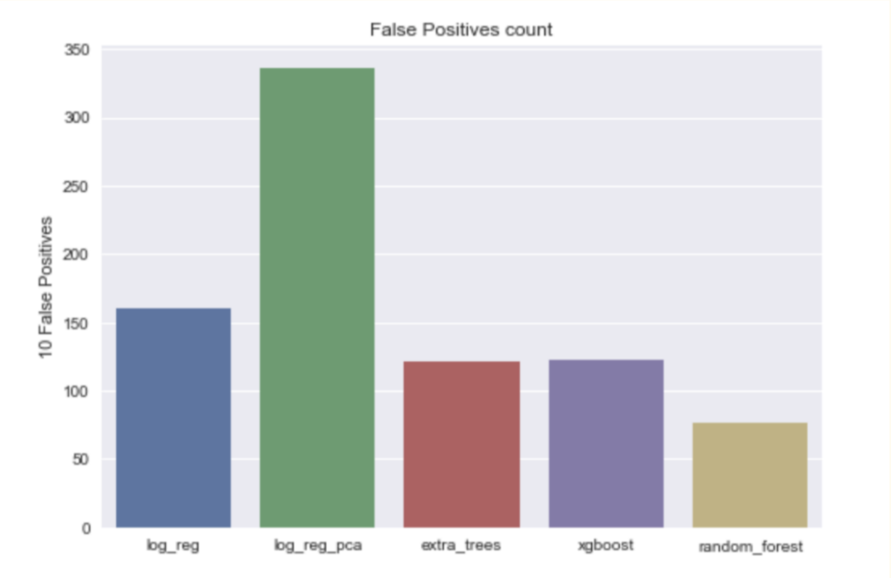


	log_reg	log_reg_pca	extra_trees	xgboost	random_forest
Specificity	0.804	0.588	0.852	**0.854**	0.906
False Positives	160	336	121	**119**	77

Specificity $\frac{TN}{TN+FP}$

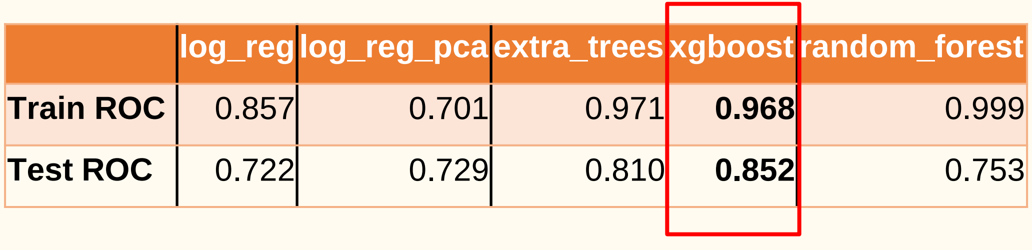
False Positive Count



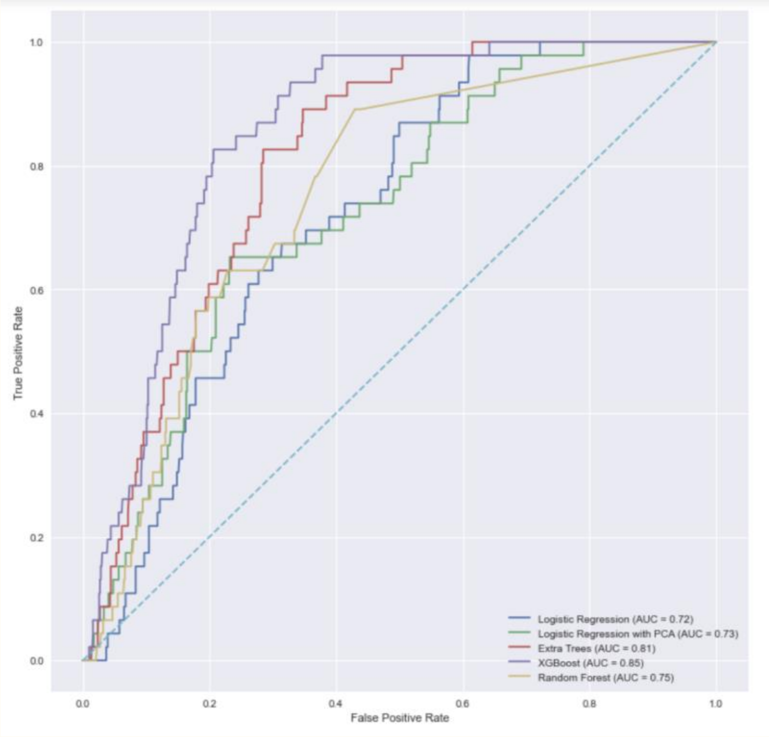


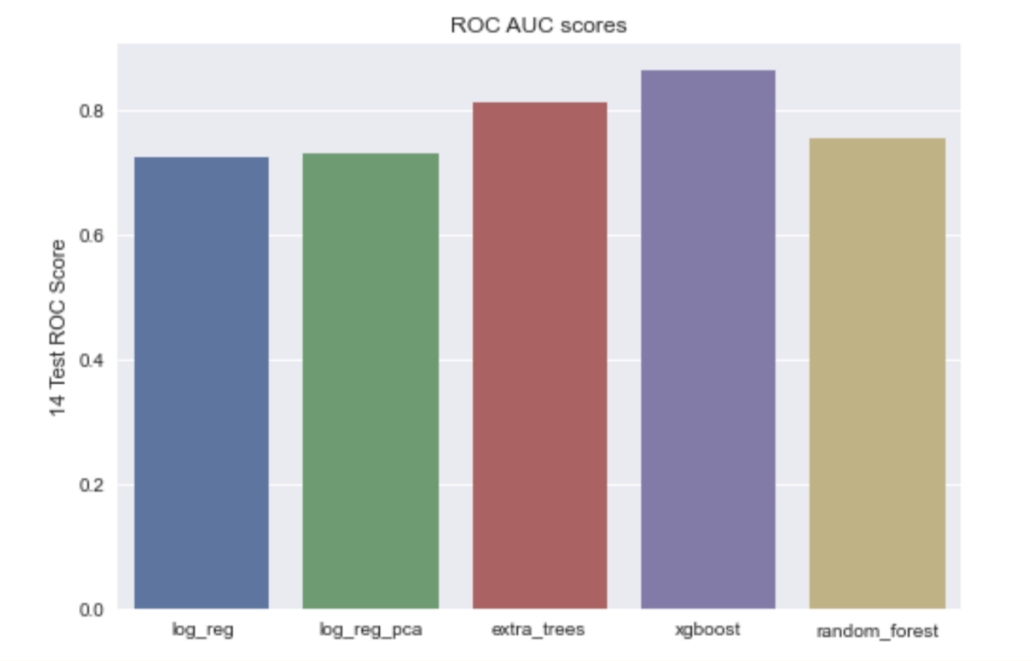
Evaluation

ROC - AUC Score



	log_reg	log_reg_pca	extra_trees	xgboost	random_forest
Train ROC	0.857	0.701	0.971	**0.968**	0.999
Test ROC	0.722	0.729	0.810	**0.852**	0.753

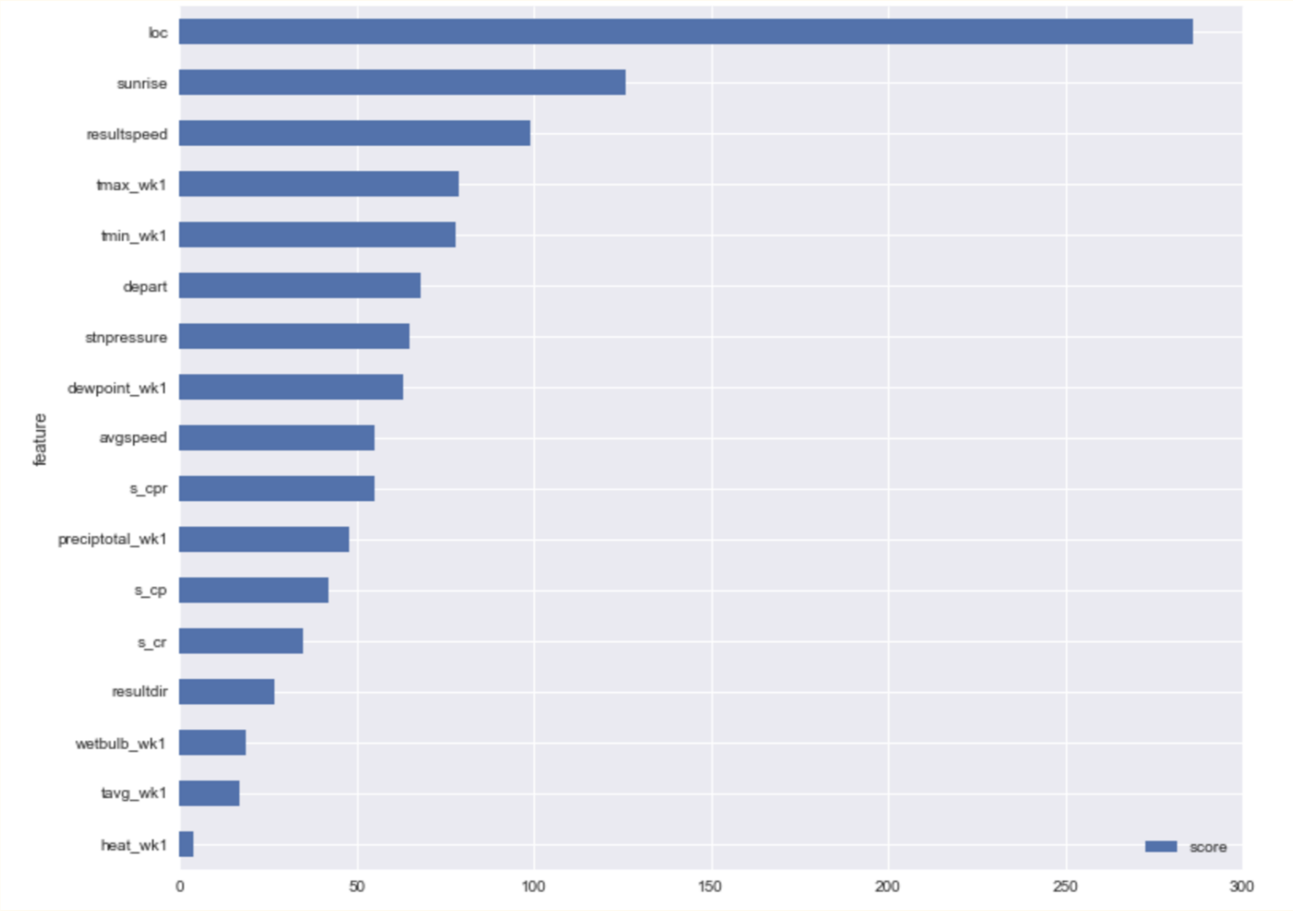




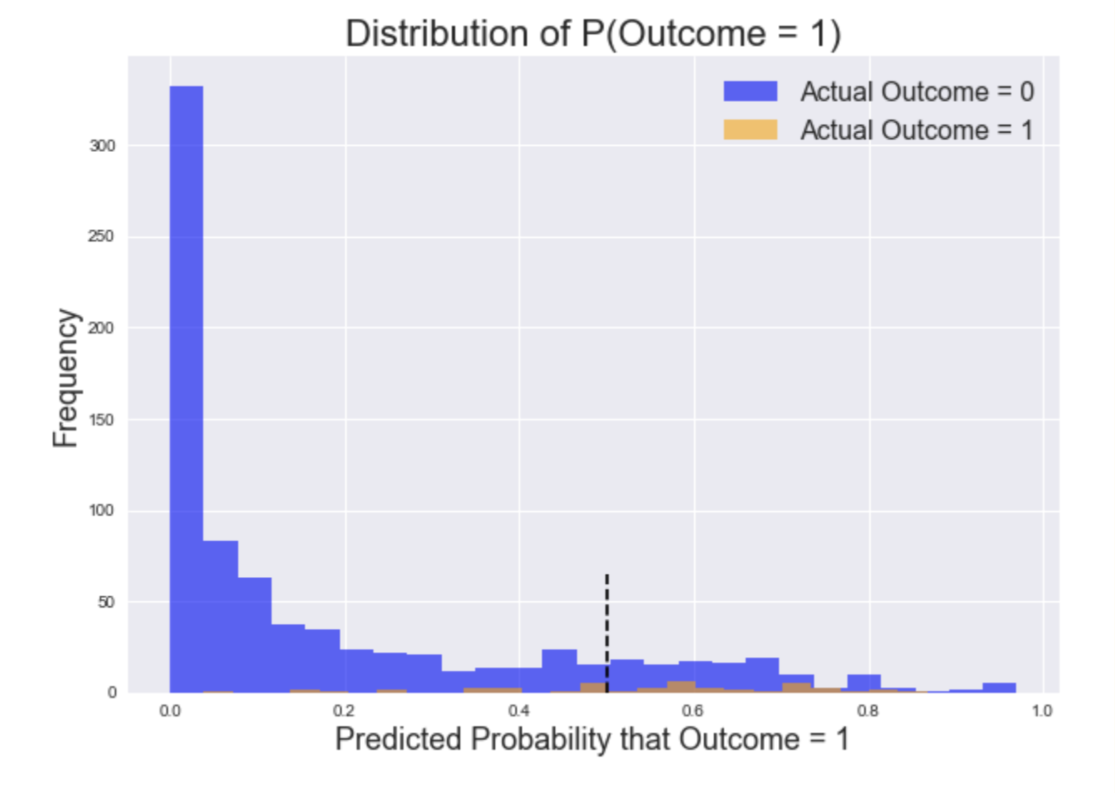
Evaluation

Top 5 Features:

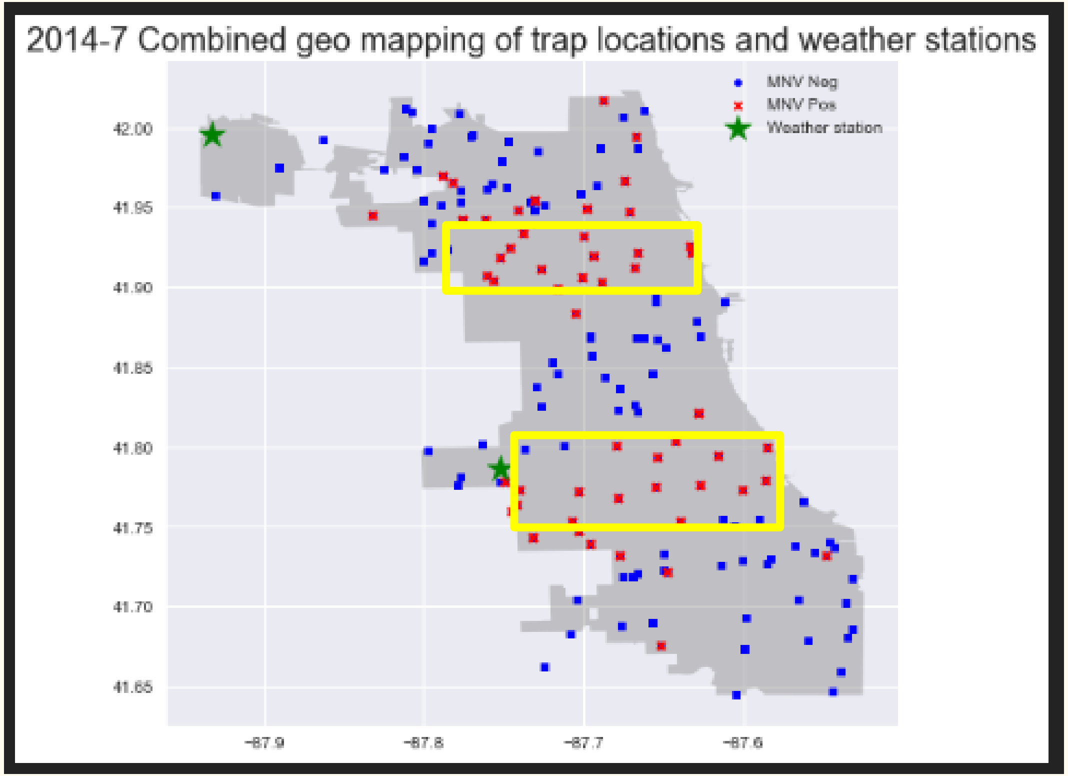
- loc
- sunrise
- resultspeed
- tmax_wk1
- tmin_wk1



Evaluation

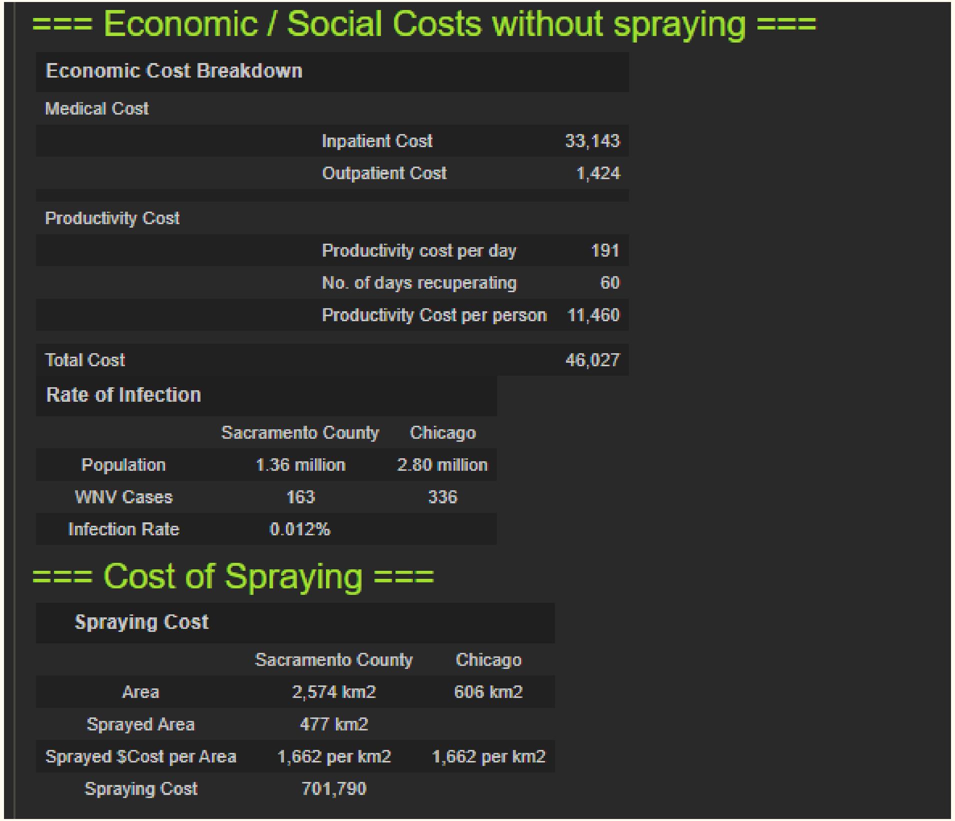

Distribution of P(Outcome = 1)
Actual Outcome = 0
Actual Outcome = 1
Frequency
300
250
200
150
100
50
0
0.0
0.2
0.4
0.6
0.8
1.0
Predicted Probability that Outcome = 1

Model Prediction



XG Boost Model Prediction for 2014 - July

Cost Benefit Analysis



=== Economic / Social Costs without spraying ===

Economic Cost Breakdown		
Medical Cost		
	Inpatient Cost	33,143
	Outpatient Cost	1,424
Productivity Cost		
	Productivity cost per day	191
	No. of days recuperating	60
	Productivity Cost per person	11,460
Total Cost		46,027

Rate of Infection		
	Sacramento County	Chicago
Population	1.36 million	2.80 million
WNV Cases	163	336
Infection Rate	0.012%	

=== Cost of Spraying ===

Spraying Cost		
	Sacramento County	Chicago
Area	2,574 km2	606 km2
Sprayed Area	477 km2	
Sprayed $Cost per Area	1,662 per km2	1,662 per km2
Spraying Cost	701,790	

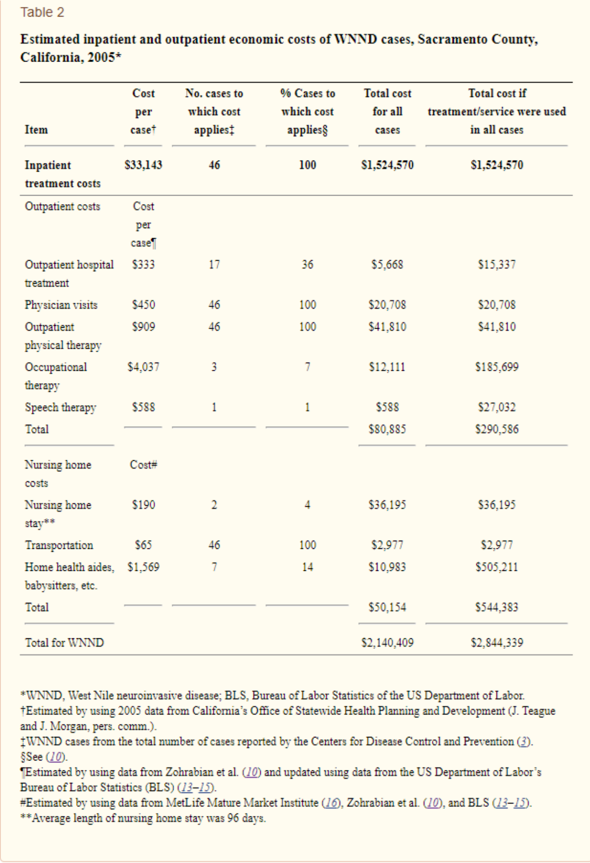


Table 2

Estimated inpatient and outpatient economic costs of WNND cases, Sacramento County, California, 2005*

Item	**Cost per case†**	**No. cases to which cost applies‡**	**% Cases to which cost applies§**	**Total cost for all cases**	**Total cost if treatment/service were used in all cases**
Inpatient treatment costs	**$33,143**	**46**	**100**	**$1,524,570**	**$1,524,570**
Outpatient costs	Cost per case¶				
Outpatient hospital treatment	$333	17	36	$5,668	$15,337
Physician visits	$450	46	100	$20,708	$20,708
Outpatient physical therapy	$909	46	100	$41,810	$41,810
Occupational therapy	$4,037	3	7	$12,111	$185,699
Speech therapy	$588	1	1	$588	$27,032
Total				$80,885	$290,586
Nursing home costs	Cost#				
Nursing home stay**	$190	2	4	$36,195	$36,195
Transportation	$65	46	100	$2,977	$2,977
Home health aides, babysitters, etc.	$1,569	7	14	$10,983	$505,211
Total				$50,154	$544,383
Total for WNND				$2,140,409	$2,844,339

*WNND, West Nile neuroinvasive disease; BLS, Bureau of Labor Statistics of the US Department of Labor.

†Estimated by using 2005 data from California's Office of Statewide Health Planning and Development (J. Teague and J. Morgan, pers. comm.).

‡WNND cases from the total number of cases reported by the Centers for Disease Control and Prevention (3).

§See (10).

¶Estimated by using data from Zohrabian et al. (10) and updated using data from the US Department of Labor's Bureau of Labor Statistics (BLS) (13–15).

#Estimated by using data from MetLife Mature Market Institute (16), Zohrabian et al. (10), and BLS (13–15).

**Average length of nursing home stay was 96 days.

Conclusion

XG Boost model was selected to predict the WNV presence in the 2014 traps.

We drilled down into years of july as the outbreak starts to occur, should spray in mid june

Cost of spraying vs cost of non-spraying

$1.01 million vs $15.40 million

- Adopt genetically modified mosquitoes

Thank You