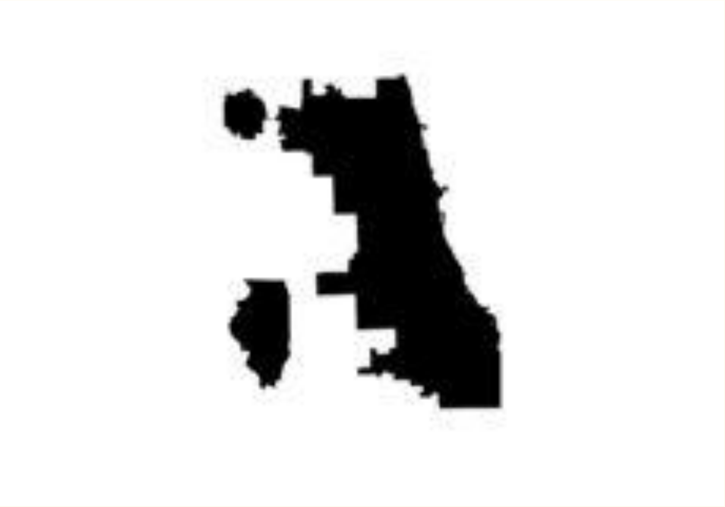
WEST NILE VIRUS

Pesticides - To spray or not to spray?

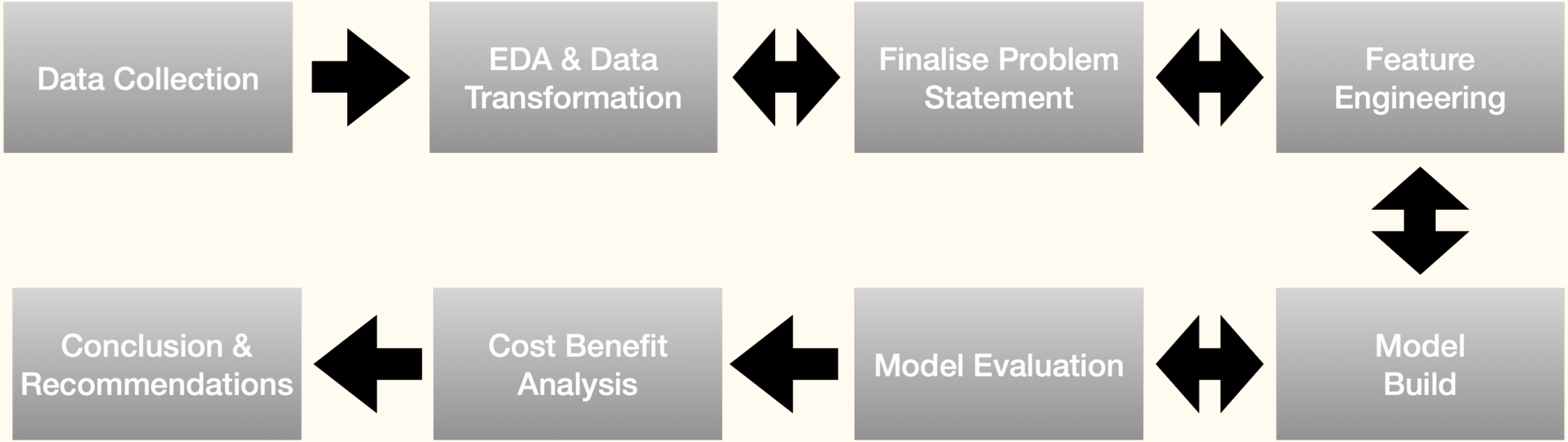
DSI 16 Project 4 : Dominic Ong / Vikaskalia / Peter Wong / Jeriel Wong / Cheyanne Wong

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

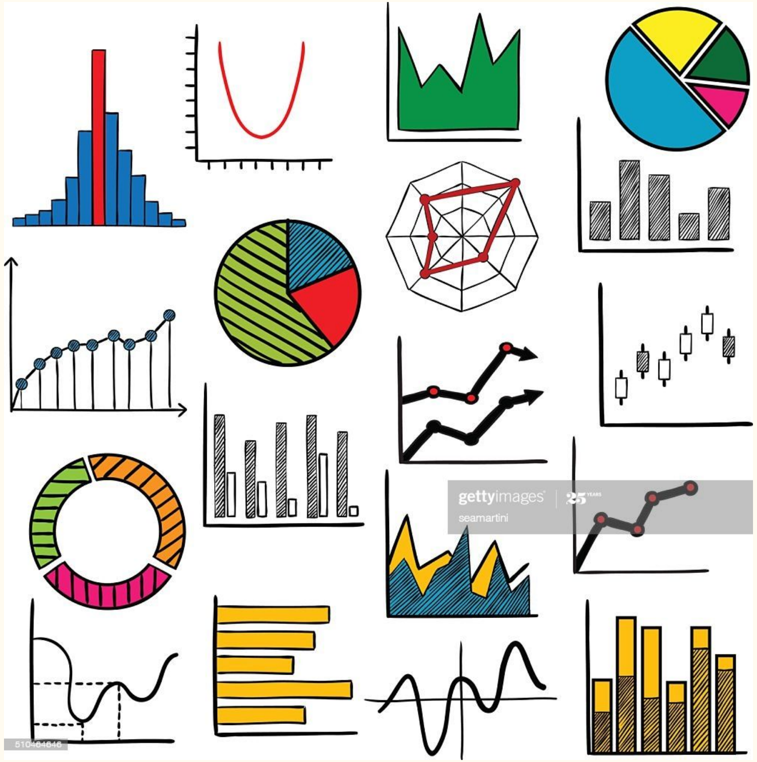
Merging Rows > 50 Mosquitos

Merge Train & Weather dataset

Data Imputation

Weather Station 1 & 2 Ffill

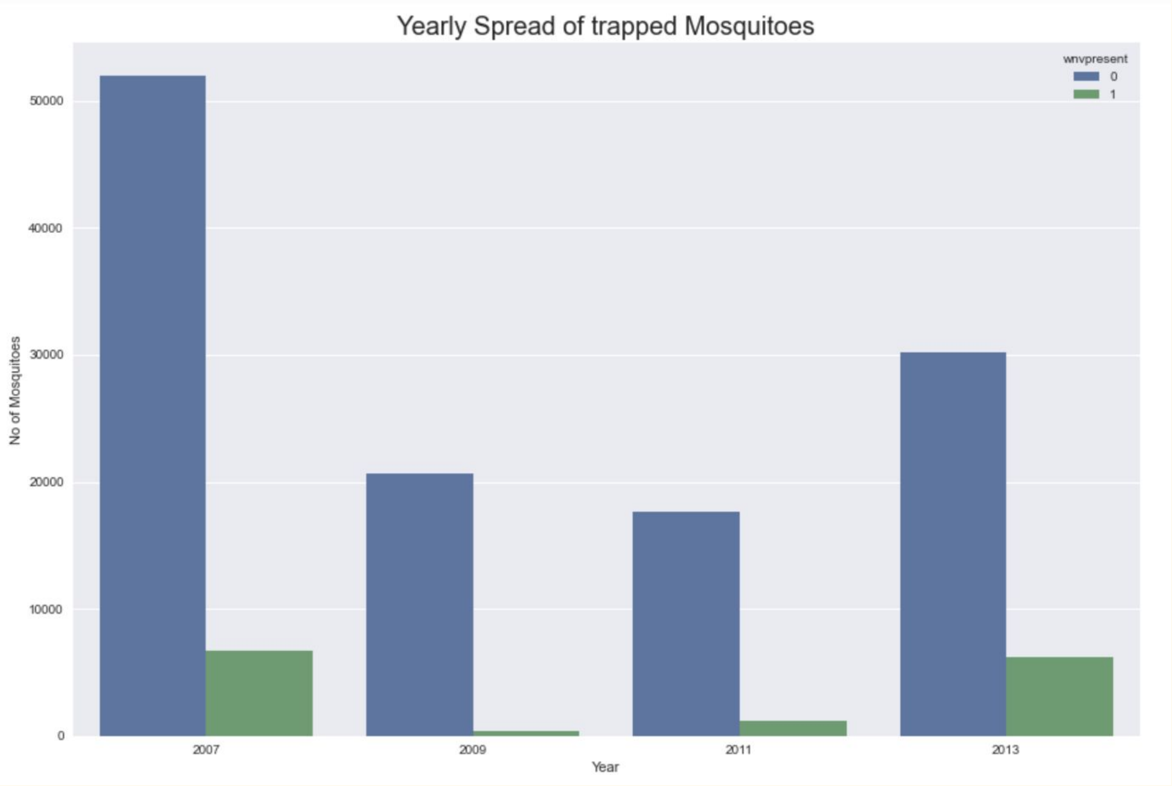
EDA



EDA

Spread of Trapped Mosquitoes By Year

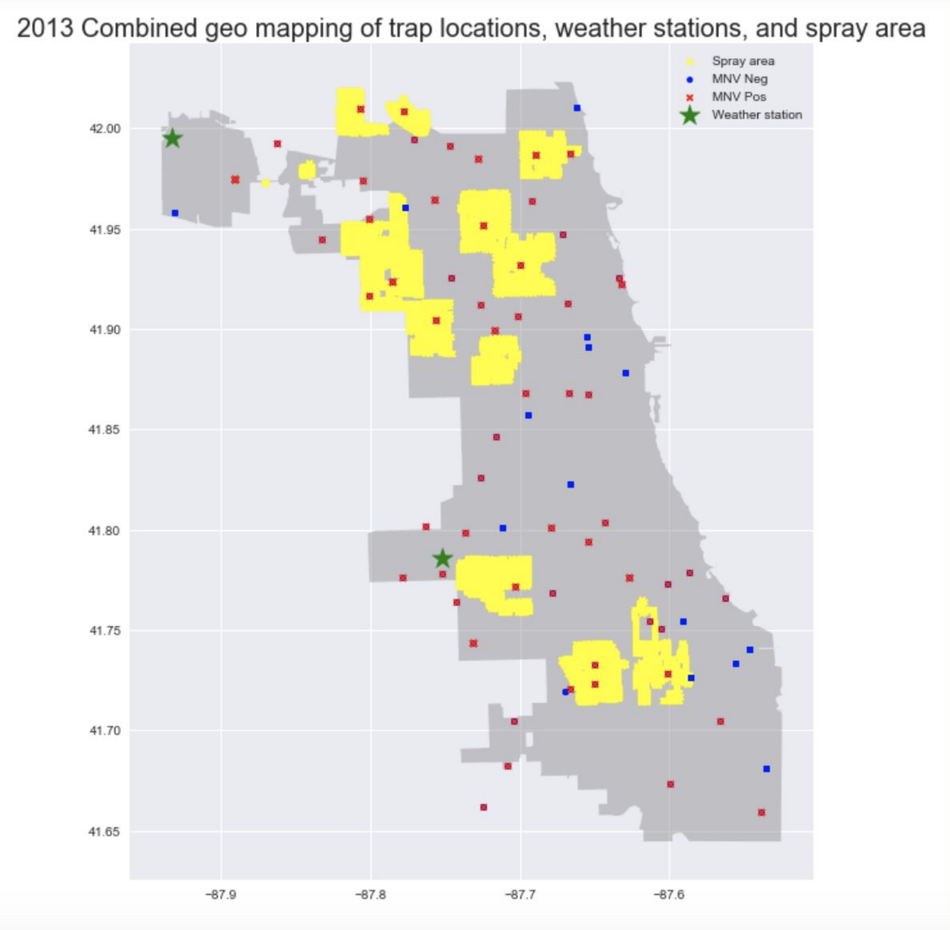
Even though the total number of mosquitoes caught in 2013 was lower than that of 2013, the percentage of WNV presence went up in 2013.



EDA

2013 Trap locations and Spray Area

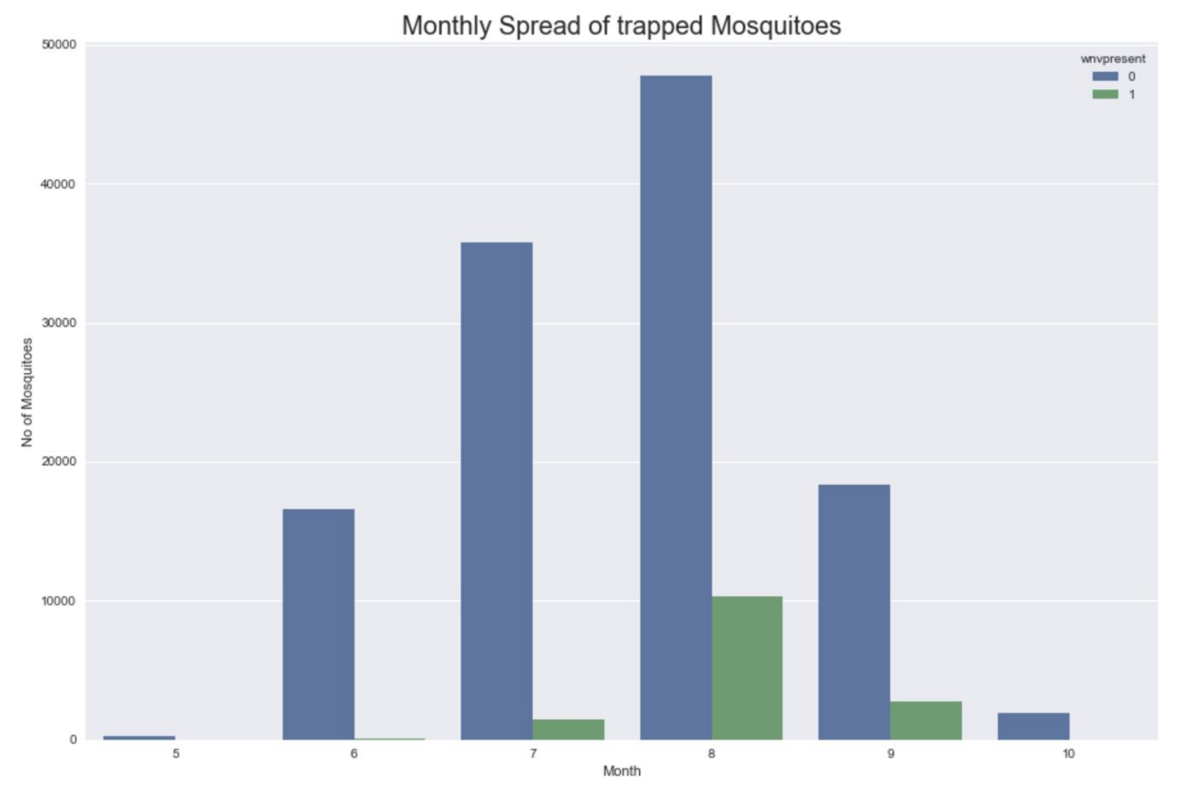
In 2013, WNV presence was found in most traps across the city. The area near Station 1 in the northern region seems to be a hotspot for WNV presence. The spraying of pesticide is concentrated in this region.



EDA

Spread of Trapped Mosquitoes By Month

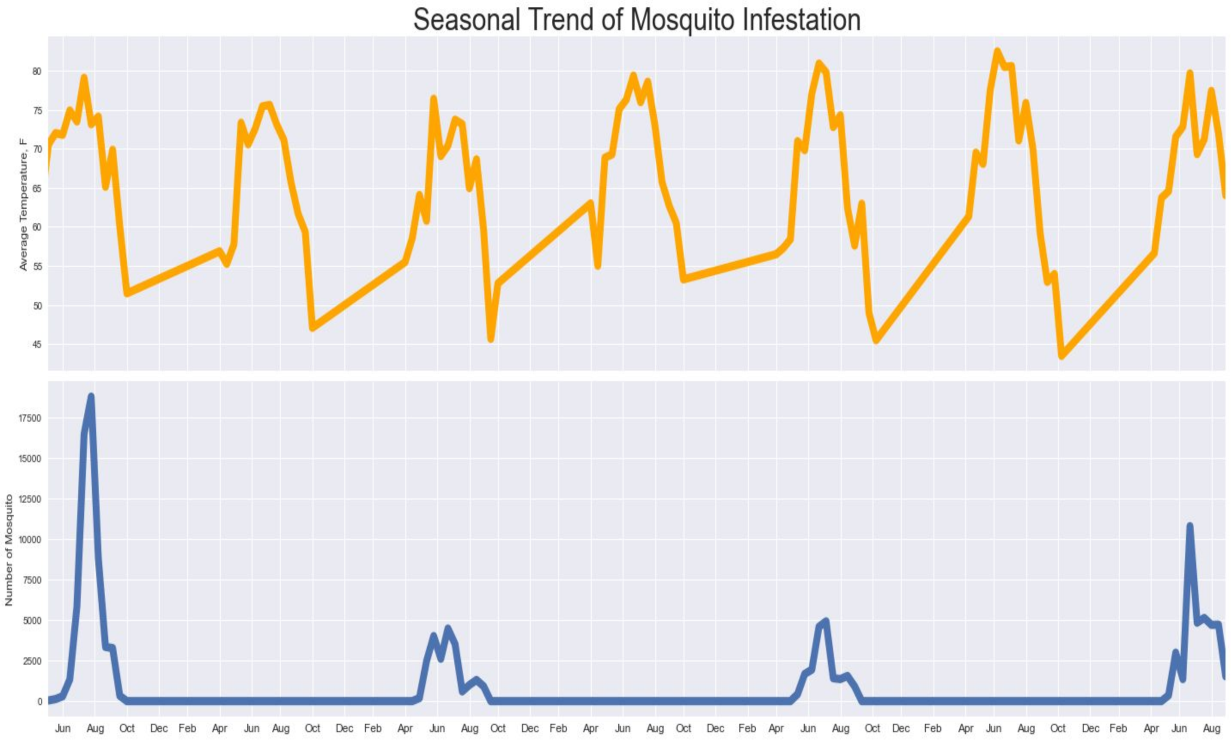
Number of mosquitoes trapped was the highest in the month of August where the weather is hot and humid. The presence of WNV was also higher in this month.



EDA

Seasonal Trend of Mosquitoes Infestation With Ave Temperature

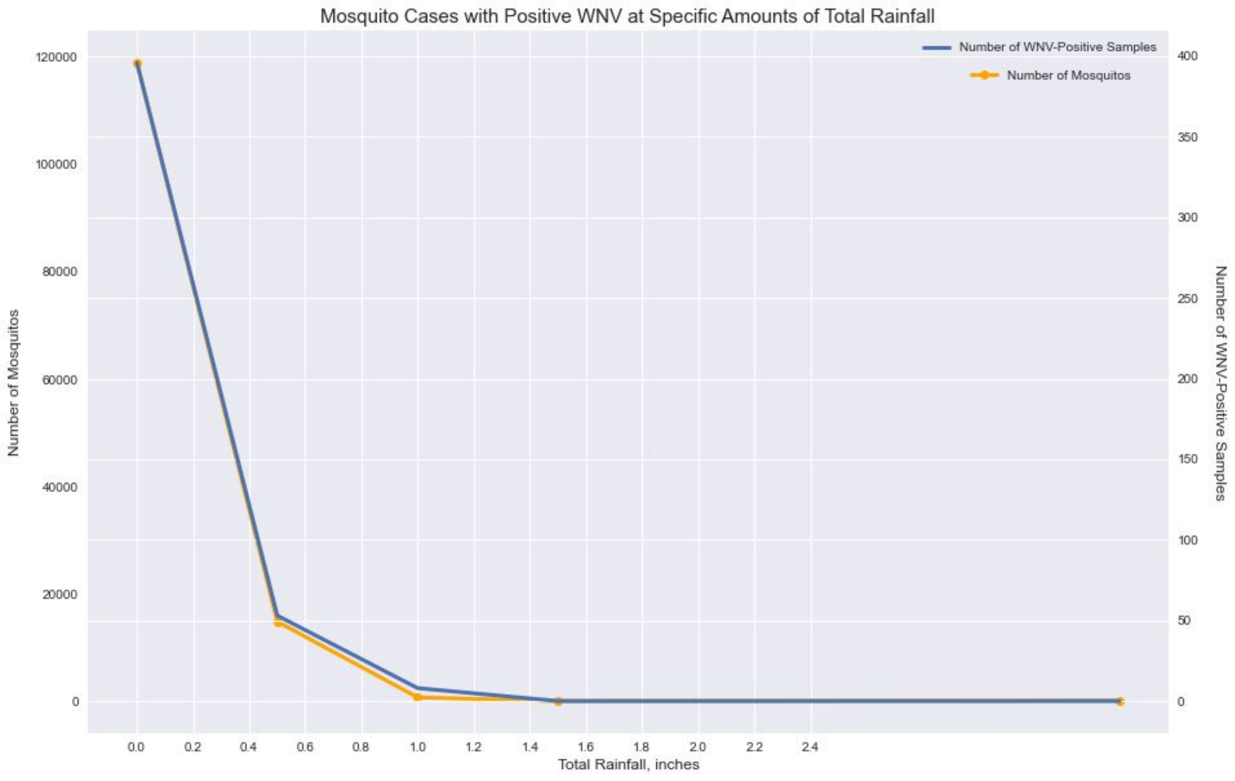
The above graph shows that mosquitoes prefer the higher temperatures as when temperature increase so does the number of mosquitoes.



EDA

No. of Mosquitoes Cases with Total Rainfall

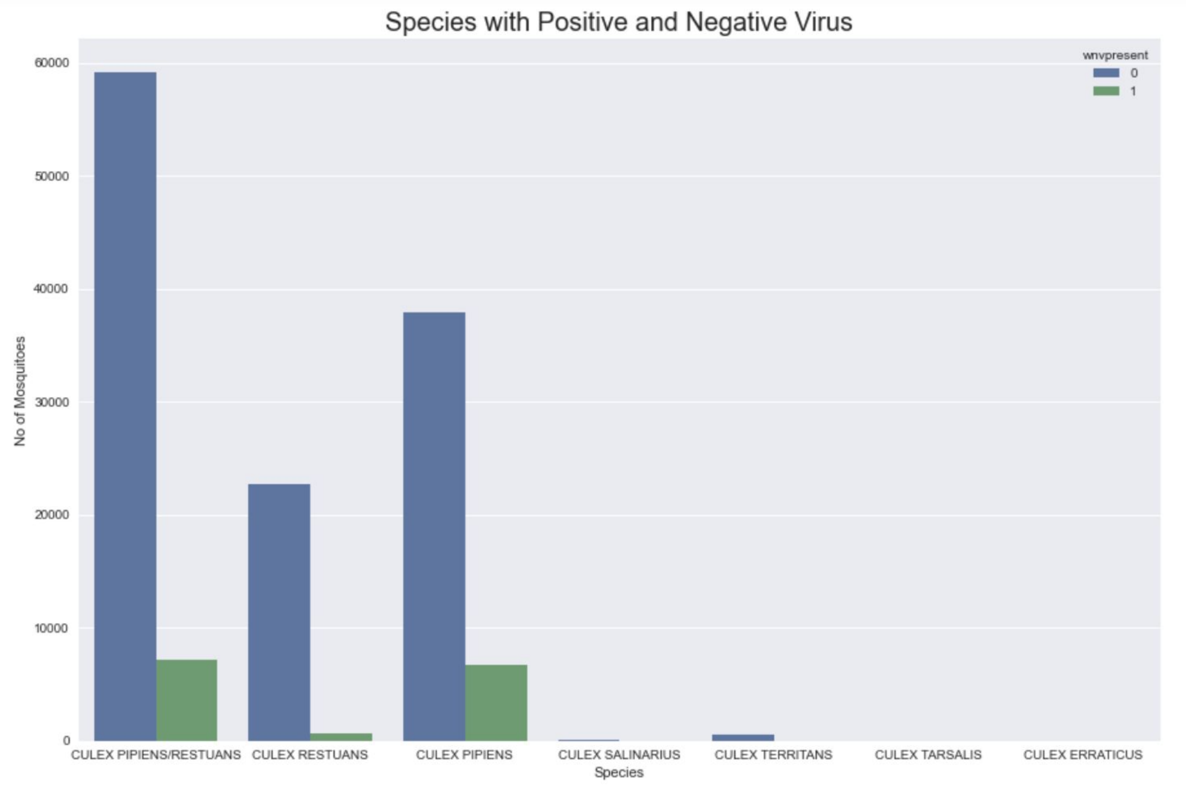
Total rainfall (precipitation) is inversely proportional to both the number of mosquitos and number of WNV-positive traps.



EDA

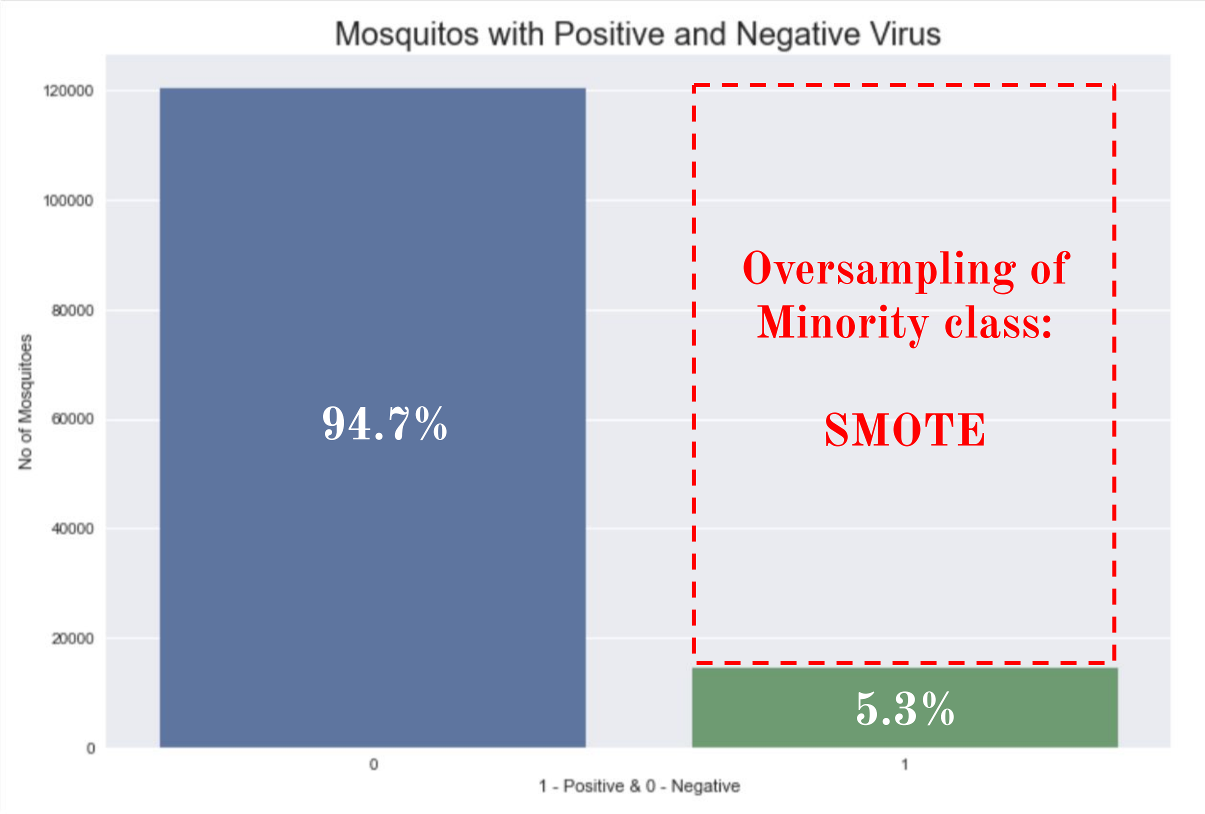
Mosquito Species with Positive & Negative Virus

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

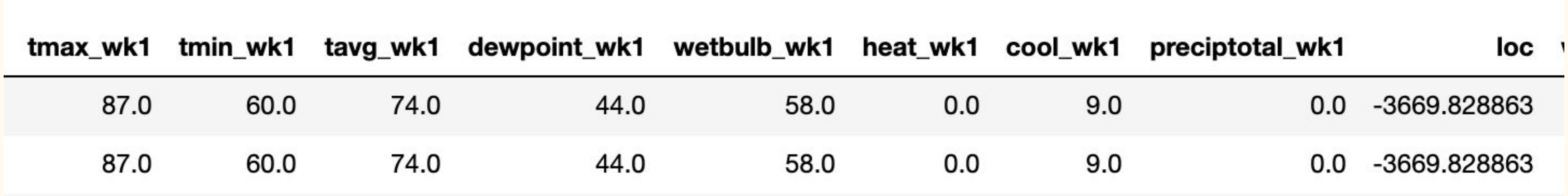
Time-lagged Weather Conditions

Interaction Terms

Principal Component Analysis

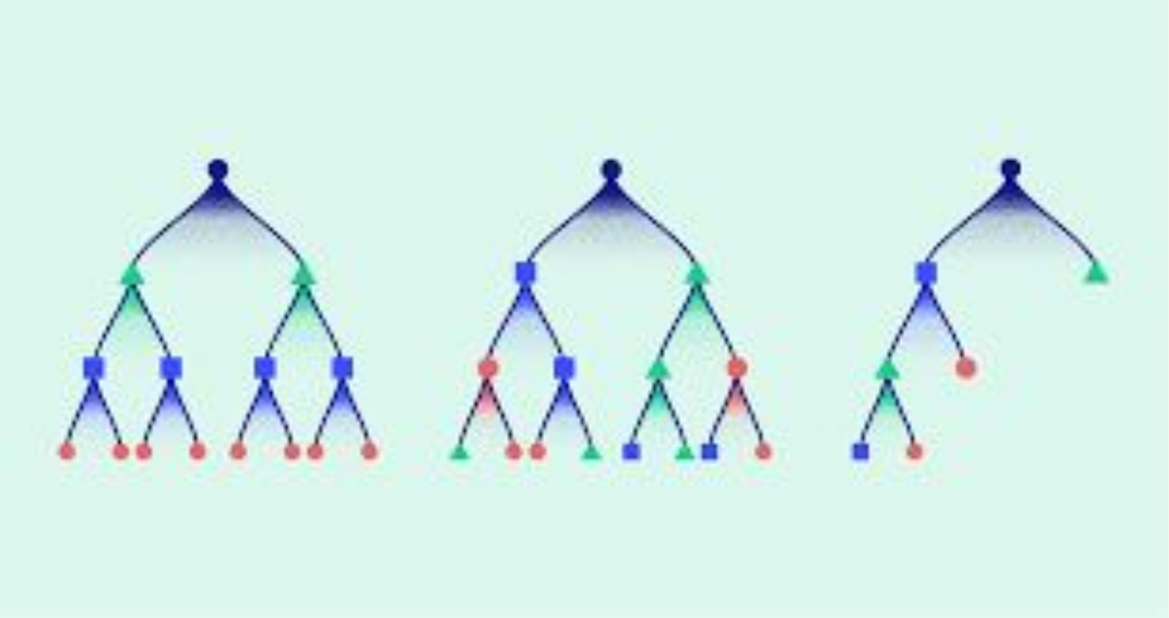
SMOTE

Multicollinearity Reduction



tmax_wk1	tmin_wk1	tavg_wk1	dewpoint_wk1	wetbulb_wk1	heat_wk1	cool_wk1	preciptotal_wk1	loc
87.0	60.0	74.0	44.0	58.0	0.0	9.0	0.0	-3669.828863
87.0	60.0	74.0	44.0	58.0	0.0	9.0	0.0	-3669.828863

Modeling



Modeling Approach

Model Types

- Logistic Regression
- Random Forest
- XGBoost

Tuning Techniques

- Pipeline
- GridSearch
- PCA

CRITERIA
ANALYSIS
PERFORMANCE
100 %
APPROACH
EVALUATION
IMPROVEMENT
RESULTS
FEEDBACK
ASSESSMENT

Evaluation Approach

Metrics

- Accuracy
- ROC_AUC
- Specificity

Methods

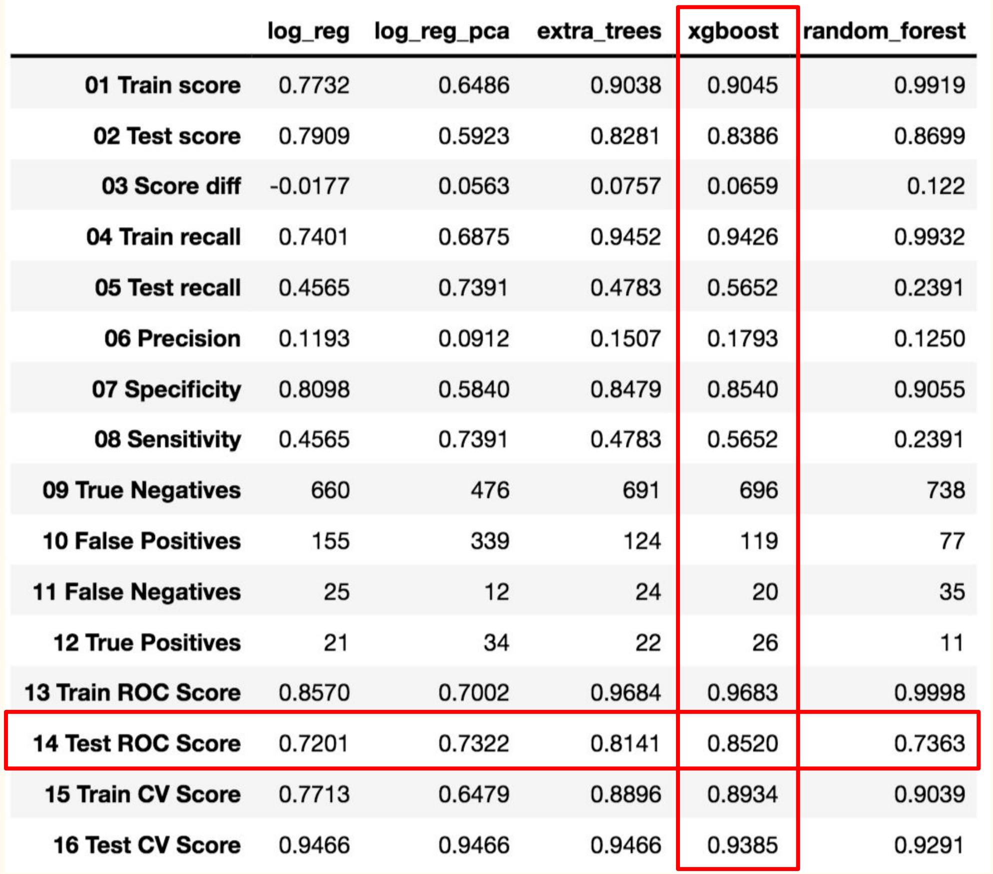
- Cross Validation (Kfold)
- Confusion Matrix
- Feature Importance Analysis
- ROC_AUC Curve
- Misclassification Analysis

Model

Evaluation Metrics:

- Accuracy
- Specificity
- ROC-AUC Score

XGBoost is our best model!



	log_reg	log_reg_pca	extra_trees	xgboost	random_forest
01 Train score	0.7732	0.6486	0.9038	0.9045	0.9919
02 Test score	0.7909	0.5923	0.8281	0.8386	0.8699
03 Score diff	-0.0177	0.0563	0.0757	0.0659	0.122
04 Train recall	0.7401	0.6875	0.9452	0.9426	0.9932
05 Test recall	0.4565	0.7391	0.4783	0.5652	0.2391
06 Precision	0.1193	0.0912	0.1507	0.1793	0.1250
07 Specificity	0.8098	0.5840	0.8479	0.8540	0.9055
08 Sensitivity	0.4565	0.7391	0.4783	0.5652	0.2391
09 True Negatives	660	476	691	696	738
10 False Positives	155	339	124	119	77
11 False Negatives	25	12	24	20	35
12 True Positives	21	34	22	26	11
13 Train ROC Score	0.8570	0.7002	0.9684	0.9683	0.9998
14 Test ROC Score	0.7201	0.7322	0.8141	0.8520	0.7363
15 Train CV Score	0.7713	0.6479	0.8896	0.8934	0.9039
16 Test CV Score	0.9466	0.9466	0.9466	0.9385	0.9291

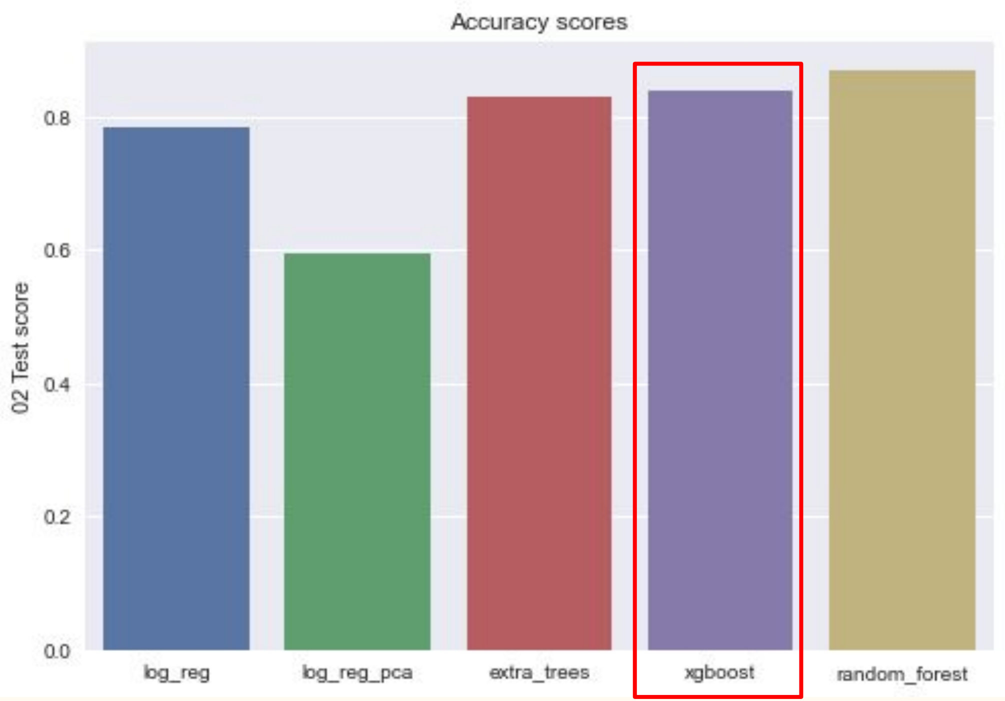
Evaluation

	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

Accuracy

Ratio of correctly predicted observation to the total observations

$$\frac{TN+TP}{TN+FP+TP+FN}$$

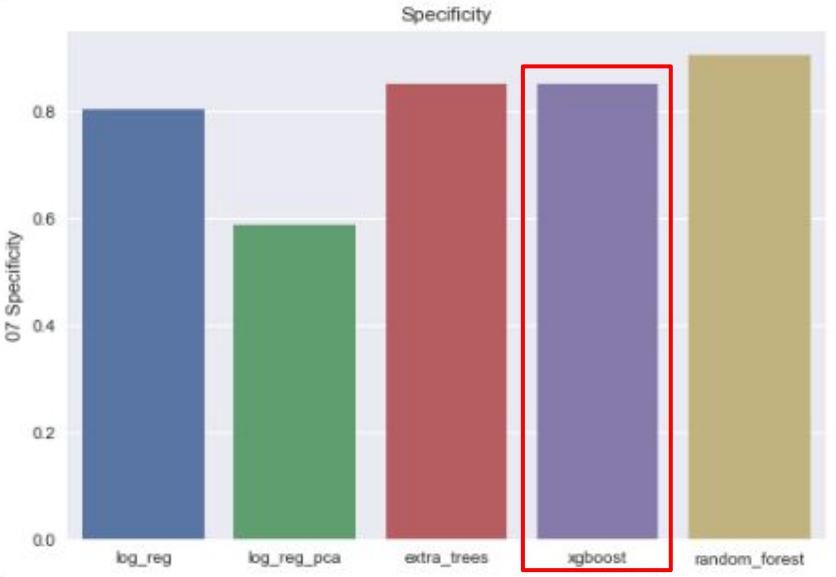


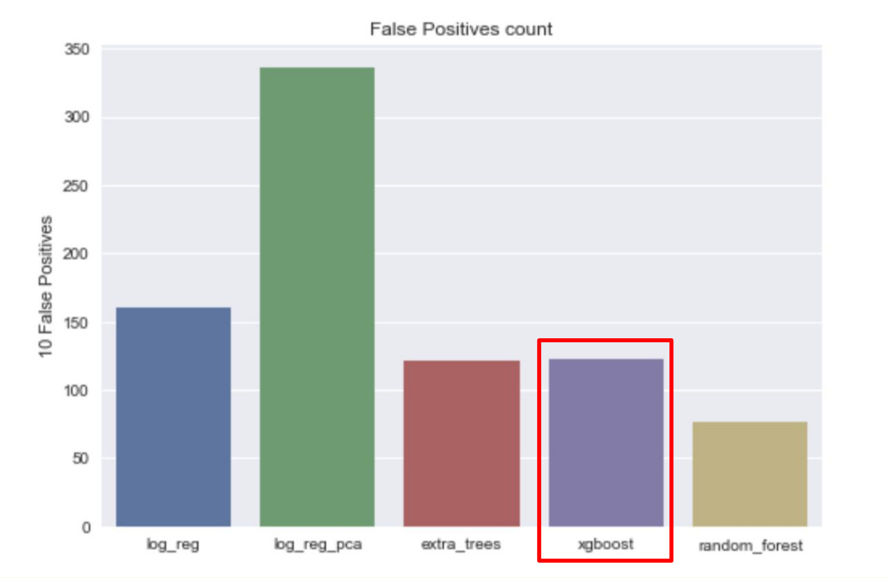
Evaluation

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

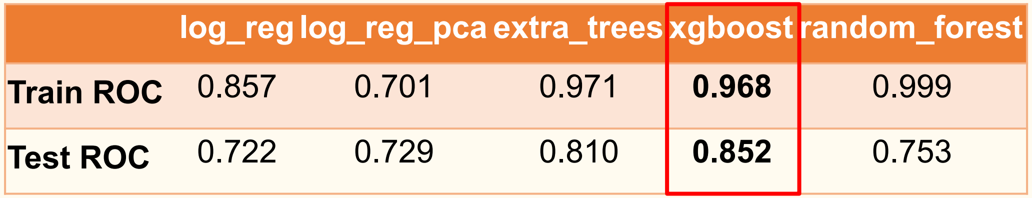
False Positive Count (WNV Mosquitoes)



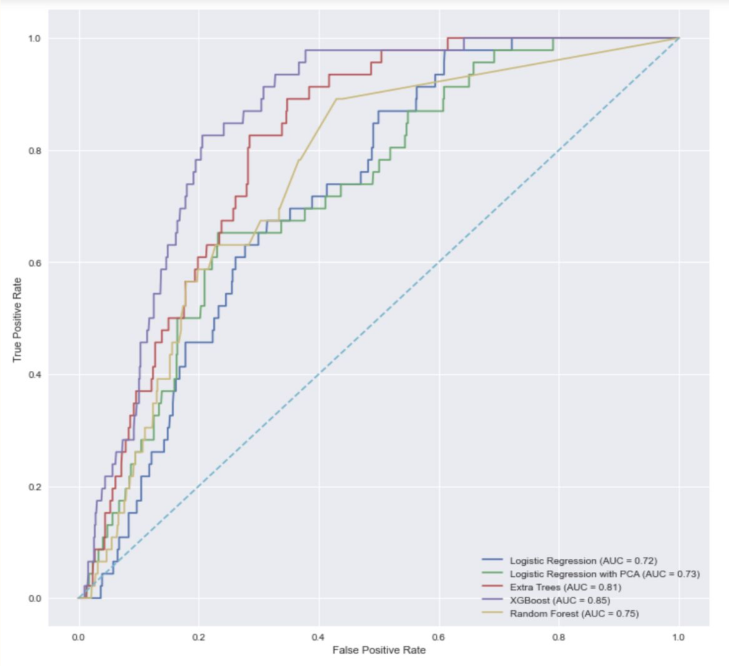


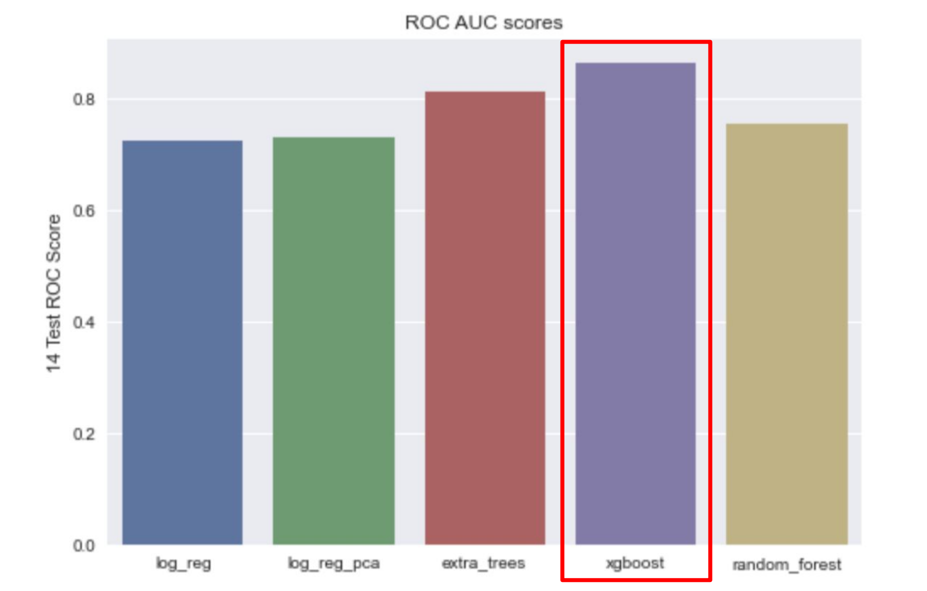
Evaluation

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

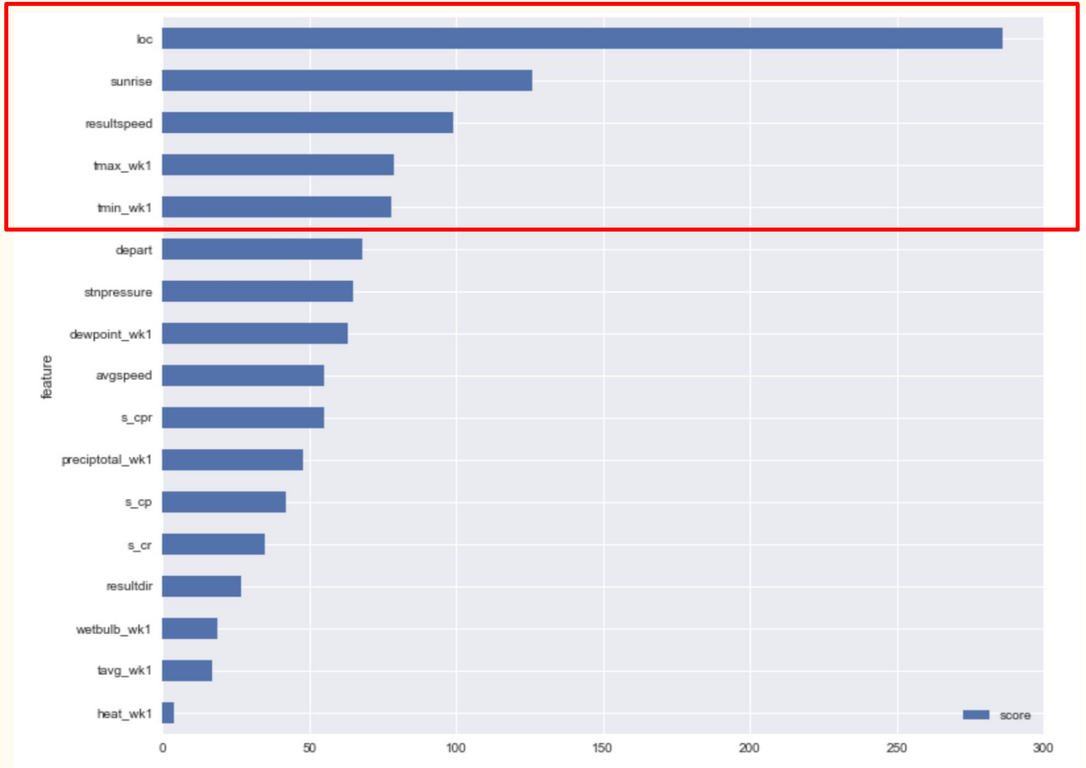




Evaluation

Feature Importance Top 5 Features:

- loc
- sunrise
- resultspeed
- tmax_wk1
- tmin_wk1

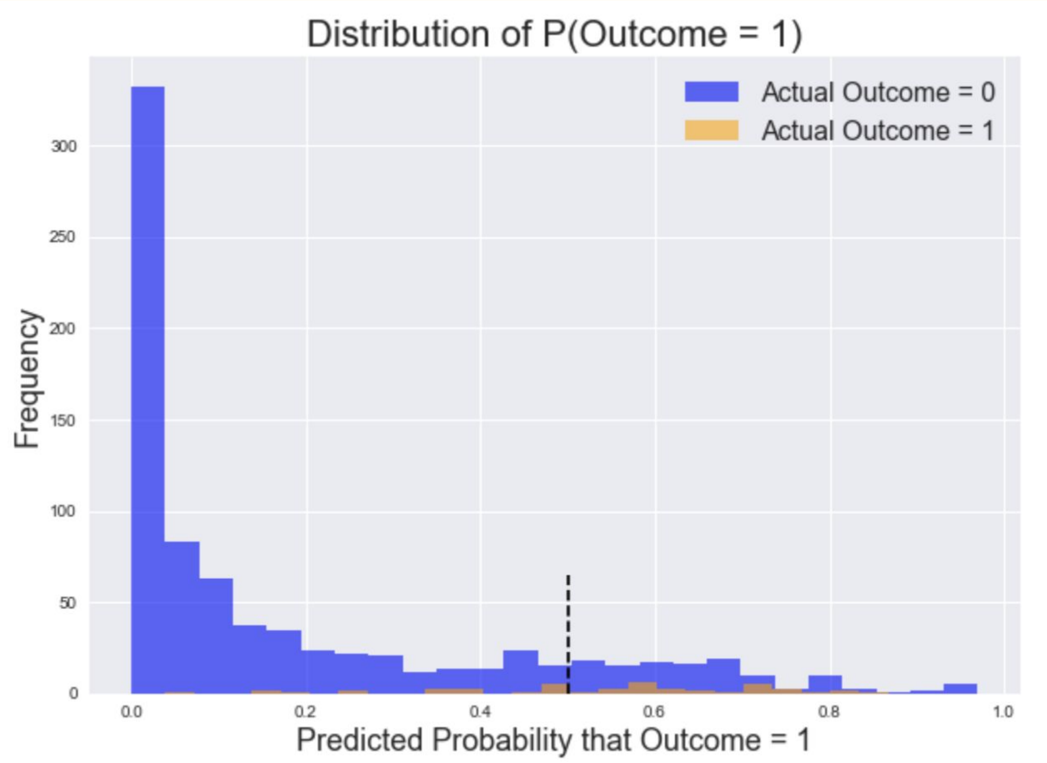


Importance plot provides a score that indicates how useful each feature was in the construction of the boosted decision trees within the model. The more an attribute is used to make key decisions with decision trees, the higher its relative importance.

Evaluation

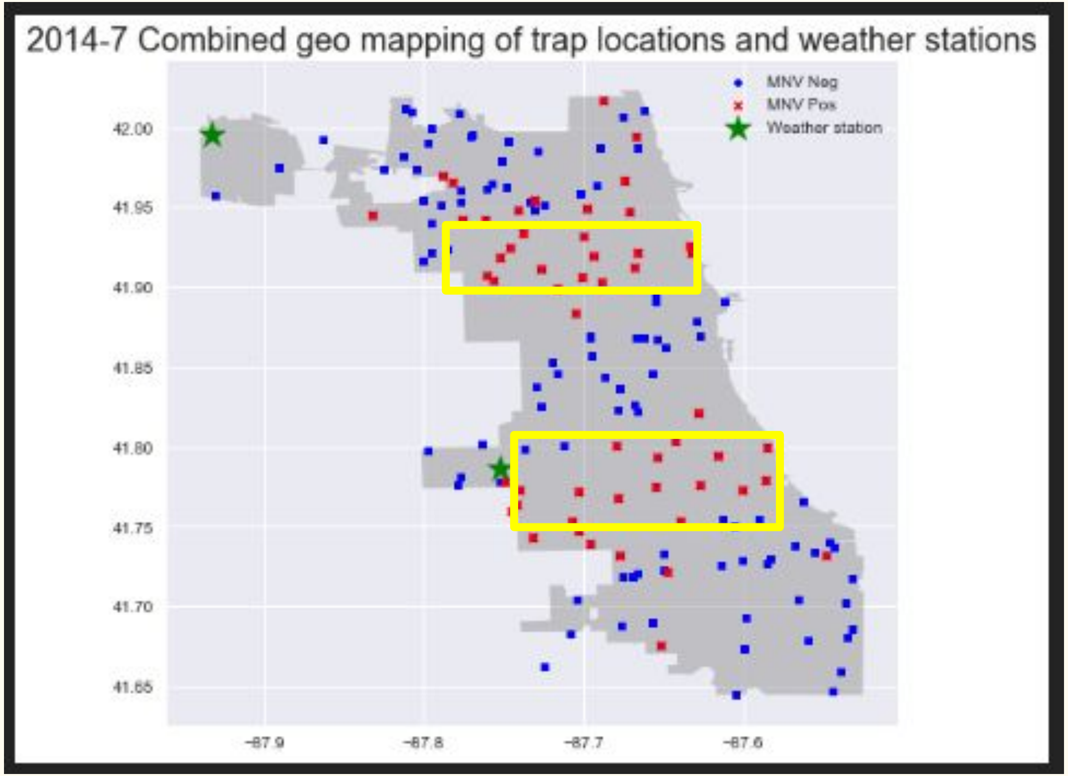
Distribution of Probability:

- 119 False Positives
- 20 False Negatives



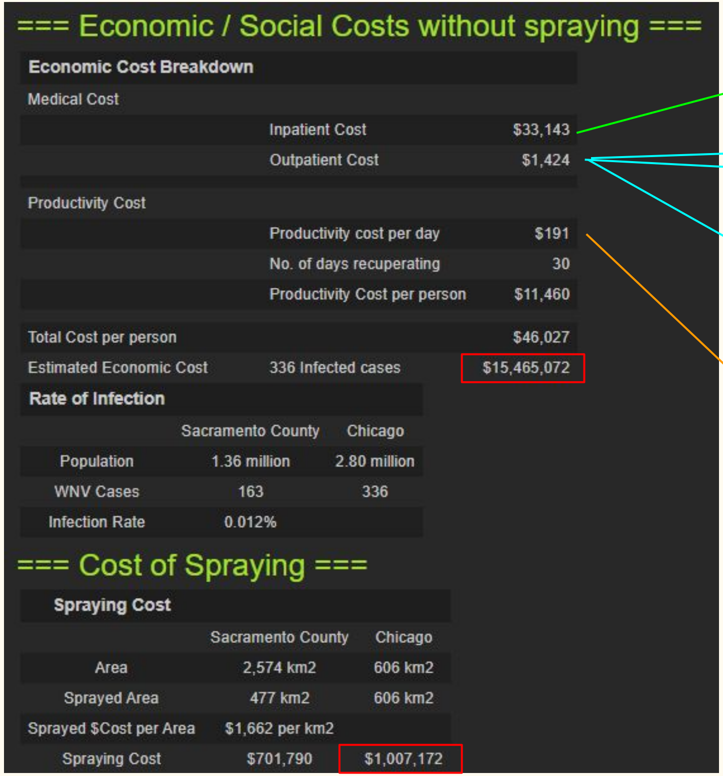
Minimizing the # of False Negatives is of greater importance in this problem. Predicting Mosquitoes to have WNV when do they actually do not have it (False Positives) is less of a concern than predicting Mosquitoes not to have WNV when they actually do have it (False Negatives).

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	30
	Productivity Cost per person	$11,460
Total Cost per person		$46,027
Estimated Economic Cost	336 Infected cases	$15,465,072

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	606 km2
Sprayed $Cost per Area	$1,662 per km2	
Spraying Cost	$701,790	$1,007,172

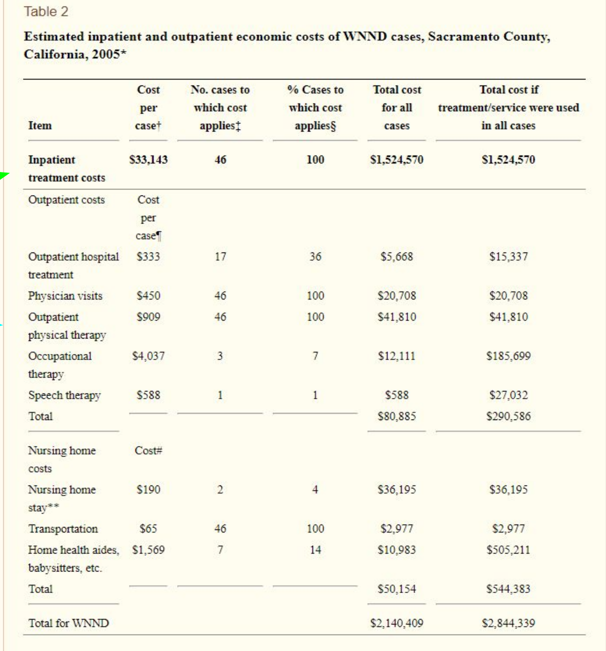


Table 2

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

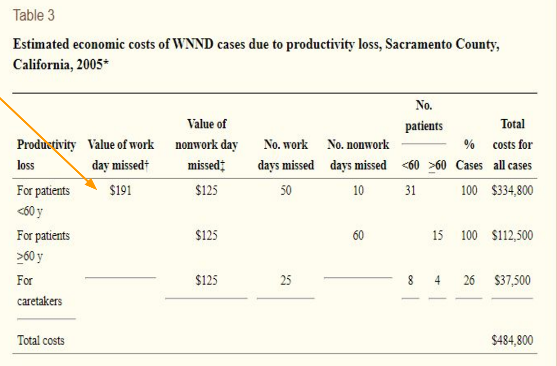


Table 3

Estimated economic costs of WNND cases due to productivity loss, Sacramento County, California, 2005*

Productivity loss	Value of work day missed†	Value of nonwork day missed‡	No. work days missed	No. nonwork days missed	No. patients <60	No. patients ≥60	% Cases	Total costs for all cases
For patients <60 y	$191	$125	50	10	31		100	$334,800
For patients ≥60 y		$125		60		15	100	$112,500
For caretakers		$125	25		8	4	26	$37,500
Total costs								$484,800

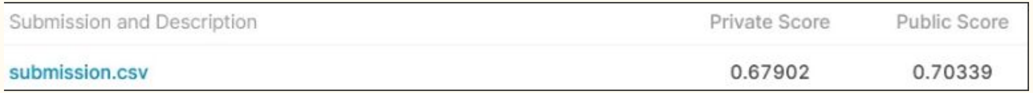
Conclusion

Business Recommendations

- Conduct aerial spraying
- Social Education
- Birds/Pests Monitoring

Further Exploration

- Hyperparameter tuning
- Time lag weather data
- Poisson Regression modeling
- Post spray effectiveness



Submission and Description	Private Score	Public Score
submission.csv	0.67902	0.70339

Thank You