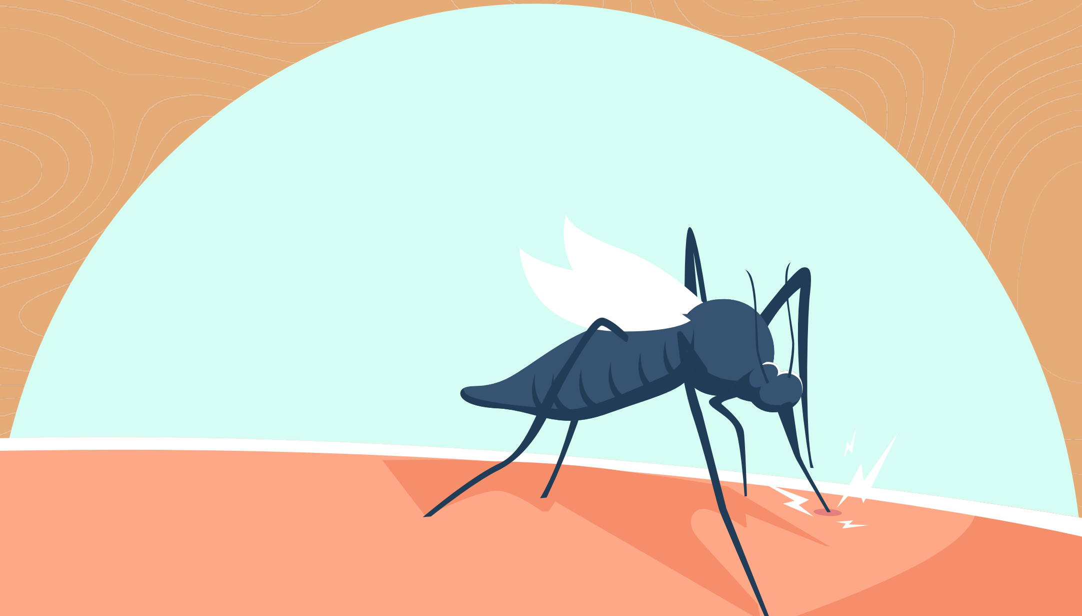
PREDICTING THE PRESENCE OF WEST NILE VIRUS

CHICAGO

Report generated for use by the Centre of Disease Control and Prevention (CDC)



By: Hui Min, Shi Min, Saloni, Haziq, Shu Kai

TABLE OF CONTENTS

01

BACKGROUND & PROBLEM STATEMENT

Saloni

02

DATA CLEANING & EDA

Hui Min

03

MODEL SELECTION & EVALUATION

Haziq

04

COST-BENEFIT ANALYSIS

Shi Min

05

CONCLUSION & RECOMMENDATIONS

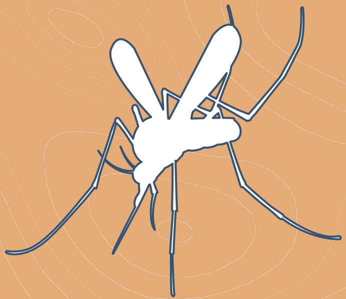
Shu Kai

01

BACKGROUND & PROBLEM STATEMENT

ABOUT THE WEST NILE VIRUS

CAUSE



EFFECTS

- PERSISTENT FEVER
- SERIOUS NEUROLOGICAL ILLNESS
- DEATH

ABOUT THE WEST NILE VIRUS

DETECTION & TREATMENT

ANTIBODY SERUM TEST

NO VACCINE OR SPECIFIC MEDICINE AVAILABLE

MEASURES

2002

First human case detected in chicago

2004

City of Chicago and CDPH establishes comprehensive surveillance and control program

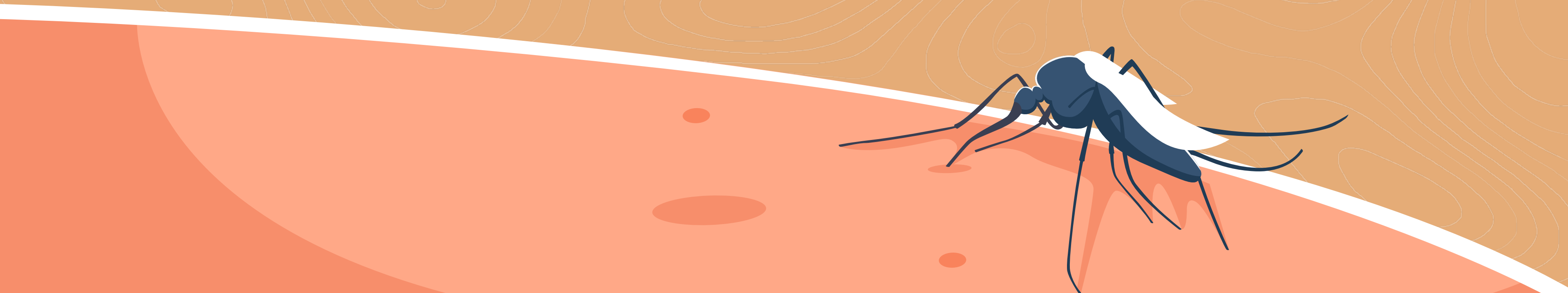
PRESENT

Mosquito traps are laid every week in late spring across the city

PROBLEM STATEMENT

A more accurate method of predicting outbreaks of West Nile virus is required so that the City of Chicago and CPHD can allocate resources efficiently and effectively towards preventing transmission of this potentially deadly virus.

As such, given weather, trap and spray data, we intend to predict when and where different species of mosquitoes will test positive for West Nile virus.



02 DATA CLEANING & EDA

DATASETS USED

	2007	2008	2009	2010	2011	2012	2013
MAIN DATASET (TRAIN) Records the location of mosquito traps, number and species of mosquitoes caught, and whether West Nile Virus is present in the mosquitoes	May to Oct		May to Oct		Jun to Sep		Jun to Sep
WEATHER DATASET Records the weather condition in Chicago from 2 weather stations	May to Sep	May to Sep	May to Sep	May to Sep	May to Sep	May to Sep	May to Sep
SPRAY DATASET Records the date, time and location where the pesticides are sprayed					Aug to Sep		Jul to Sep

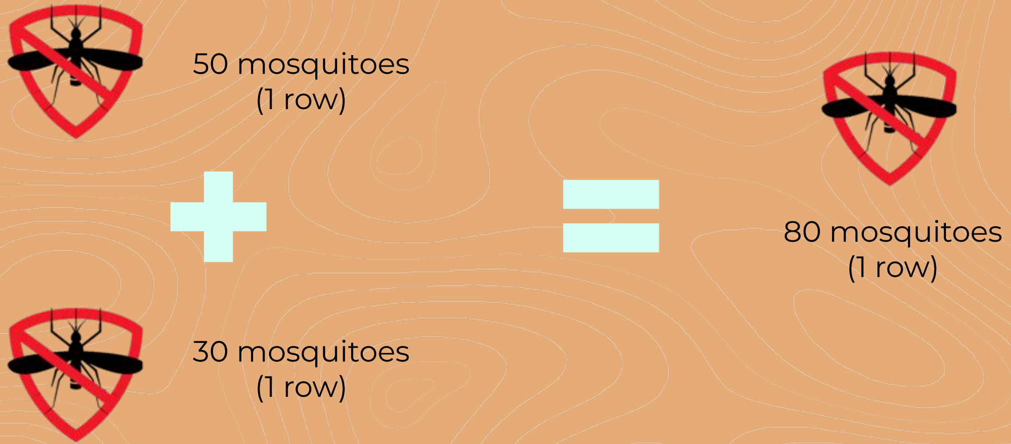
DATA CLEANING

MAIN DATASET (TRAIN)

WEATHER DATASET

SPRAY DATASET

- Dropped 'Address', 'Block', 'Street', 'AddressNumberAndStreet' and 'AddressAccuracy' columns.
- Traps that captured > 50 mosquitoes for any day were split into multiple rows. Such records were combined to form a single record.



50 mosquitoes
(1 row)

+

30 mosquitoes
(1 row)

=

80 mosquitoes
(1 row)

DATA CLEANING

MAIN DATASET (TRAIN)

WEATHER DATASET

SPRAY DATASET

- Dropped ‘Time’ column.
- Dropped exact duplicate records.

DATA CLEANING

MAIN DATASET (TRAIN)

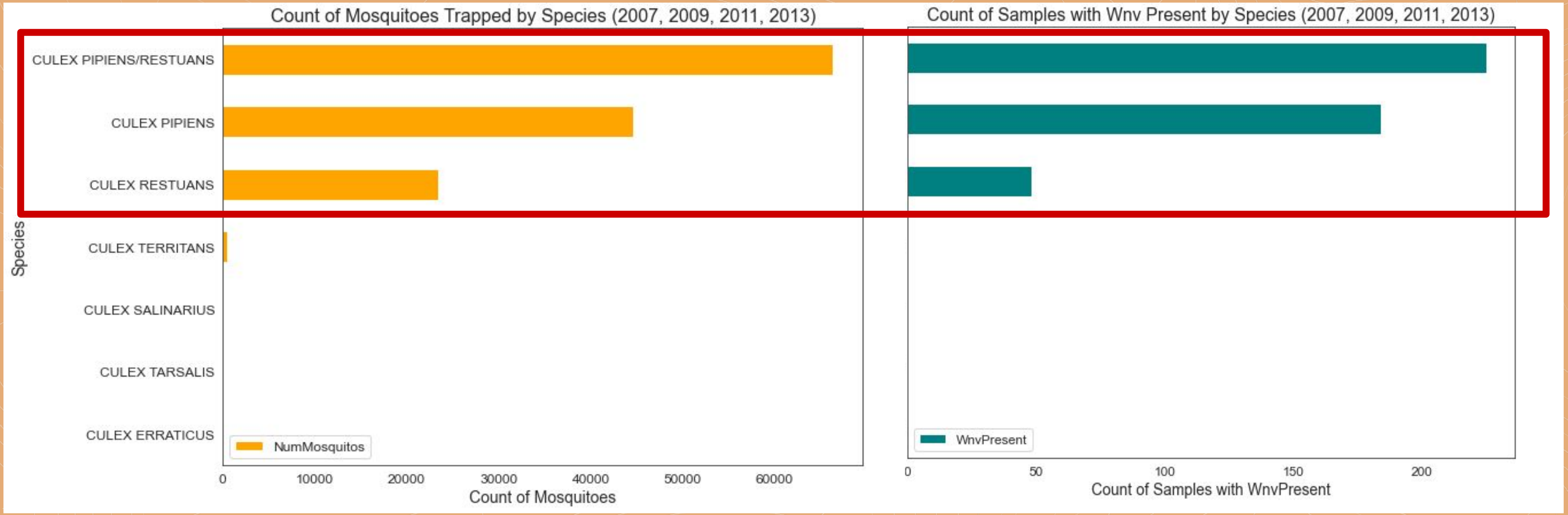
WEATHER DATASET

SPRAY DATASET

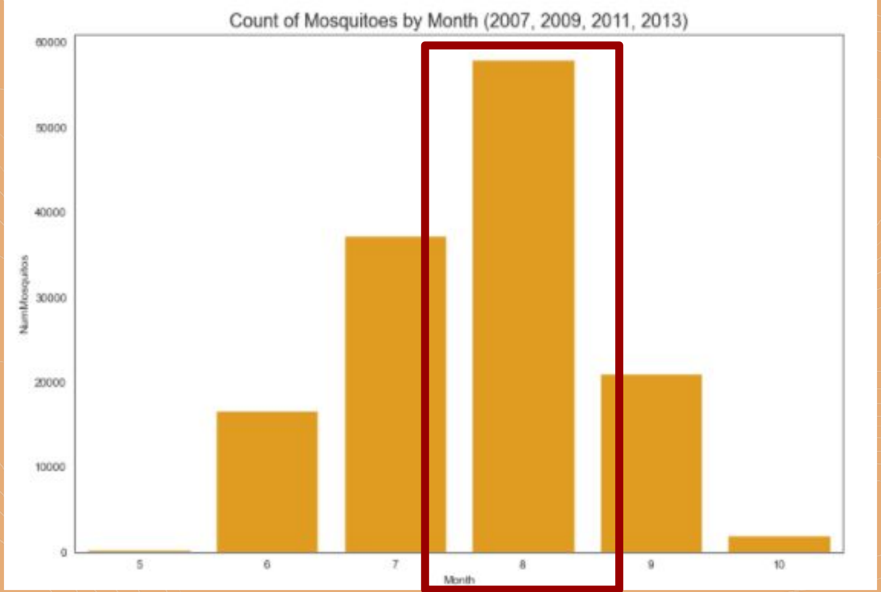
- Dropped ‘CodeSum’, ‘SnowFall’, ‘Depth’, ‘Water1’ and ‘Depart’.
- Imputed null values for ‘Tavg’ with the average of ‘Tmax’ and ‘Tmin’.
- Replaced ‘T’ in ‘PrecipTotal’ with ‘0’.
- Applied forward filling method for null values in ‘Cool’, ‘Heat’, ‘SeaLevel’, ‘WetBulb’, ‘StnPressure’, ‘AvgSpeed’ and ‘PrecipTotal’ (i.e. previous day's readings in the respective Stations, as there is likely a high autocorrelation).
- Imputed ‘Sunrise’ and ‘Sunset’ timings for Station 2 using Station 1's values.
- Computed the distance of each trap to Station 1 and 2, and assigned weather information of the nearest station to each trap record.

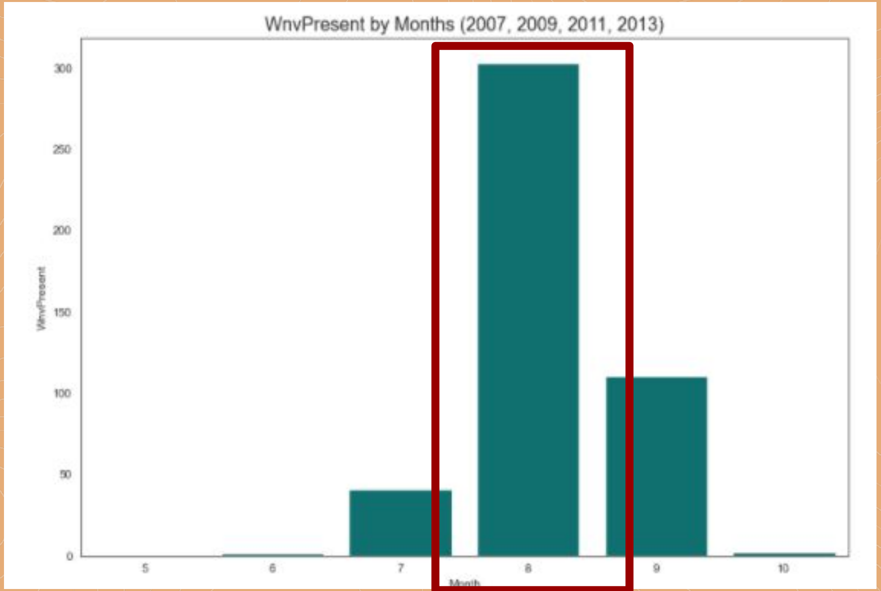
CULEX PIPIENS AND CULEX RESTUANS ARE THE 2 DOMINANT MOSQUITO SPECIES IN CHICAGO

THEY ARE THE ONLY SPECIES FOUND TO BE CARRYING WEST NILE VIRUS

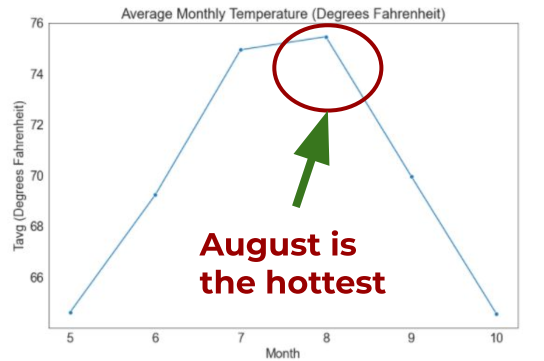


AUGUST APPEAR TO HAVE THE HIGHEST COUNT OF MOSQUITOES AND PRESENCE OF WEST NILE VIRUS...



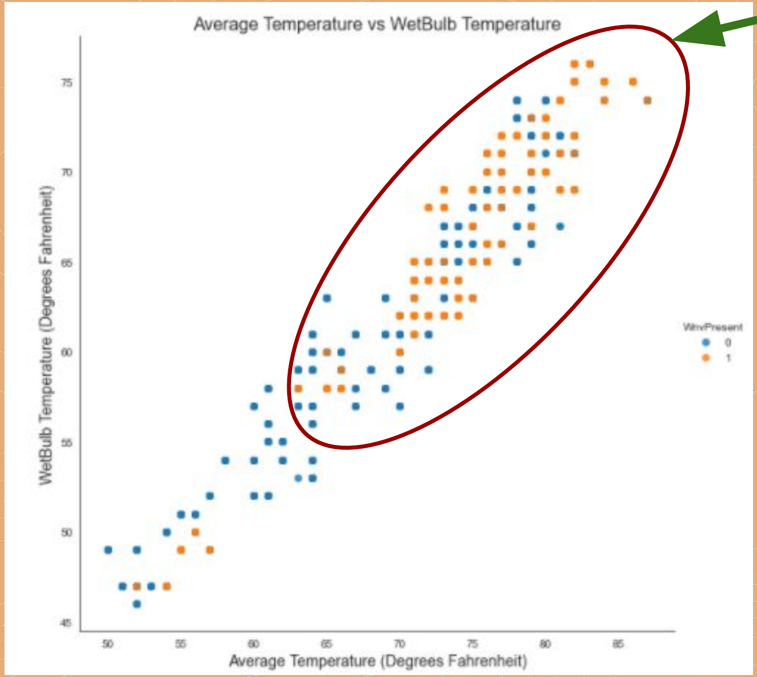


COULD IT BE DUE TO CLIMATE CONDITIONS IN AUGUST?



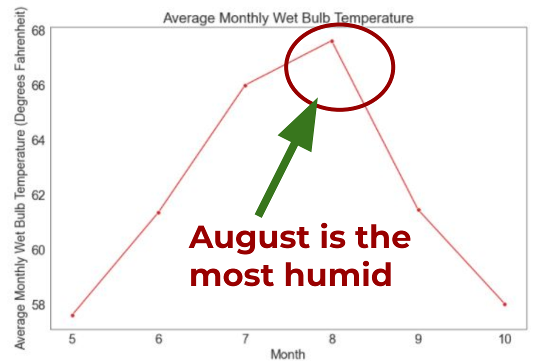
August is the hottest

HOT



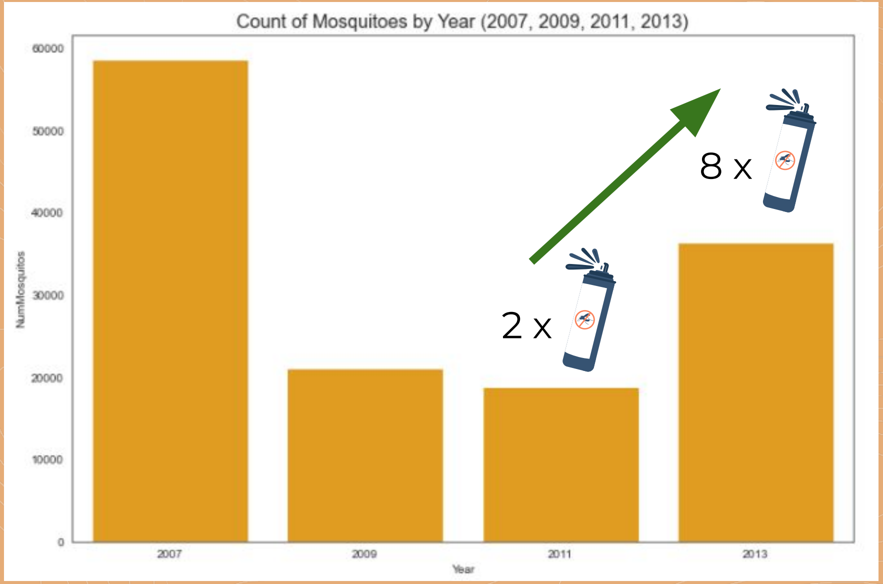
Mosquitoes that carry WNV appear to cluster at higher temperature and humidity range

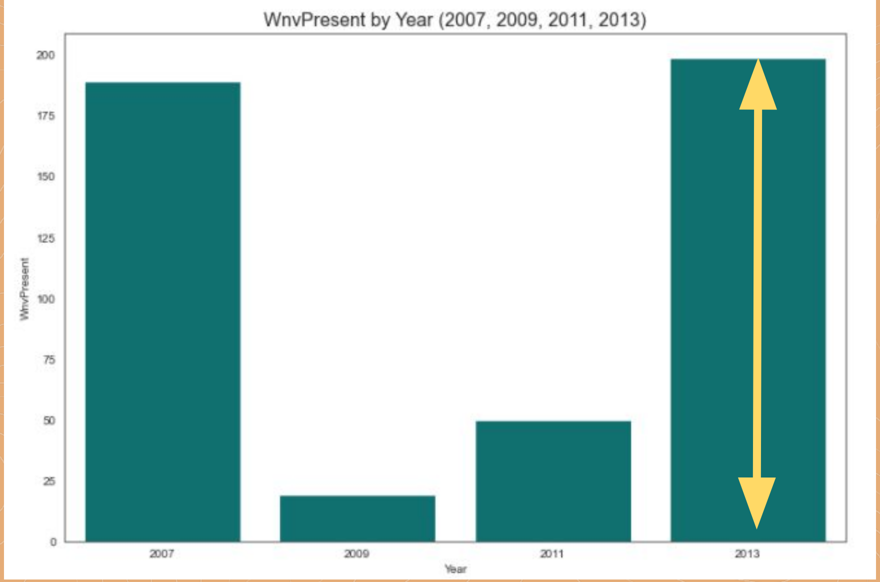
HUMID



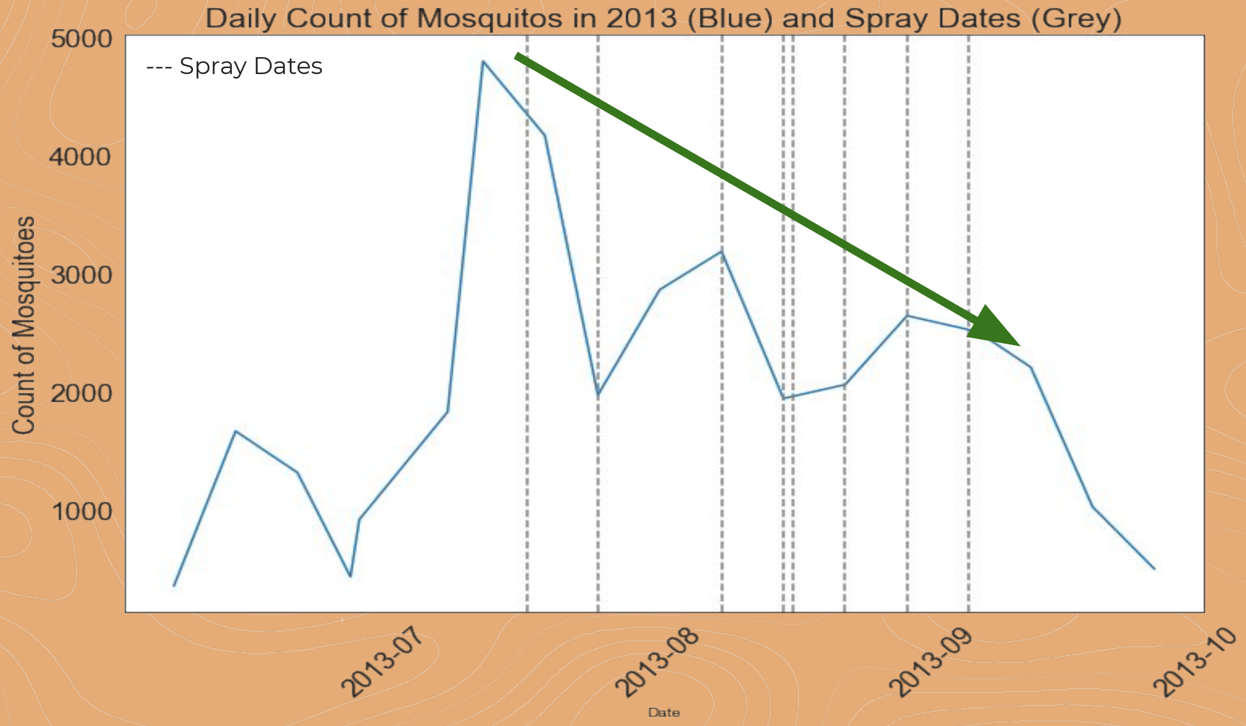
August is the most humid

MOSQUITO COUNT AND PRESENCE OF WEST NILE VIRUS IN 2013 WERE STILL HIGH DESPITE MORE FREQUENT SPRAYING ATTEMPTS...



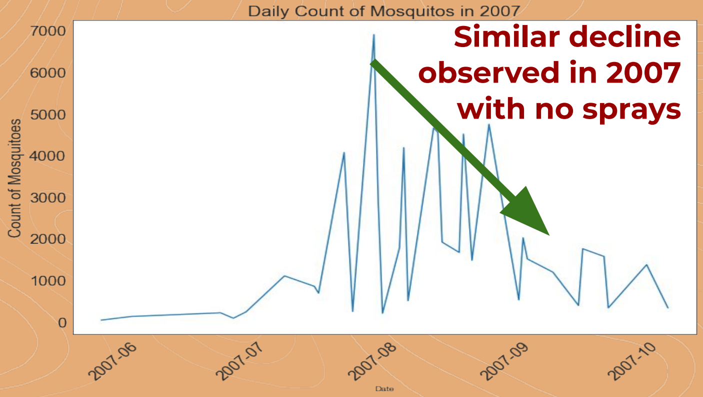


SPRAYS IN 2013 APPEAR TO REDUCE MOSQUITO COUNT, BUT...



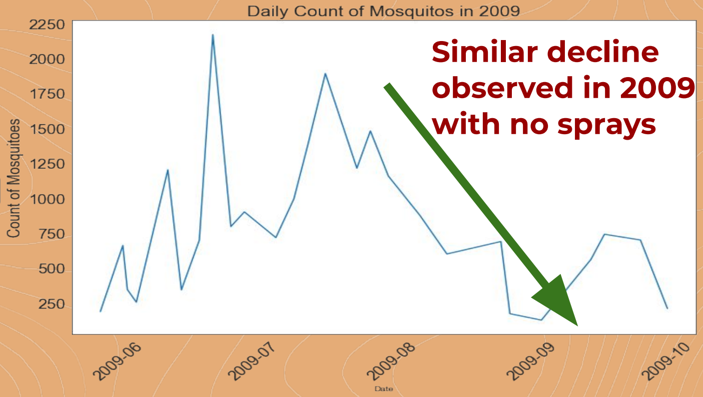
...WAS IT REALLY DUE TO THE SPRAY?

2007
(NO SPRAY)


Daily Count of Mosquitos in 2007

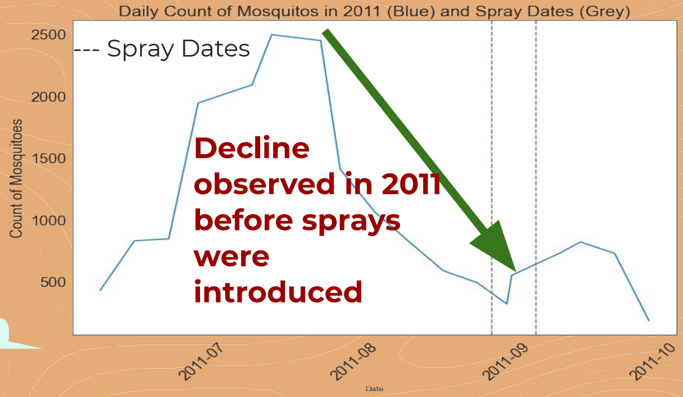
Similar decline observed in 2007 with no sprays

2009
(NO SPRAY)


Daily Count of Mosquitos in 2009

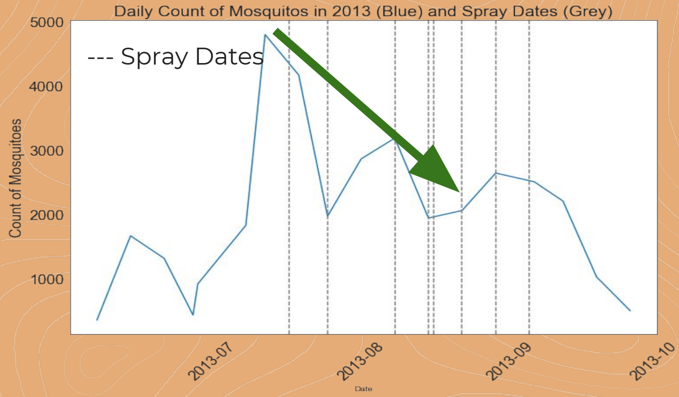
Similar decline observed in 2009 with no sprays

2011
(SPRAY)

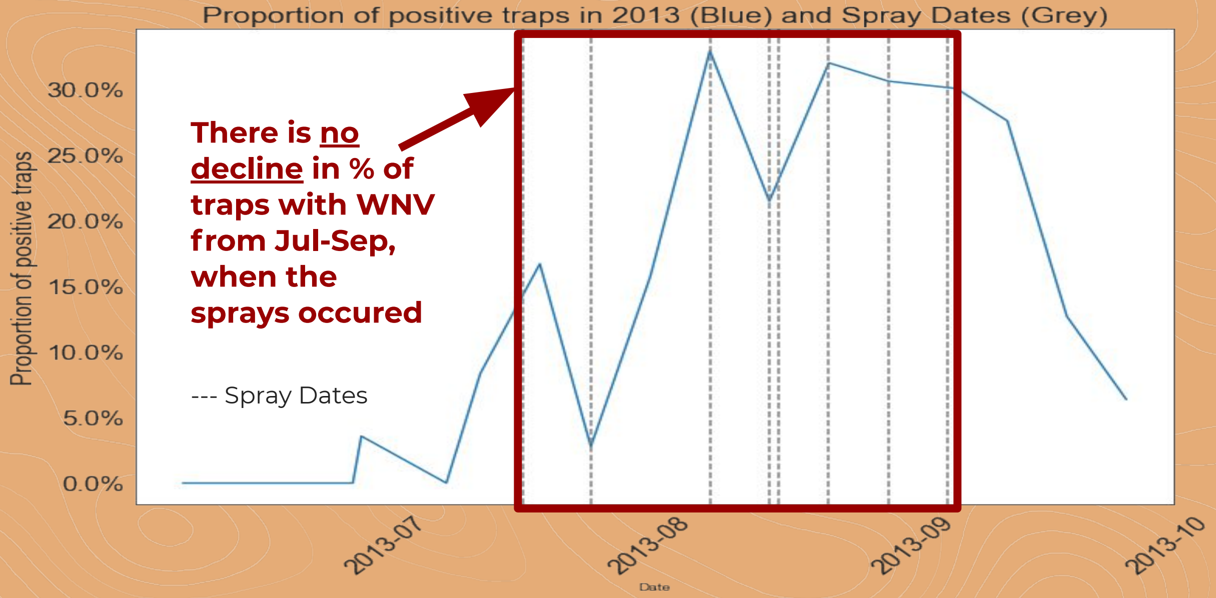

Daily Count of Mosquitos in 2011 (Blue) and Spray Dates (Grey)

Decline observed in 2011 before sprays were introduced

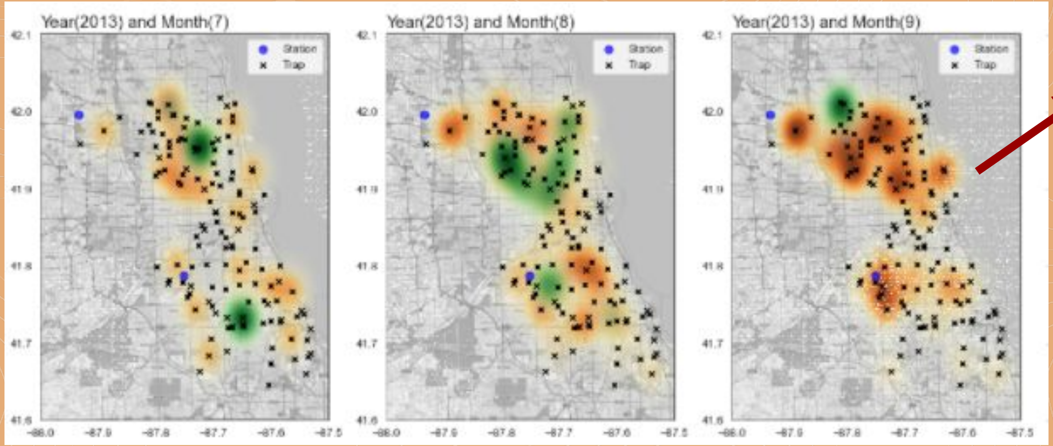
2013
(SPRAY)


Daily Count of Mosquitos in 2013 (Blue) and Spray Dates (Grey)

LIMITED EFFECT OF SPRAY ON WEST NILE VIRUS PRESENCE IN 2013



There is **no decline** in % of traps with WNV from Jul-Sep, when the sprays occured



Higher WNV presence after spraying

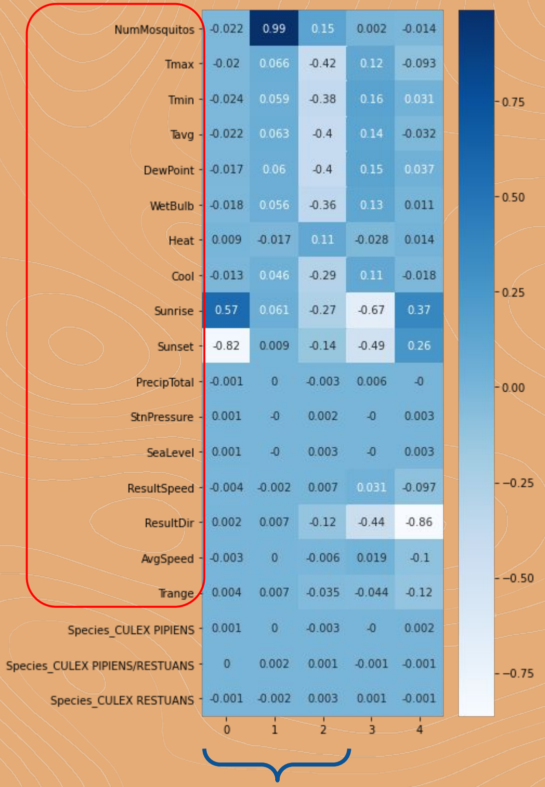
Legend

- ● Stations
- ● Presence & estimated density of WNV
- ● Presence & estimated density of spray
- ✖ Trap

03

MODEL SELECTION & EVALUATION

FEATURE SELECTION (PRINCIPAL COMPONENT ANALYSIS)



	0	1	2	3	4
NumMosquitos	-0.022	0.99	0.15	0.002	-0.014
Tmax	-0.02	0.066	-0.42	0.12	-0.093
Tmin	-0.024	0.059	-0.38	0.16	0.031
Tavg	-0.022	0.063	-0.4	0.14	-0.032
DewPoint	-0.017	0.06	-0.4	0.15	0.037
WetBulb	-0.018	0.056	-0.36	0.13	0.011
Heat	0.009	-0.017	0.11	-0.028	0.014
Cool	-0.013	0.046	-0.29	0.11	-0.018
Sunrise	0.57	0.061	-0.27	-0.67	0.37
Sunset	-0.82	0.009	-0.14	-0.49	0.26
PrecipTotal	-0.001	0	-0.003	0.006	-0
StnPressure	0.001	-0	0.002	-0	0.003
SeaLevel	0.001	-0	0.003	-0	0.003
ResultSpeed	-0.004	-0.002	0.007	0.031	-0.097
ResultDir	0.002	0.007	-0.12	-0.44	-0.86
AvgSpeed	-0.003	0	-0.006	0.019	-0.1
Trange	0.004	0.007	-0.035	-0.044	-0.12
Species_CULEX PIPIENS	0.001	0	-0.003	-0	0.002
Species_CULEX PIPIENS/RESTUANS	0	0.002	0.001	-0.001	-0.001
Species_CULEX RESTUANS	-0.001	-0.002	0.003	0.001	-0.001

First 3 components (96.6%)

- Dropped correlated features such as Temperature-related.
- Dropped minimal impact features such as Year.
- Dummified categorical features.

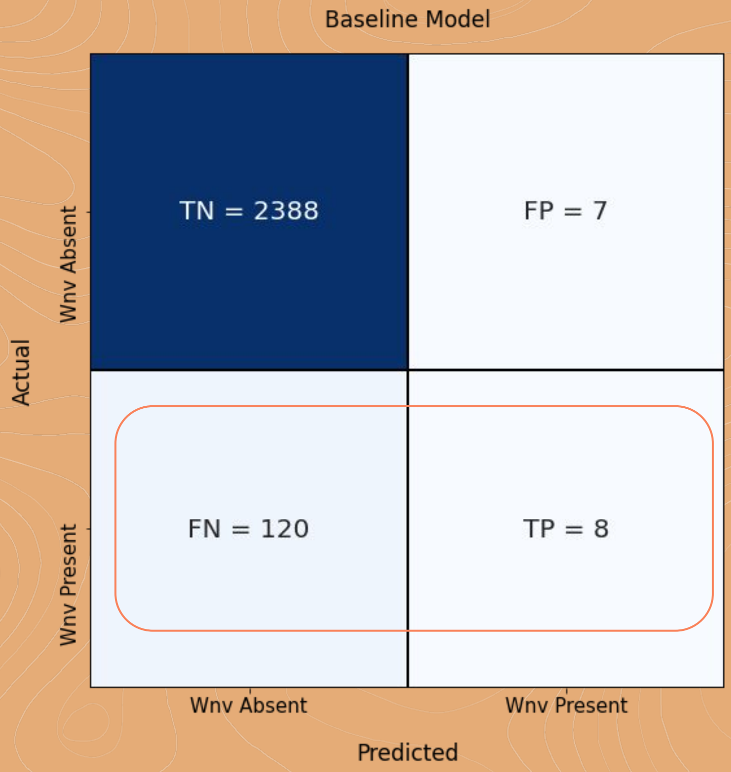
BASELINE MODEL

- **Logistic Regression**
- **Accuracy = 95%**

WHY HIGH ACCURACY, YET POOR PREDICTION OF WNV?

- Imbalanced dataset
- Test Set:
  - 2395 rows Wnv Absent
  - 128 rows Wnv Present

120 cases with WnvPresent undetected



Baseline Model

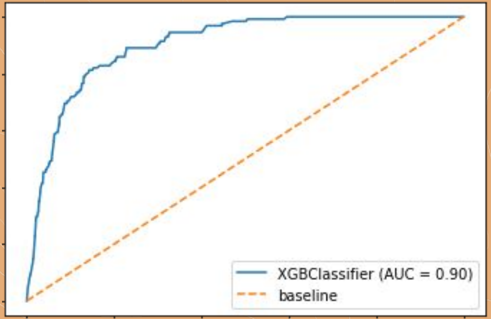
Actual

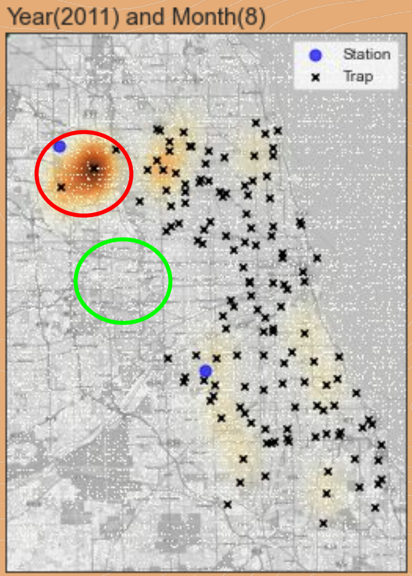
	Wnv Absent	Wnv Present
Wnv Absent	TN = 2388	FP = 7
Wnv Present	FN = 120	TP = 8

Predicted

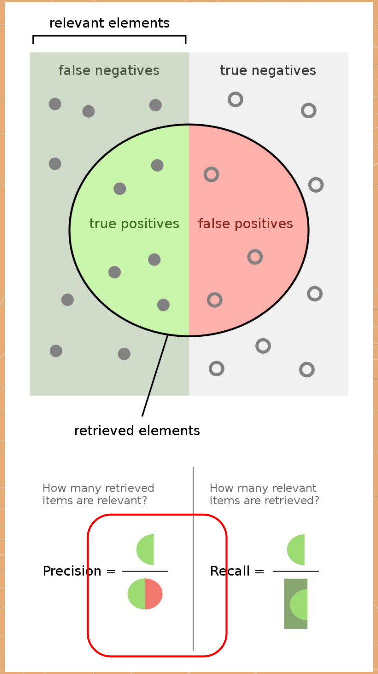
METRIC SELECTION

1. AUC





2. RECALL



MODEL SELECTION

- **Models considered**
  - Logistic Regression, SVC, AdaBoost, GradientBoost, XGBoost

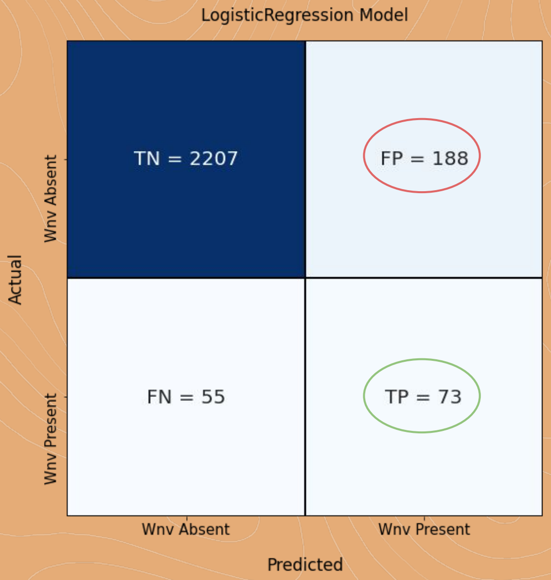
- **Pipeline:**
  - Resampling (SMOTE) ⇒ Scaling (StandardScaler) ⇒ Classifier Model

- **GridsearchCV**
  - Used for hyperparameter tuning

WHAT MAKES A GOOD MODEL?

1. **AUC** : The higher the better at classifying between classes
2. **Recall** : The higher the better (percentage of WnvPresent predicted correctly)

MODEL COMPARISON

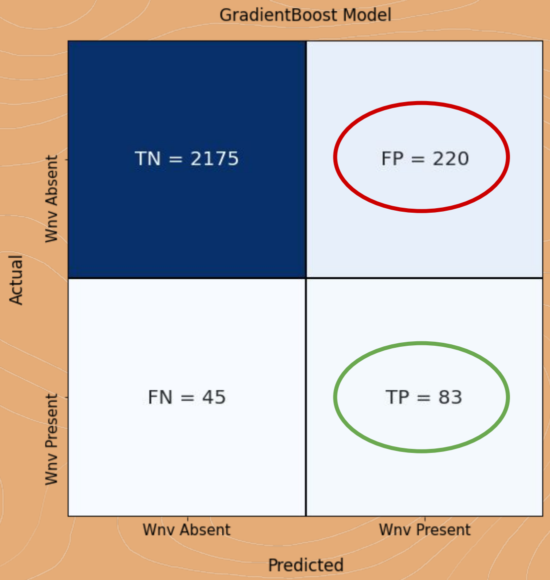

LogisticRegression Model

Actual \ Predicted	Wnv Absent	Wnv Present
Wnv Absent	TN = 2207	FP = 188
Wnv Present	FN = 55	TP = 73

AUC = 0.74

Recall = 0.57

Accuracy = 0.90

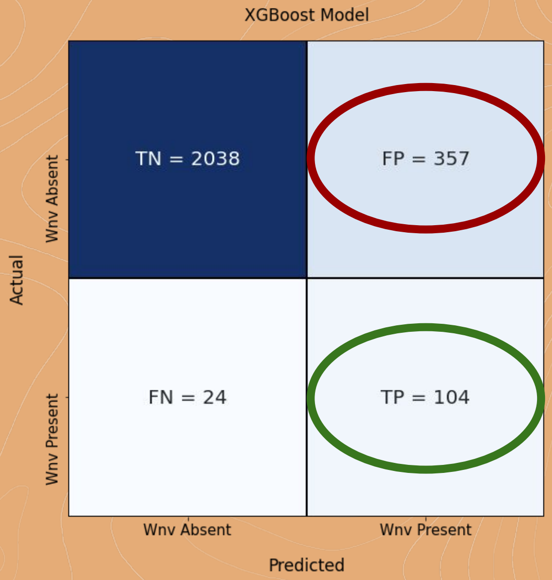

GradientBoost Model

Actual \ Predicted	Wnv Absent	Wnv Present
Wnv Absent	TN = 2175	FP = 220
Wnv Present	FN = 45	TP = 83

AUC = 0.77 ↑

Recall = 0.65 ↑

Accuracy = 0.89 ↓


XGBoost Model

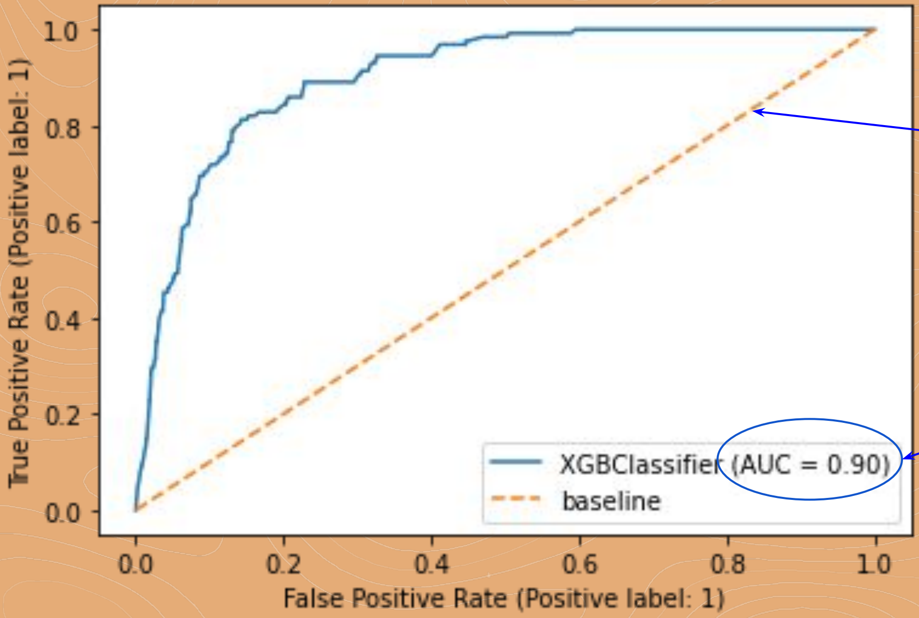
Actual \ Predicted	Wnv Absent	Wnv Present
Wnv Absent	TN = 2038	FP = 357
Wnv Present	FN = 24	TP = 104

AUC = 0.83 ↑

Recall = 0.81 ↑

Accuracy = 0.85 ↓

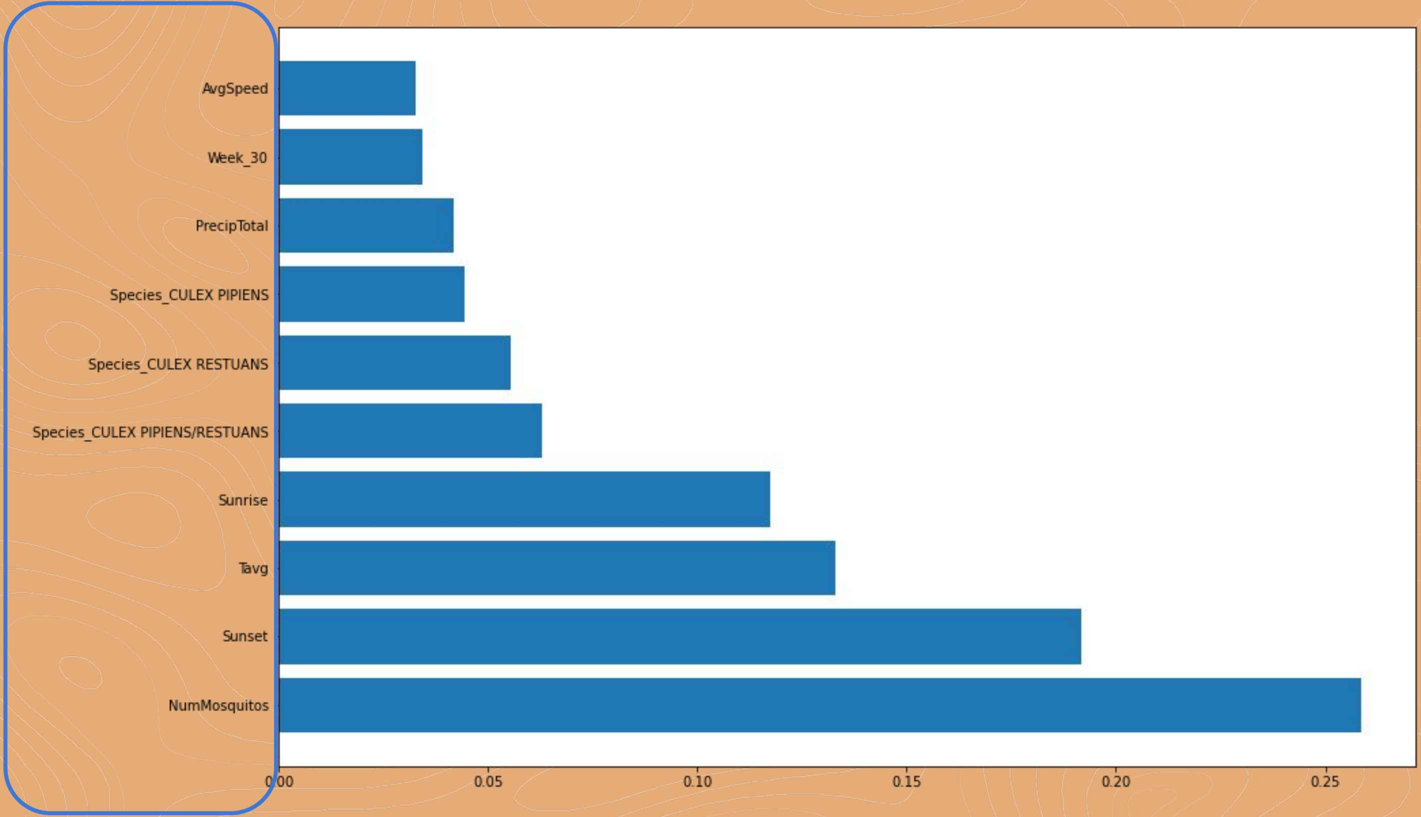
ROC-AUC CURVE



Unable to distinguish

90% chance of distinguishing WnV = 0 / Wnv = 1

FEATURE IMPORTANCE



04

COST-BENEFIT ANALYSIS

WHAT ARE THE COSTS AND BENEFITS?

WE NEED TO WEIGH THE COSTS AND BENEFITS

COSTS

- ECOLOGICAL COSTS
- COST OF SURVEILLANCE
- COST OF SPRAY

BENEFITS

- REDUCE SUFFERING
- PREVENT LOSS OF INCOME
- AVOID MEDICAL COSTS

BEFORE WE START THE ESTIMATION PROPER...

- Due to the lack of updated, Chicago-specific data, we've used previous studies on the West Nile Virus in California. As they are quite dated, we have made inflation adjustments where necessary.

- While more accurate cost estimates can be obtained once we have access to better data, this analysis should give a good enough sense of whether the spraying effort, supported by our model, is worth the cost

COST ESTIMATION

ECOLOGICAL COSTS

SPRAYS HAVE BEEN KNOWN TO KILL OTHER HARMLESS ANIMALS

COST OF SPRAY

$245/KM^2[2]

COST OF SURVEILLANCE

$82/TRAP[3]

(PER WEEKLY SURVEILLANCE)

1. In 2005, aerial spray was utilised in Sacremento County, 6 times over an area of 477km2. The total cost was USD700k, including labour costs (Source: Economic Cost Analysis of West Nile Virus Outbreak, Sacramento County, California, USA, 2005). While this is arguably a more expensive method vs truck spraying, it was also 16 years ago. With inflation considered, we think this can be a good proxy
2. Cost of surveilling and testing each mosquito trap from 2004-2012 in California was estimated to be $72 (Source: US National Library of Medicine) . Assuming 1.5% inflation/year from 2012-2021, the estimated cost in 2021 terms would be $82

TOTAL ESTIMATED ANNUAL COST OF OUR PROPOSAL

ANNUAL COST OF SURVEILLANCE

PROPOSE WEEKLY SURVEILLANCE FROM JUN-SEP, WITH EXISTING 80 TRAPS

$82 X 80 TRAPS X 16 = $105,000

ANNUAL COST OF SPRAY

__MODEL RECALL SCORE:__

81% FOR POSITIVE,

85% FOR NEGATIVE

__ACTUAL % OF TRAPS POSITIVE:__

~30% IN JUL-AUG 2013, TO BE USED AS AN ESTIMATE

__% OF TRAPS THAT MODEL WILL IDENTIFY AS POSITIVE IN JUL-AUG:__

81% * 30% + 15% * 70% = 35%

__COST OF PROPOSED WEEKLY SPRAY FROM JUL-AUG__

35% X 606KM^2 X $245 X 8 = $416,000

TOTAL ANNUAL COST: $521,000

BENEFITS ESTIMATION

REDUCE SUFFERING

79% MILD SYMPTOMS[1]

20% ACHES, VOMITING, DIARRHEA, OR RASH. FATIGUE LASTS FOR WEEKS TO MONTHS

1% BRAIN INFLAMMATION. PERMANENT EFFECTS TO THE CENTRAL NERVOUS SYSTEM

PREVENT LOSS OF INCOME

$500

/INFECTED PERSON[2]

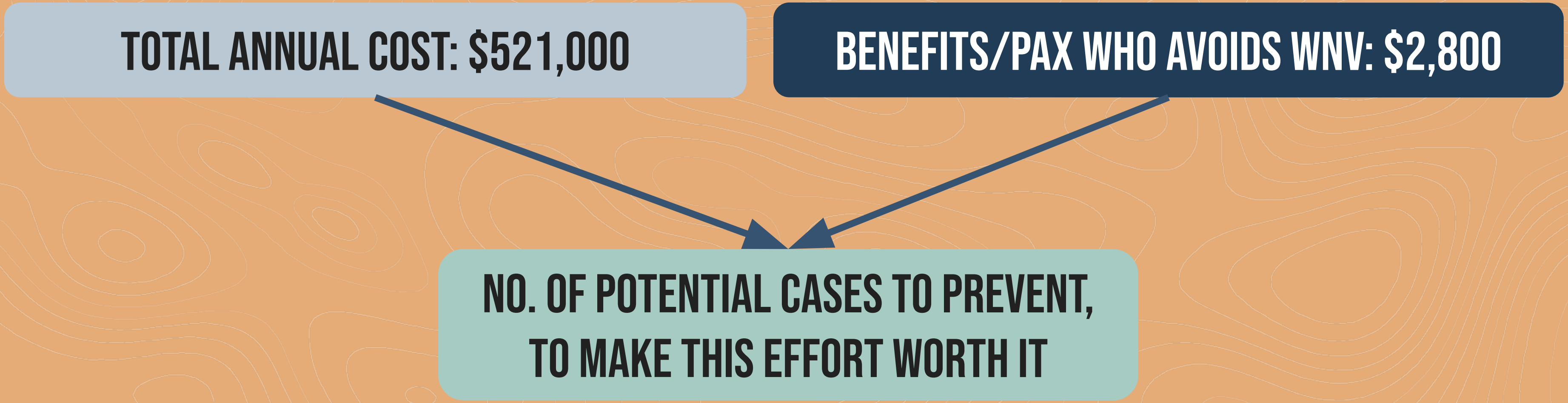
AVOID MEDICAL COSTS

$2,300

/INFECTED PERSON[3]

1. Proportion of infected persons that suffer mild, moderate and severe symptoms are obtained from Centre of Disease Control (Source)
2. Median per capita income in Chicago is $37,100year. (Source: US Census 2019). Assuming 5 days of sick leave on average (same assumption as Source: Economic Cost Analysis of West Nile Virus Outbreak, Sacramento County, California, USA, 2005)
3. In 2005, medical costs for mild and moderate cases (WNF) cost $300 per patient, moderate cases cost $6,317 while serious cases (WNND)cost $33,143. Applying these costs with inflation rate of 1.5% per year, it would be $380, $8016 and $42000 respectively in 2021 terms. Applying to the ratios of mild/moderate : severe, the average medical cost per infected person works out to be $800 (Source: Economic Cost Analysis of West Nile Virus Outbreak, Sacramento County, California, USA, 2005)

WHAT WOULD IT TAKE TO MAKE SPRAYING WORTH IT



TOTAL ANNUAL COST: $521,000

BENEFITS/PAX WHO AVOIDS WNV: $2,800

NO. OF POTENTIAL CASES TO PREVENT, TO MAKE THIS EFFORT WORTH IT

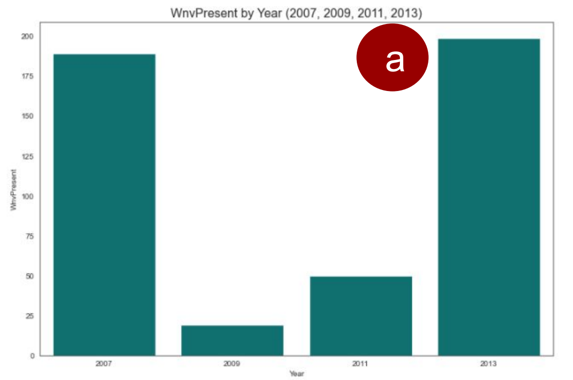
185

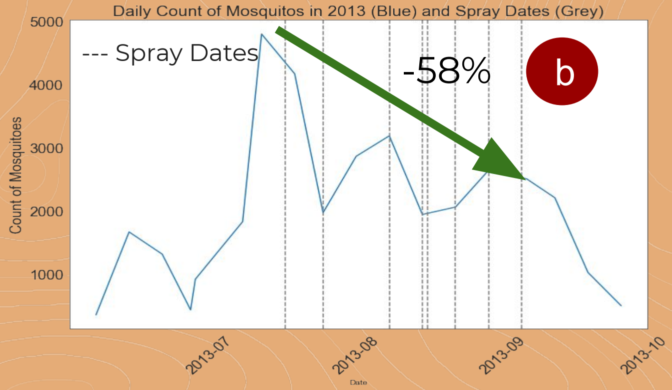
WE CAN PREVENT AT LEAST 185 POTENTIAL CASES

a Let’s use 2013 as an example: About 200 samples in 2013 were tested Wnv positive[1]

b Assume that spraying is effective in reducing mosquito counts by about 50%[2]

- If ~100 of the positive mosquitos can be eliminated by the spraying program, it seems reasonable to assume that at least 185 people can be “saved” from the virus





[1] Number of mosquitos that are positive in each positive sample could actually be more than 1, but we assume just one positive mosquito to be conservative

[2] The daily count of mosquitos appeared to have reduced by about 58% after the sprays in 2013, although it is unclear if the spray is the direct cause of it. Other studies (Source: Journal of Medical Entomology) have also show about 54% of reduction in mosquito count in treated areas vs an increase in untreated areas

05 CONCLUSIONS & RECOMMENDATIONS

CONCLUSION

- XGBoost Classification performed well compared to Gradient Boosting and Logistic Regression.

- Lower accuracy score as compared to Gradient Boost and Logistic Regression models, scored the best on the testing data with the AUC score of 0.83.

- Weather parameters as well as number of mosquitoes caught per trap are very useful in prediction.

RECOMMENDATIONS

WHY OUR MODEL HELPS TO MAKE SPRAYING WORTH IT...

An indiscriminate spraying over all locations over the jul-aug period will increase cost by ~3x vs using our model to guide on where to spray

555 cases are needed to prevent potential contraction of the virus, to make the indiscriminate spraying worth it.

In, 2002 when the virus first appeared and number of cases were at its peak, chicago only recorded **225 cases.**

...BUT OTHER MEASURES SHOULD BE USED IN CONJUNCTION

- Ramping up education on West Nile Virus and how to prevent mosquito breeding.

- Educating the public on preventing stagnant water, applying insect repellent, wearing the proper attire to reduce the chances of getting bitten by mosquitos.

FUTURE IMPROVEMENTS/OPPORTUNITIES

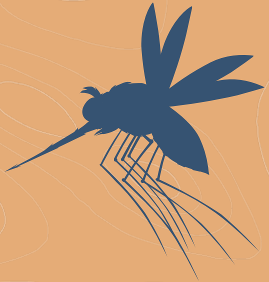
EXPLORE OTHER CLASSIFIER MODELS

- Deep Neural Networks like Keras that may have better prediction results.

BETTER DATASET?

- Annual VS alternate years
- Record number of Mosquitoes carrying WNV.
- Surveillance on birds.

THANKS!



Do you have any questions?

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Stories**

PLEASE KEEP THIS SLIDE FOR ATTRIBUTION