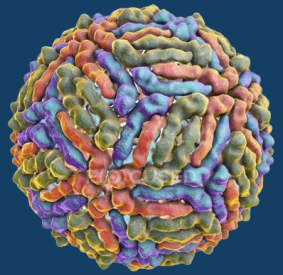
REPORT GENERATED FOR CITY OF CHICAGO

PREDICTING WEST NILE VIRUS

By: Jimmy, Maryam, Ming Jie, Priscilla & Ting Wei

ABOUT: WEST NILE VIRUS



THE VIRUS

- First isolated in West Nile district

THE TRANSMITTER

- Mosquitoes
- Mostly Culex species

THE EFFECTS

- Fever
- Inflammation
- Death

WNV: TREATMENT & CONTROL

DIAGNOSIS
Lab test on spinal fluid

TREATMENT
NO vaccine or medicine

MOSQUITO CONTROL
Trap & Spray

WNV: URGENCY

2001
First detection of WNV in dead birds

2002
635 cases, including 42 deaths

2003
A comprehensive surveillance and control system established

2011
22 human cases, including 1 death

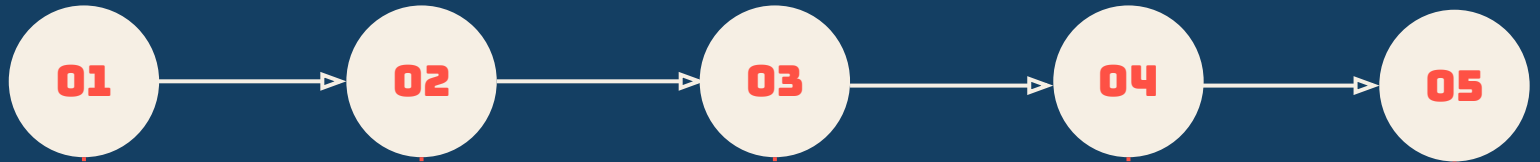
END-2012
174 cases, including 5 deaths

PROBLEM STATEMENT

We need a more accurate system of forecasting the outbreaks of West Nile Virus that will help the City of Chicago and the Chicago Department of Public Health to allocate resources more efficiently and work towards eradicating the viral epidemic.

Thus, we must build an improved model of at least 75% recall to successfully predict when and where WNV- positive mosquitoes will be found in the city.

PROJECT WORKFLOW



- IMPORT DATA
- CLEAN DATA

- VISUALISE DATA
- DRAW ANALYSES

- CREATE MODELS
- SELECT BEST MODEL

- DEPLOY MODEL
- PREDICT WNV-PRESENCE

- COST-BENEFIT ANALYSIS

DATA IMPORT
DATA CLEANING

DATASETS IMPORTED

DATASET NAME	YEAR DATA WAS COLLECTED						
	2007	2008	2009	2010	2011	2012	2013
MAIN (TRAIN)	May-Oct		May-Oct		Jun-Sep		Jun-Sep
SPRAY					Aug-Sep		July-Sep
WEATHER	May-Oct	May-Oct	May-Oct	May-Oct	May-Oct	May-Oct	May-Oct

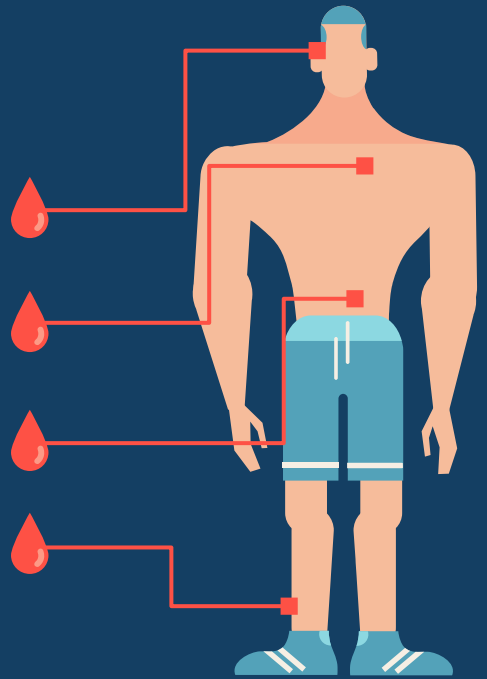
TRAIN DATA CLEANING STEPS

DROP DUPLICATES

CREATE DATE-TIME COLUMNS

DROP COLUMNS; IRRELEVANT

COMBINE RECORDS FOR TRAPS WITH >50 MOSQUITOES IN SAME DAY



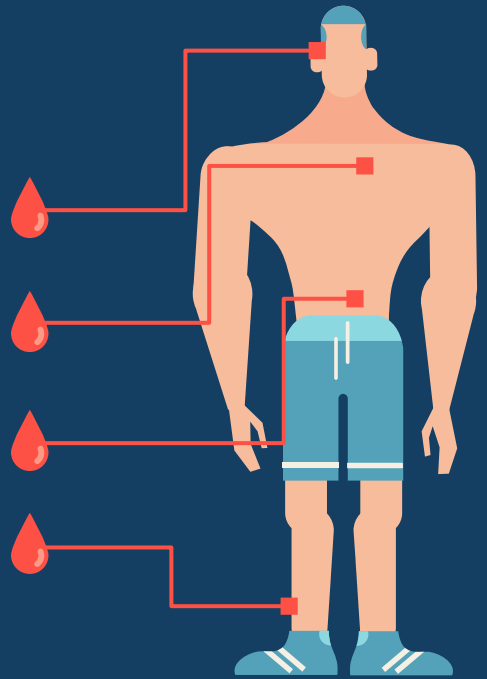
SPRAY DATA CLEANING STEPS

DROP DUPLICATES

CREATE DATE-TIME COLUMNS

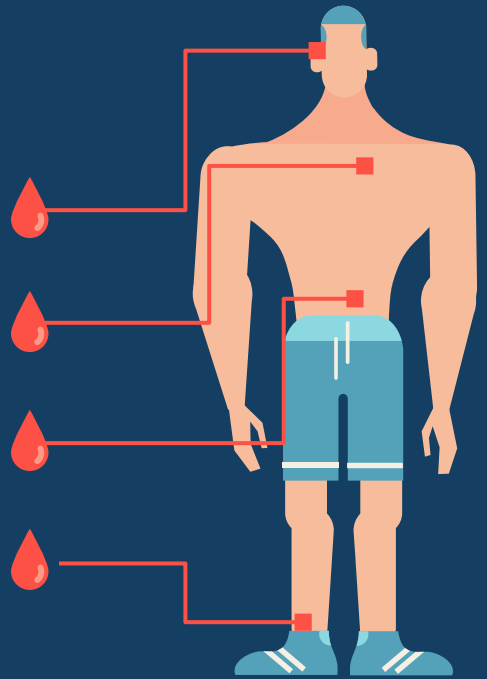
DROP ROWS

IMPUTE VALUES



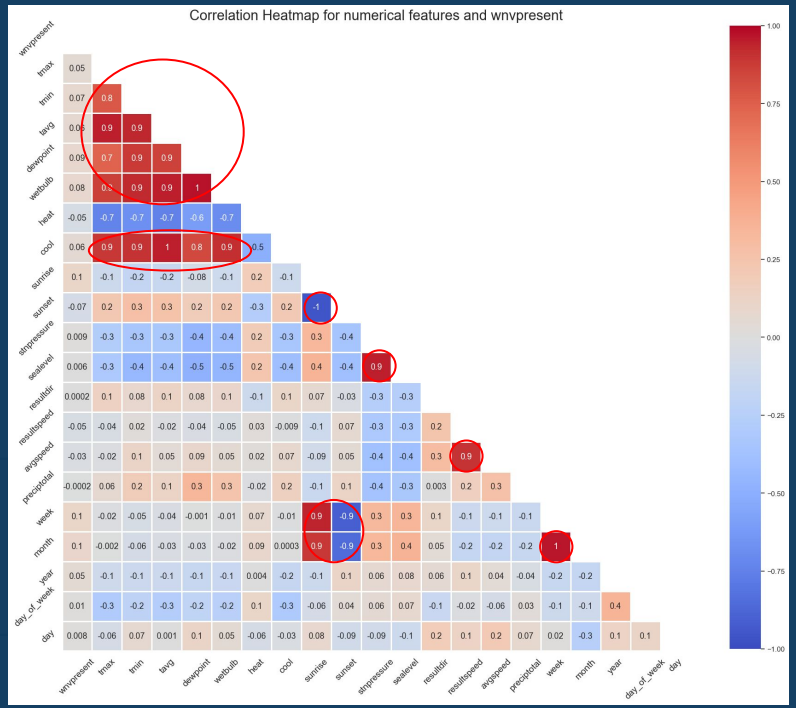
WEATHER DATA CLEANING STEPS

- LOWERCASE COLUMN NAMES
- CONVERT DATES TO DT FORMAT
- DROP ROWS WITH HIGH MISSING VALUES
- IMPUTE VALUES



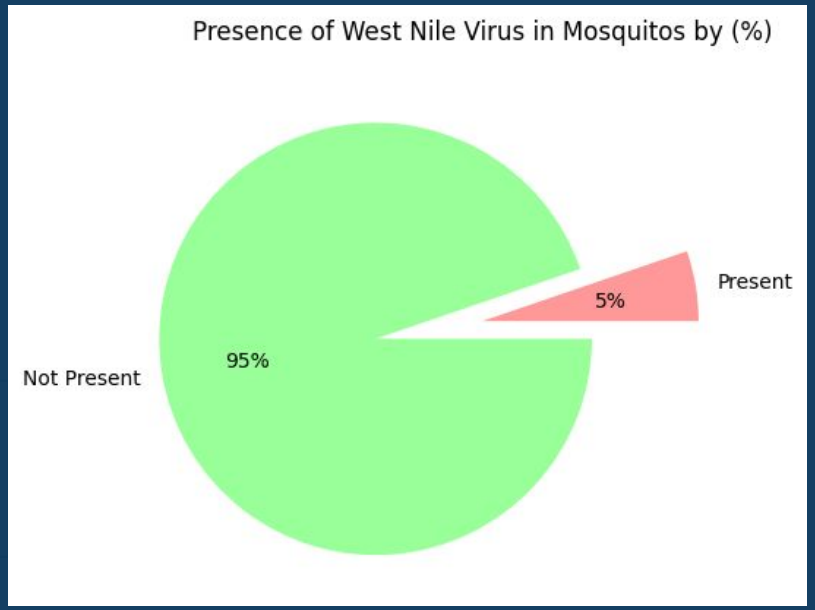
EXPLORATORY DATA ANALYSIS

HEATMAP SUMMARY



- High correlation with wnv-present:
  - Temperature
  - Relative Humidity
  - Precipitation

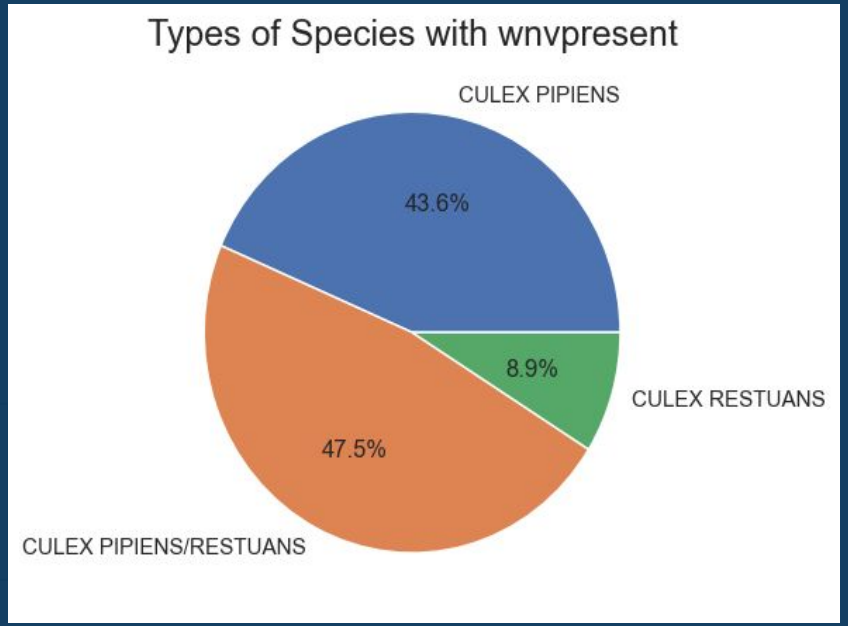
WNV-CARRIER MOSQUITOES



Mosquitoes with WNV:

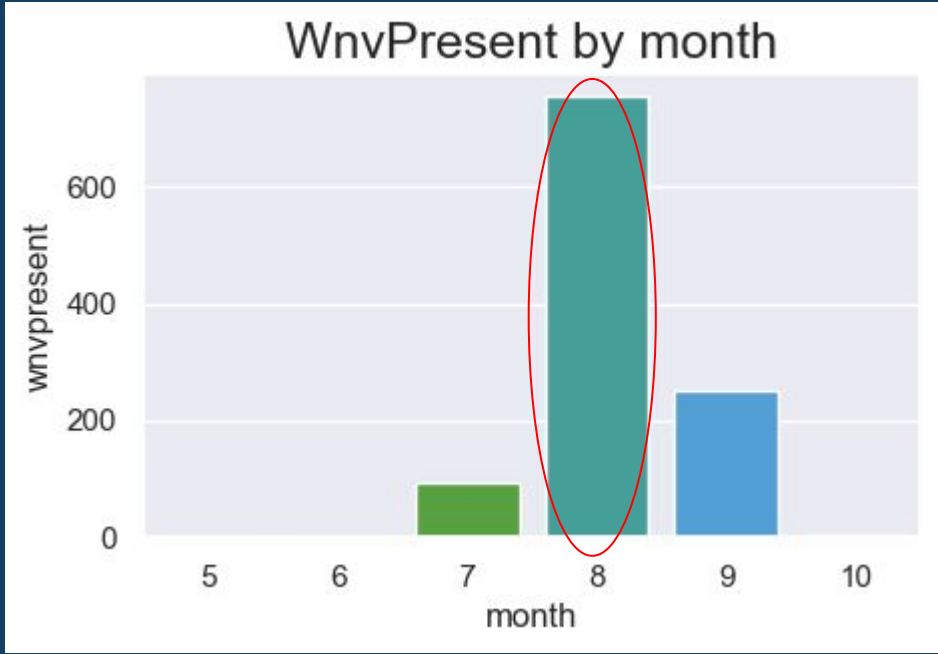
- **5%**
- **1/20** mosquitoes found are WNV-carriers

WNV-CARRIER SPECIES

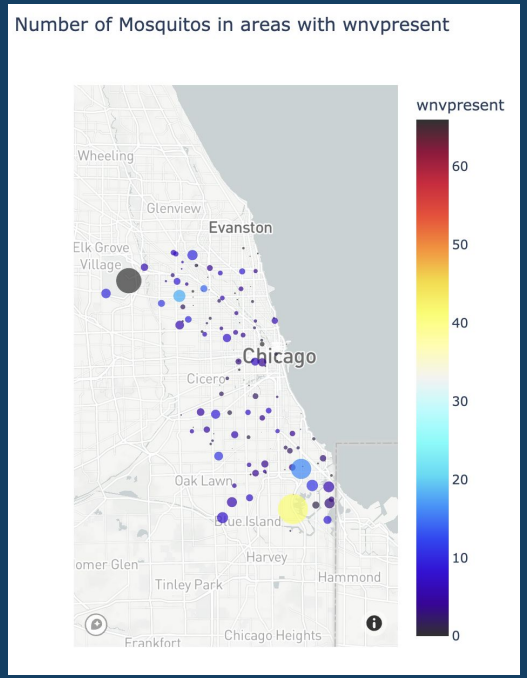


- Species with WNV:
  - Culex Pipiens
  - Culex Restuans

AUGUST HAS HIGHEST RATE OF WNV

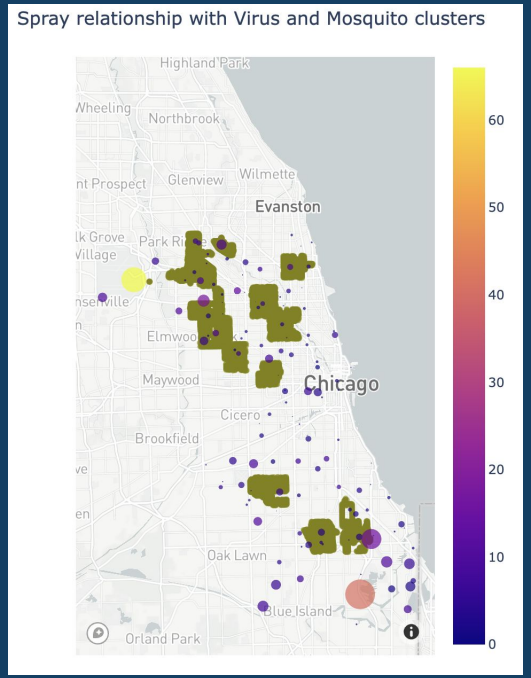


AREAS WITH HIGHEST WNV



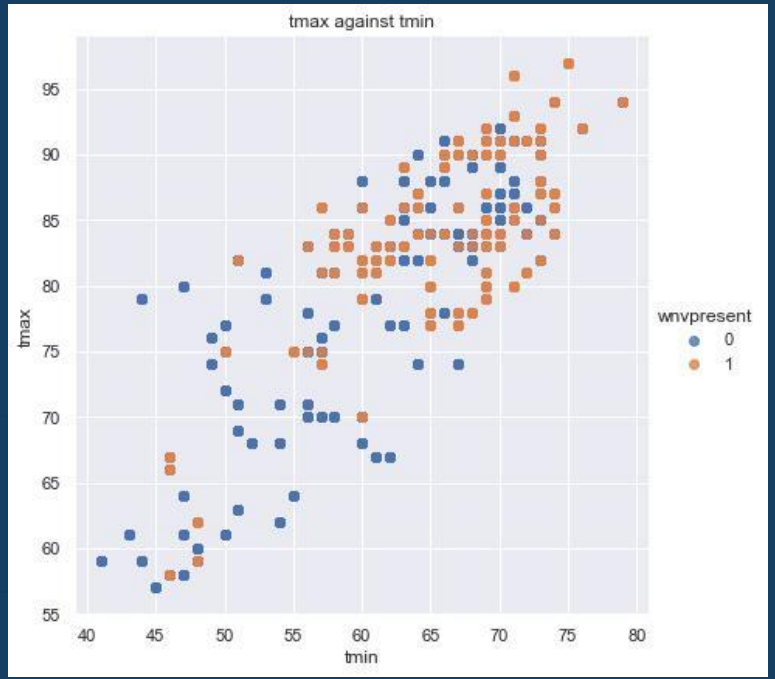
- Airport - 66 cases
- Port of Chicago - 44 cases

SPRAY EFFECTIVENESS



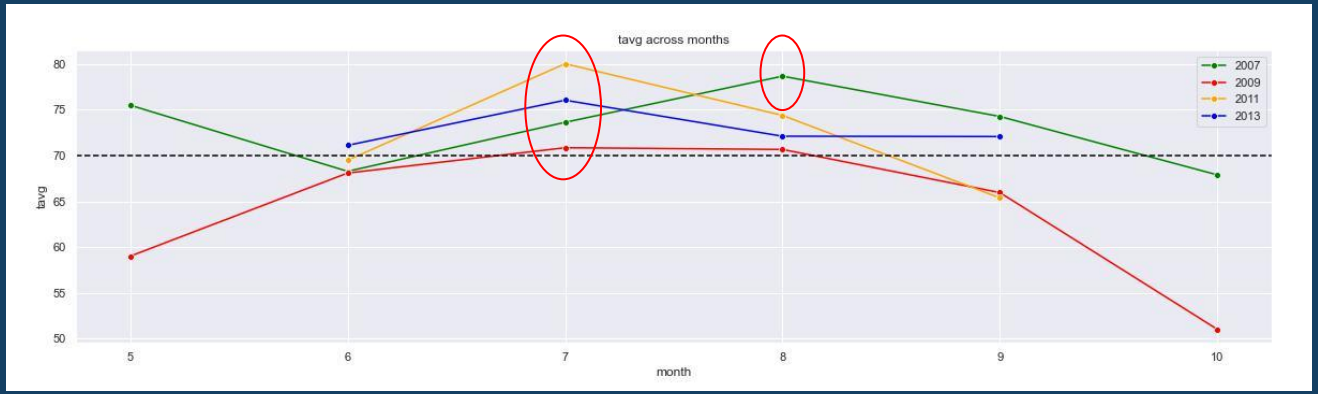
- Spray is effective
- Areas not sprayed has larger mosquito clusters & higher WNV counts

TEMPERATURE & WNV



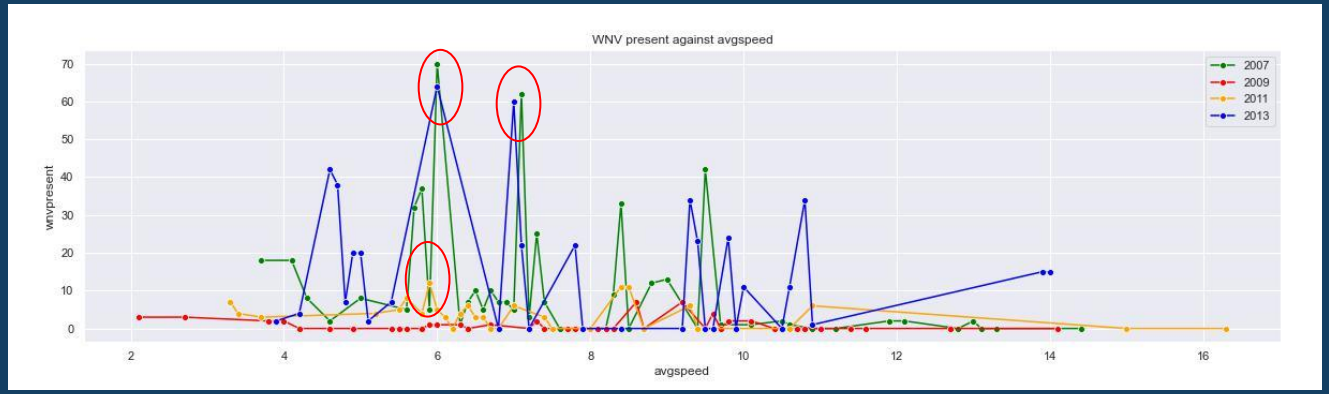
- At **higher** Tmax and Tmin temperatures, **higher** frequency of mosquitoes with wnvpresent = 1

TEMPERATURE & WNV



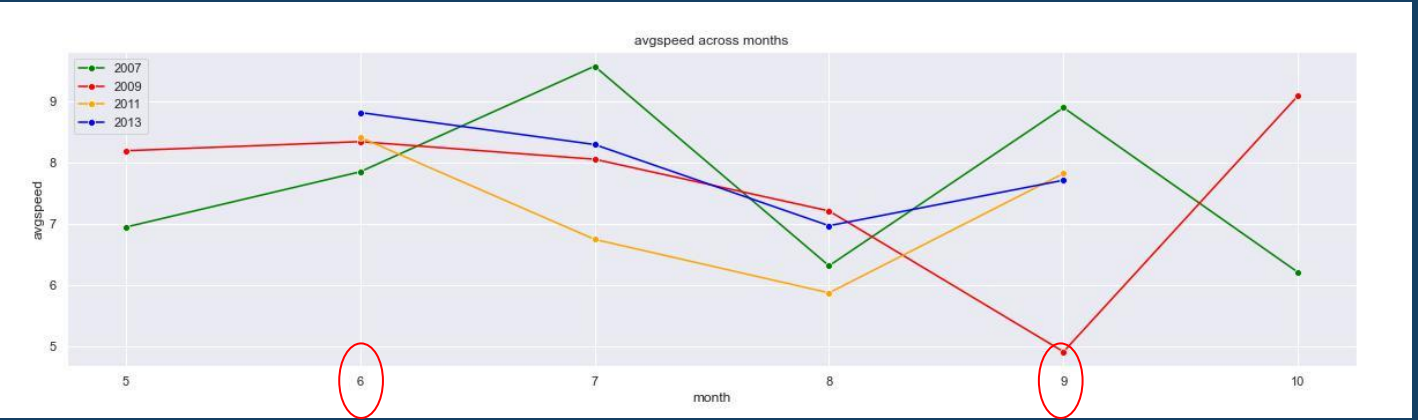
- Tavg peaks above **70°F** (21°C) in July (7) or August (8)

AVERAGE WIND SPEED & WNV



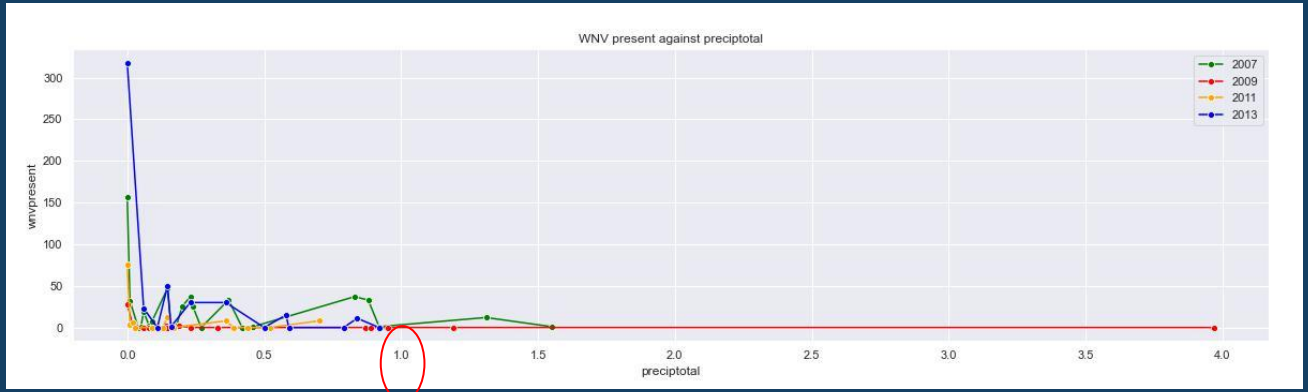
- Peaks at **lower** average wind speed between 6-8mph (light to gentle breeze)
- At lower average wind speed, higher occurrences of mosquitoes with wnvpresent = 1

AVERAGE WIND SPEED & WNV



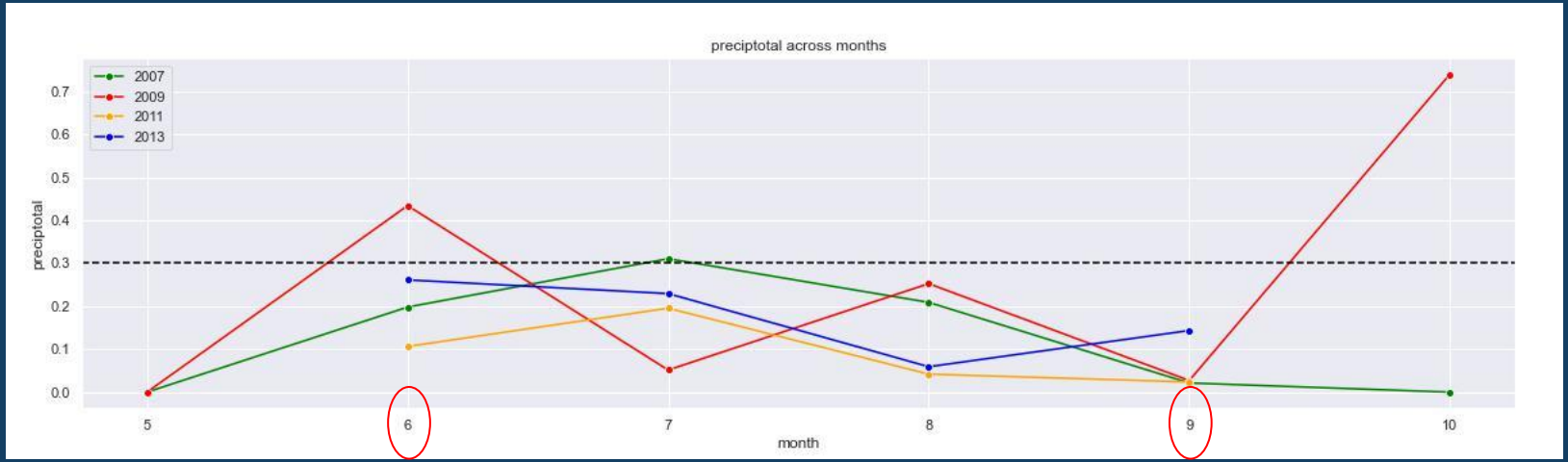
- Generally, average wind speed **decreases** from June (6) to September (9)
- Exception is seen in 2007 where there are fluctuations in average wind speed probably due to more extreme weather conditions

TOTAL PRECIPITATION & WNV



- **Low** or **no** precipitation levels (<1.0 inch) sees higher frequency of mosquitoes with wnvpresent = 1

TOTAL PRECIPITATION & WNV



- **Low** precipitation levels in June (6) to September (9)
  - at <0.3 inch

WEATHER CONDITIONS IN SUMMER

HIGH TEMP

>70 degrees fahrenheit

LOW PRECIPITATION

<0.3 inch

LOW AVE WIND SPEED

6-8 mph

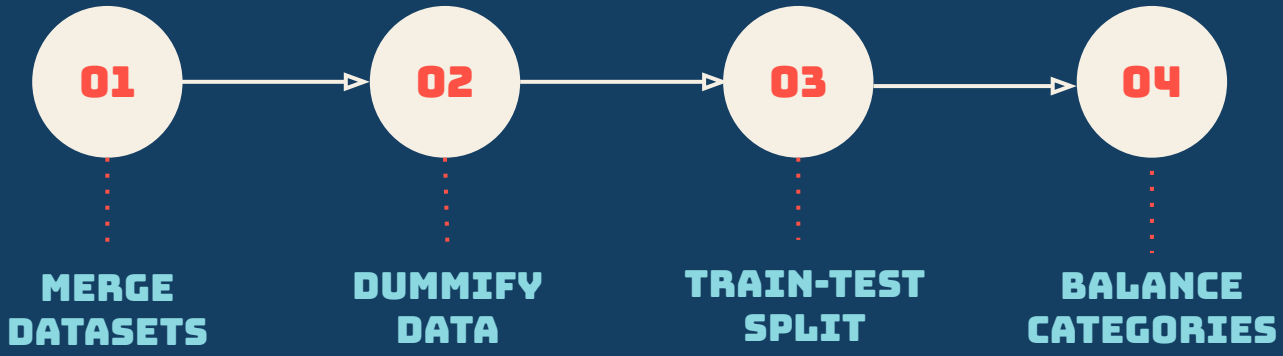
MODELLING & EVALUATION

NEW PREDICTIVE FACTORS CREATED

TRAIN DATASET
Year/Month/Day
Week of Year

WEATHER DATASET
Relative Humidity
Day/Night Length
Weather 9 Days Ago

DATA PREPARATION



MODELLING AIM

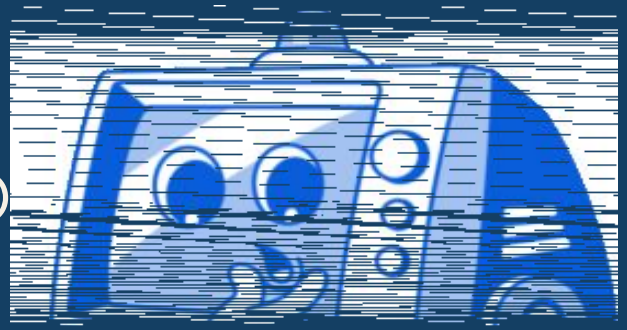
Create a model that achieves

at least 75% on recall

in predicting the presence of WNV.

MODELS BUILT

Several classification models were built and trained on the prepared dataset



- Logistic Regression (Baseline)
- Random Forest
- AdaBoost
- XGBoost

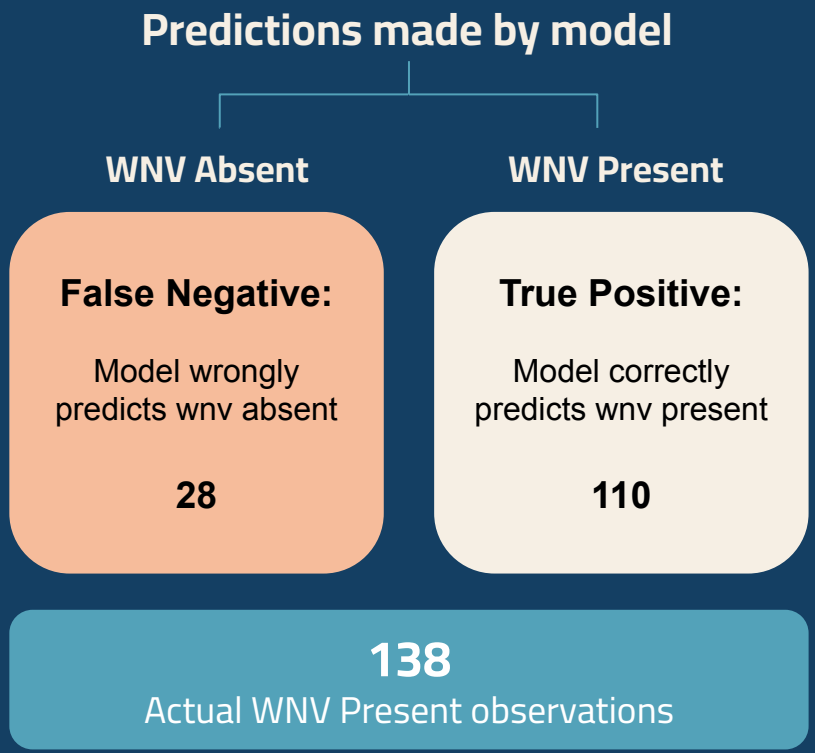
METRIC USED

- Recall was selected as the performance metric
- Measure of model performance in terms of correctly identifying the positive class out of all the actual positives
- Focus on true positive predictions

MODEL PERFORMANCE

MODEL	TRAIN RECALL	TEST RECALL
LOGISTIC REGRESSION (BASELINE)	0.82	0.68
RANDOM FOREST	**0.90**	**0.83**
ADABOOST	0.88	0.76
XGBOOST	0.93	0.58

MODEL EVALUATION



Recall:

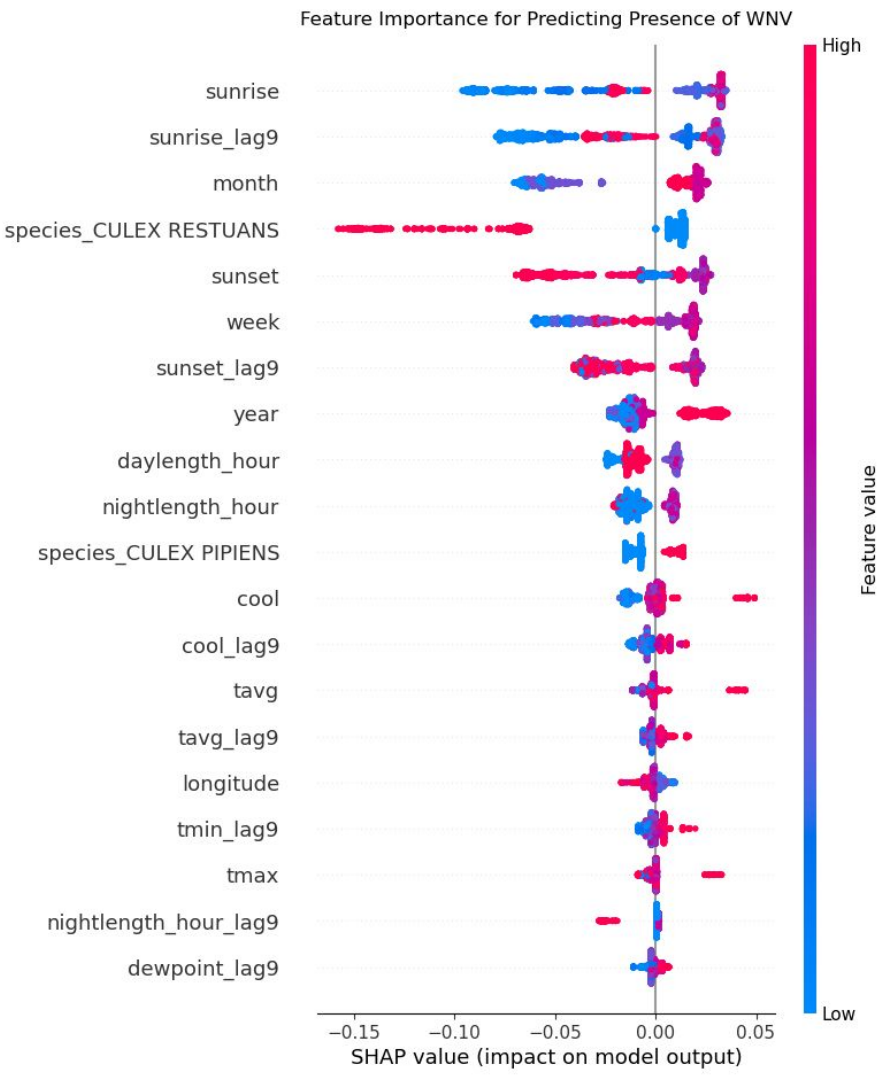
TP/(TP+FN)

= 110/138

= **0.83**

NOTABLE FACTORS

- Mosquito season predominantly in Q3
- Length of day/night affects mosquito activity
- Higher temperatures correlates with increased WNV presence
- Presence of Culex Pipiens & presence of Culex Restuans

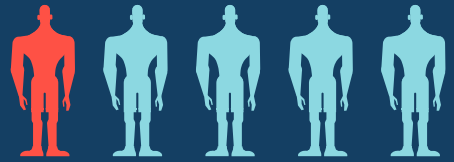


COST-BENEFIT ANALYSIS

WNV INFOGRAPHICS

1 / 150

Develops serious, sometimes fatal, illness

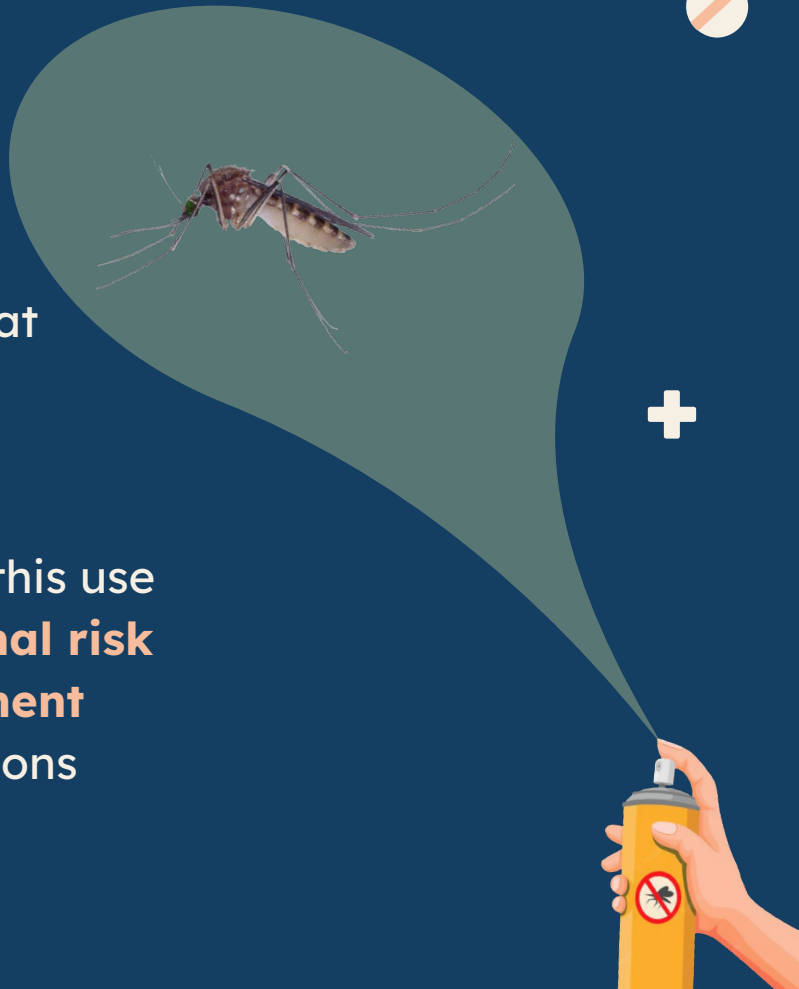


1 out of 5 develops a fever among other symptoms

No vaccine nor medication available

EXISTING EFFORTS

- **Adulticiding**
  - Aerial spraying of pesticides aimed at **eliminating adult mosquitoes**
- **Environmental Concerns**
  - Pesticides have been evaluated for this use and have been found to **pose minimal risk to human health and the environment** when used according to label directions



SPRAYING: PROS & CONS

PROS
Decrease in Mosquito populations
Reducing medical expenses associated with mosquito-borne illnesses
Preventing Loss of Productivity

CONS
Financial cost of Spraying

COST SAVINGS

$500K

TO SPRAY ENTIRE CHICAGO PER MOSQUITO SEASON

Cost can be reduced with model's guidance

$3.52/acre x 150,100 acre

$56 MILLION

SPENT YEARLY ON HOSPITALISATION

for WNV disease cases

OVER $55.5 MILLION SAVED

CONCLUSION & RECOMMENDATIONS

CONCLUSION : MISSION SUCCESS

- Model **outperforms baseline model**
- Model can **more accurately predict WNV presence**
- Predictions will **help Chicago and the CDPH more efficiently allocate resources** to effectively prevent transmission of WNV

LIMITATIONS

- Projected cost subject to change
  - Yearly review of cost-benefit analysis necessary.

- Limited features in dataset
  - Features such as number of dead birds reported has a strong impact on the spread of WNV. However, this can be easily resolved.

RECOMMENDATIONS

- Individual level:
  - Encourage civilians to strengthen personal protection against mosquitoes
- City level:
  - Enhance ground surveillance for mosquito breeding grounds and stagnant waters
- State level:
  - Further invest in research to eradicate mosquito-borne diseases

THANK YOU!

APPENDIX

MODEL OPTIMIZATION
RANDOM FOREST

- Hyperparameters of our best performing model, Random Forest, were tuned to increase performance
- Parameters optimized were max_depth, max_features, min_samples_split and n_estimators which we ran through a gridsearch.
- Rewarded with a humble improvement of 0.01 on Recall score.

RELATIVE HUMIDITY FORMULA

Relative humidity can be derived from dew point and temperature using the formula below.

17.625 and 243.04 are known as magnus coefficients obtained through experiments.

$$RH = 100 \times \left[\frac{e^{\frac{17.625 \times D_p}{243.04 + D_p}}}{e^{\frac{17.625 \times T}{243.04 + T}}} \right]$$

DAY AND NIGHT LENGTH

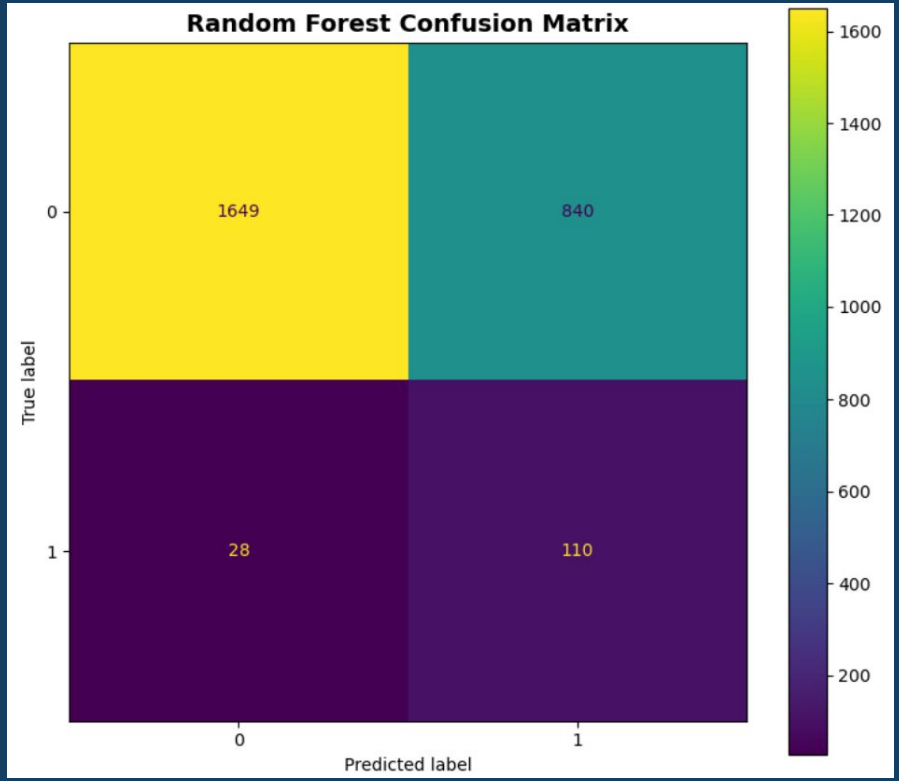
Day length = round(Sunset) - round(Sunrise) timing

Divided by 100.

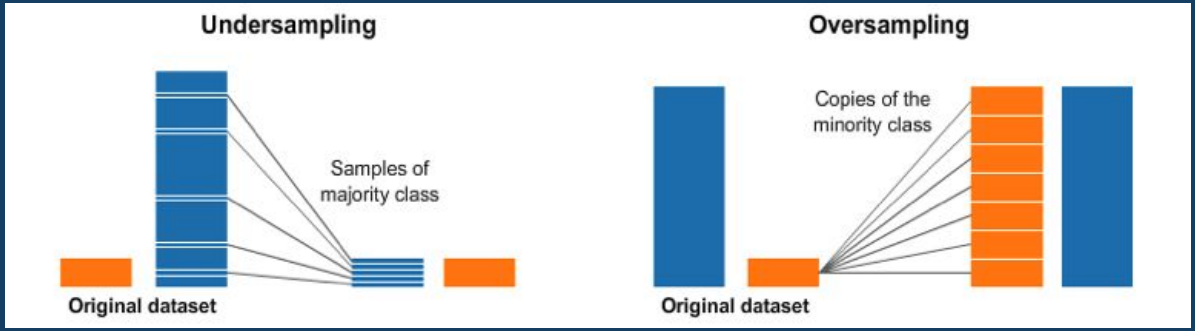
CROSS VALIDATION RECALL SCORES

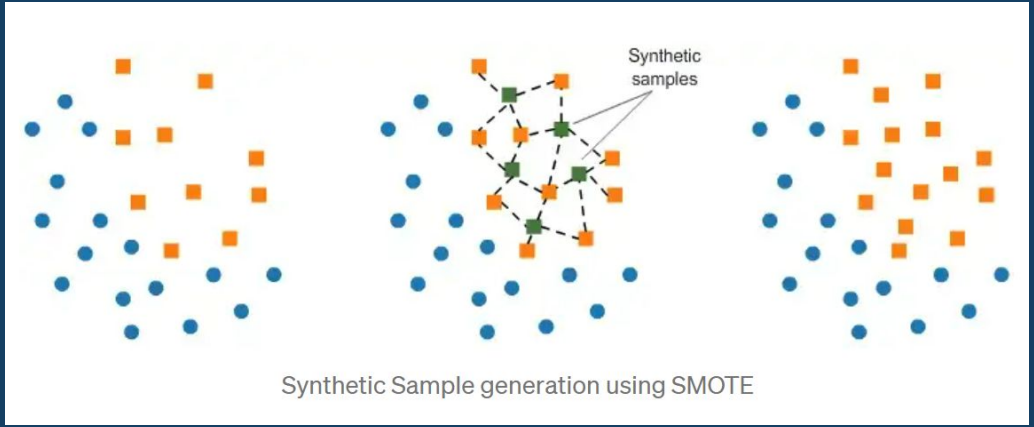
- Logreg 0.82
- RF 0.89 (pre tuning)
- Adaboost 0.88
- XGboost 0.92
- 3 Folds, Recall Score

CONFUSION MATRIX



SMOTE


Undersampling
Samples of
majority class
Original dataset
Oversampling
Copies of the
minority class
Original dataset


Synthetic
samples
Synthetic Sample generation using SMOTE

COST OF SPRAYING BREAKDOWN

Zenivex = $299.80/gl
1 gallon = 128 fluid ounce
1 fluid ounce = $2.34
1.5 fluid ounce = $3.51
$3.52 per acre
150,100 acres x \$3.51 = **$526,851**

[Source](https://www.gfmosquito.com/wp-content/uploads/2017/07/2017-ND-Mosq.-Control-Quotes-Tabulation.pdf)

$56 MILLION?!

AMJ TROP MED HYG

Table 6

Total estimated costs for United States hospitalized West Nile virus cases and death from 1999 through 2012 by cost category from simulation model in 2012 USD

Cost category	Mean*	95% CI	Median*	Range
Total acute medical care	$252,115,100	($158,022,000–$458,998,400)	$230,879,300	($115,644,400–$2,822,846,000)
Total acute lost productivity†	$22,081,260	($9,550,370–$63,069,700)	$16,144,050	($7,070,480–$2,643,251,000)
Total long-term medical care	$27,570,280	($11,566,780–$56,221,870)	$25,468,510	($6,087,800–$118,883,900)
Total long-term lost productivity	$26,866,800	($13,526,800–$48,279,320)	$25,416,720	($7,790,800–$85,567,700)
Total lifetime lost productivity caused by deaths‡	$449,464,800	(NA)	$449,464,800	(NA)

RANDOMFOREST

Works well with non-linear data

Lower risk of overfitting

Runs efficiently on a large dataset.

RANDOM FOREST CLASSIFIER