West Nile Virus Prediction

Lynette Ng
Daphne Kwok
Samuel Cheah
Albin Wan

Presentation Flow

1. Background
2. EDA
3. Modelling
4. Conclusions

What is West Nile Virus?

Mosquito-borne virus

Spreads by the bite

1 in 5 fall mildly ill
1 in 150 severely ill

Originated from West Nile area of Africa in 1937

First US case of detected in 1999

Mosquitoes pick the virus from infected birds

Problem statement

Predict the occurrence of virus

Analyze to see if there are main predictors to the virus

Recommend best practices to reduce human contraction of virus

What data are we working with?

Weather.csv

- Weather features (i.e. temperature, dewpoint, sunrise, precipitation)
- 2007-2014

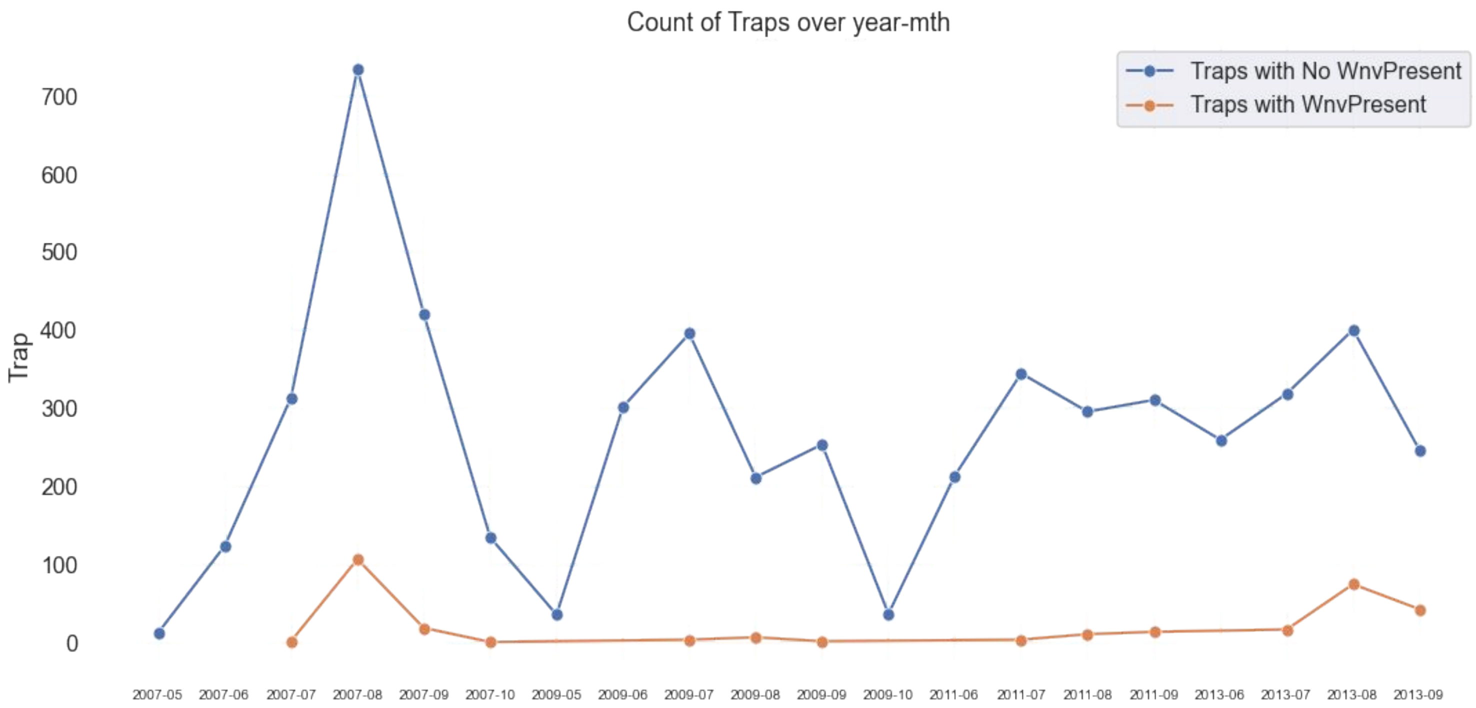
Spray.csv

- Latitude and longitudes
- 2011 and 2013

Train/test.csv

- Location of traps and number of mosquitos
- Presence of WNV
- 2007-2014

EDA - Target Class

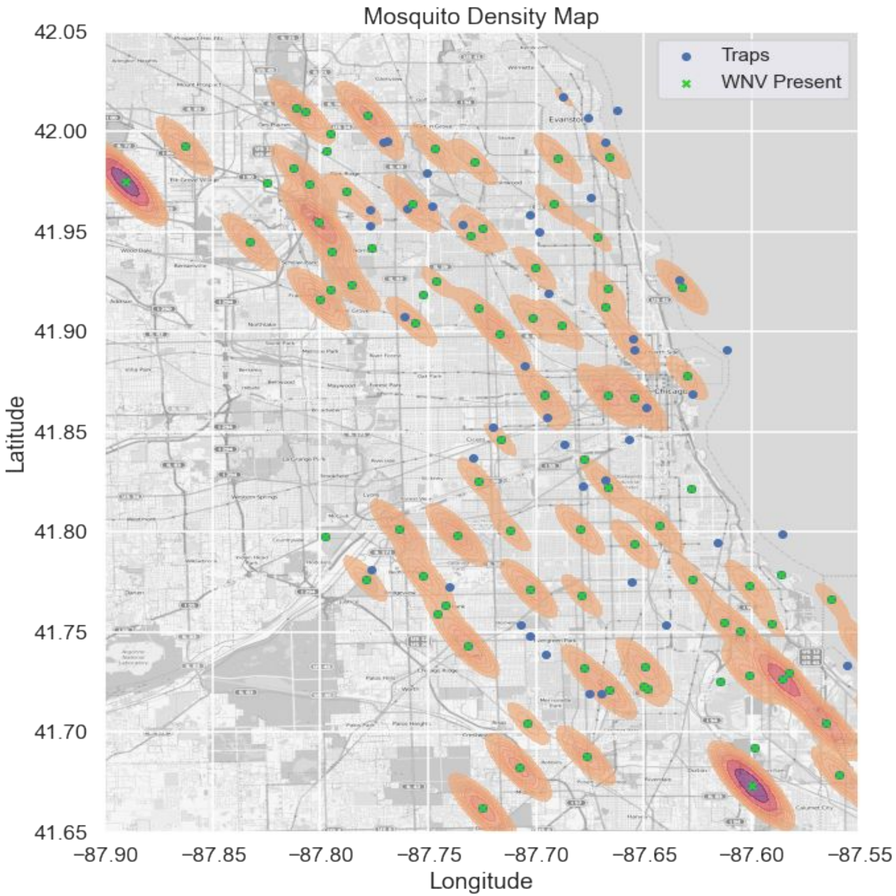


5% unbalanced class

EDA - Mosquito density

- Areas shaded in peach shows high concentration of mosquitoes

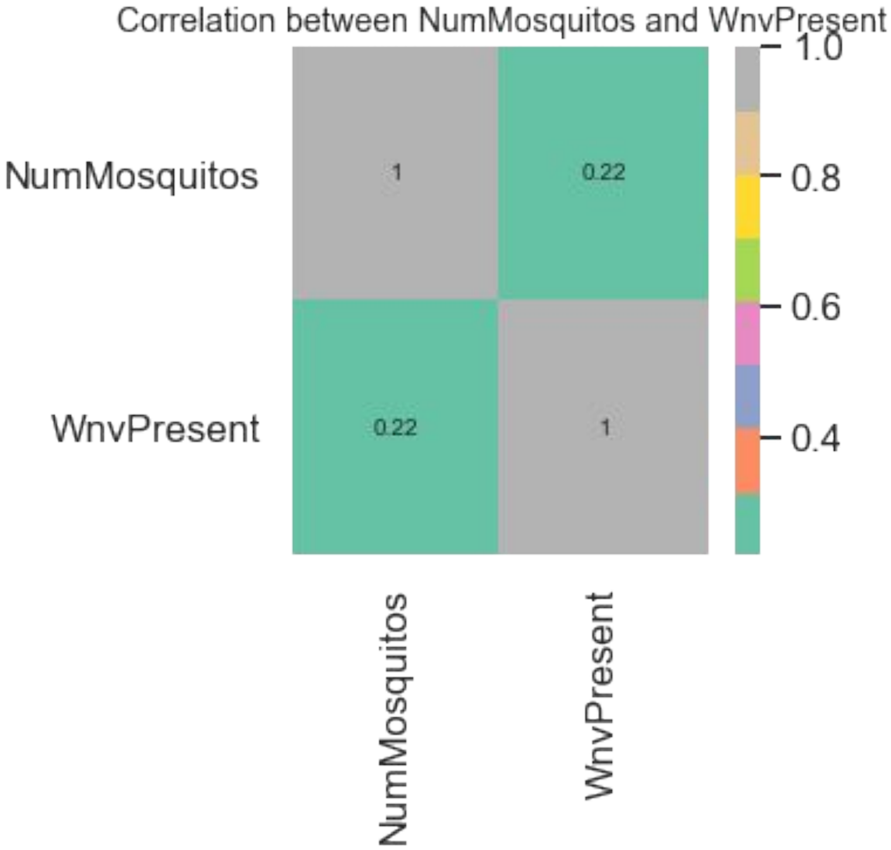
- Areas of high mosquito population tend to overlap with WNVpresent



EDA - Mosquito density

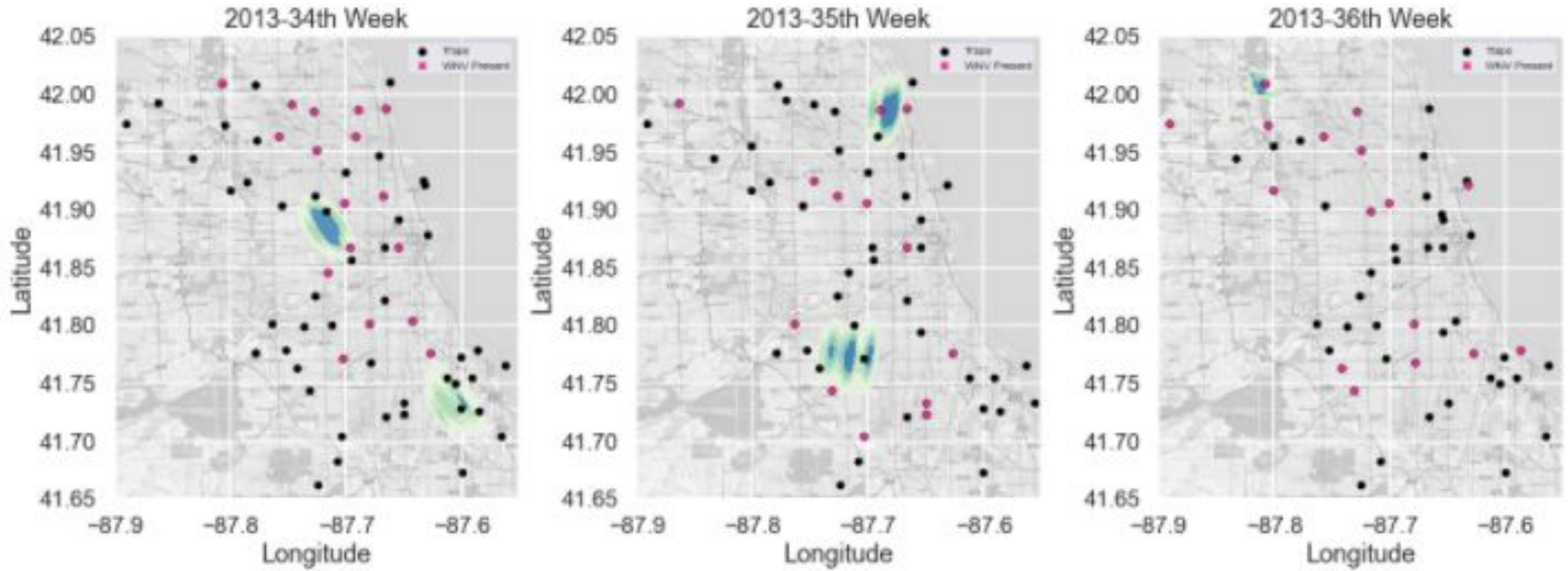
- Positively correlated, but weak

- Indicates that there could be some positive relationship between NumMosquitos and WnvPresent



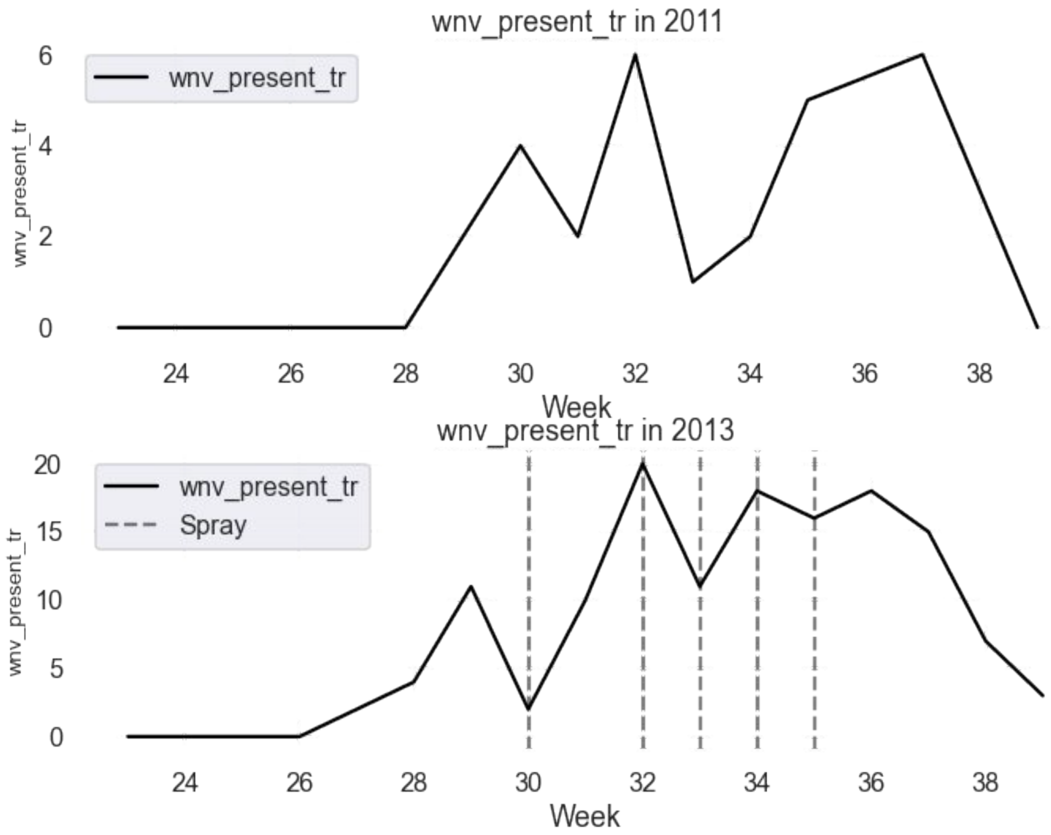
EDA - Spray map

- Sprays not well targeted in areas with WNV
- Slight week on week reduction of WNV presence



EDA - Spray: Location Analysis

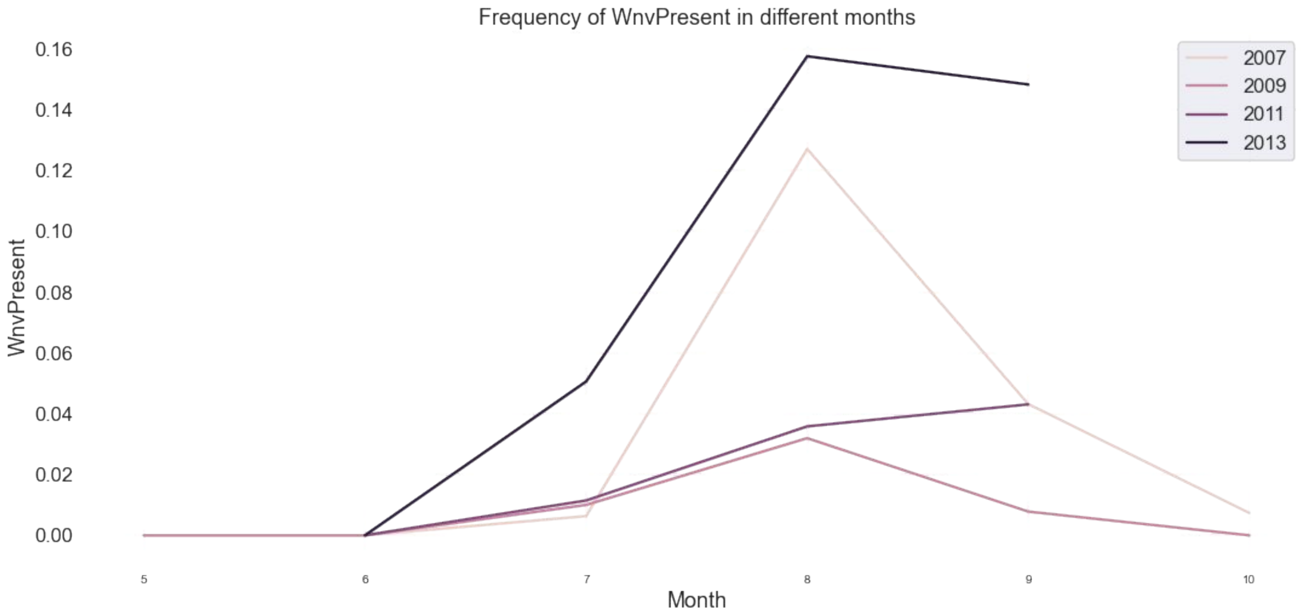
- Dotted line shows sprays that **overlapped with traps**
- In 2011, no sprays were located within a 120m range of the traps
- Sprays might have been targeted at the wrong areas, no consistent trend of reduction of WNV post spray



EDA - Spray - Timing Analysis

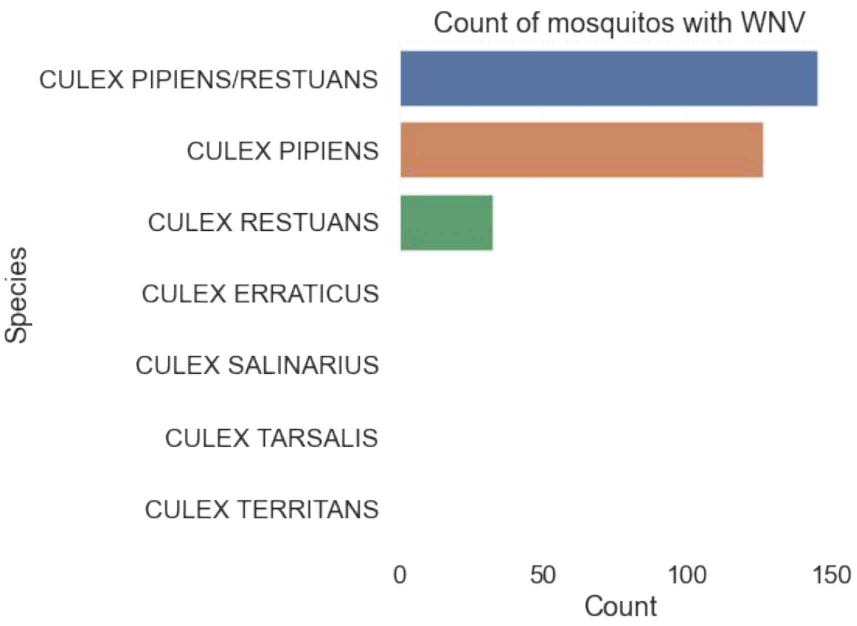
- Sprays in 2011 were made late in the year (week 35/36
- Timing of sprays were mitigative rather than preventative
- Could maybe be made at earlier breeding stages

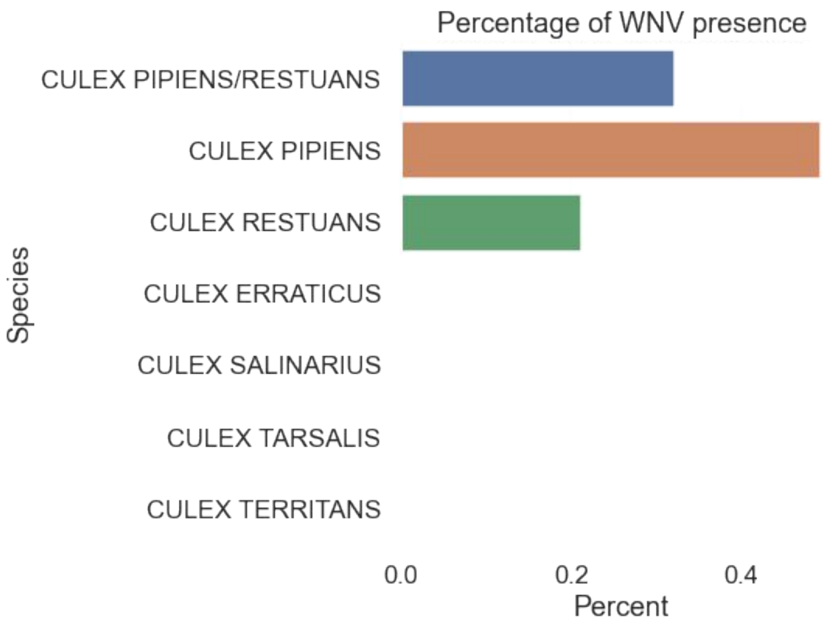
EDA - WNV against Month



Spikes in August for West Nile Virus

EDA - Species

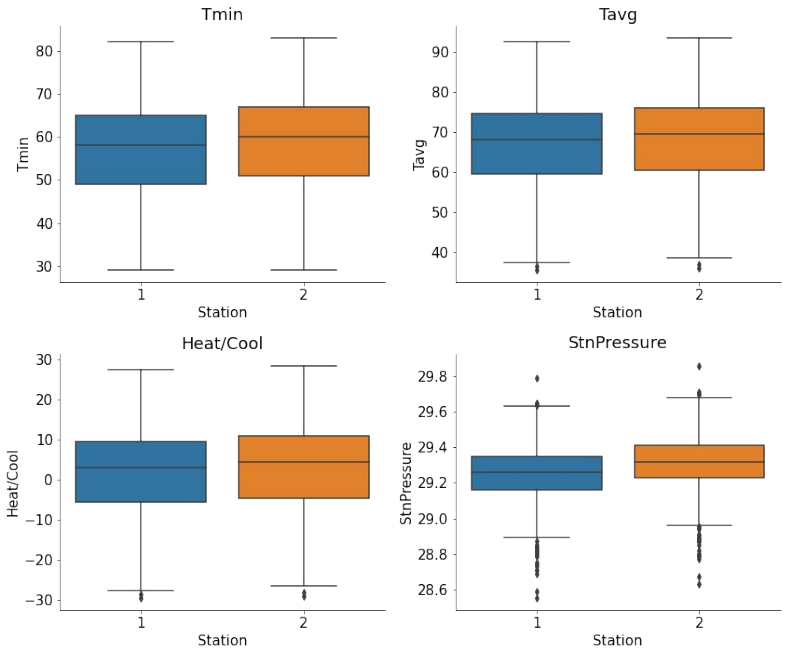




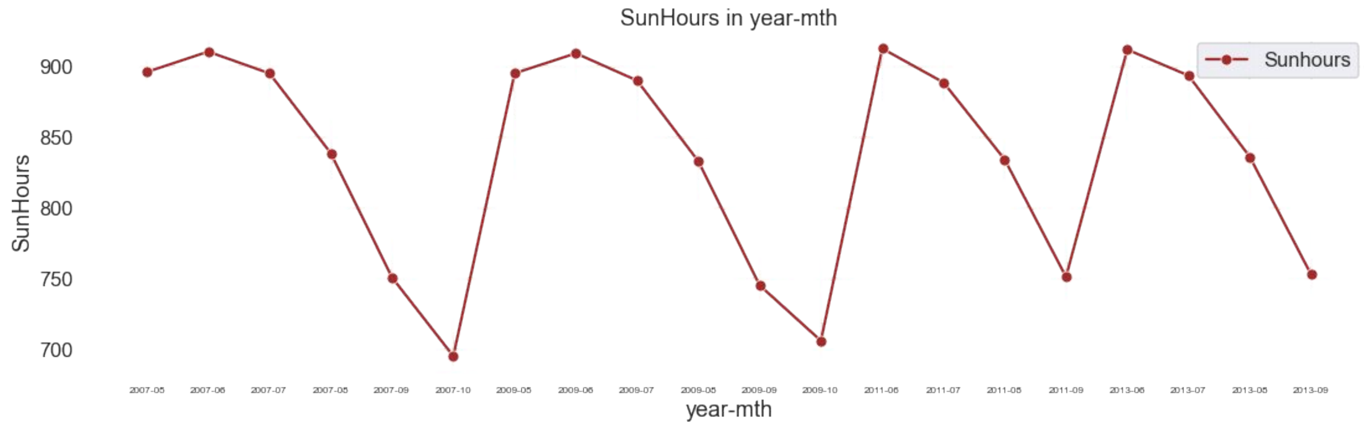
Culex Pipiens and Restuans culprits for the spread of virus

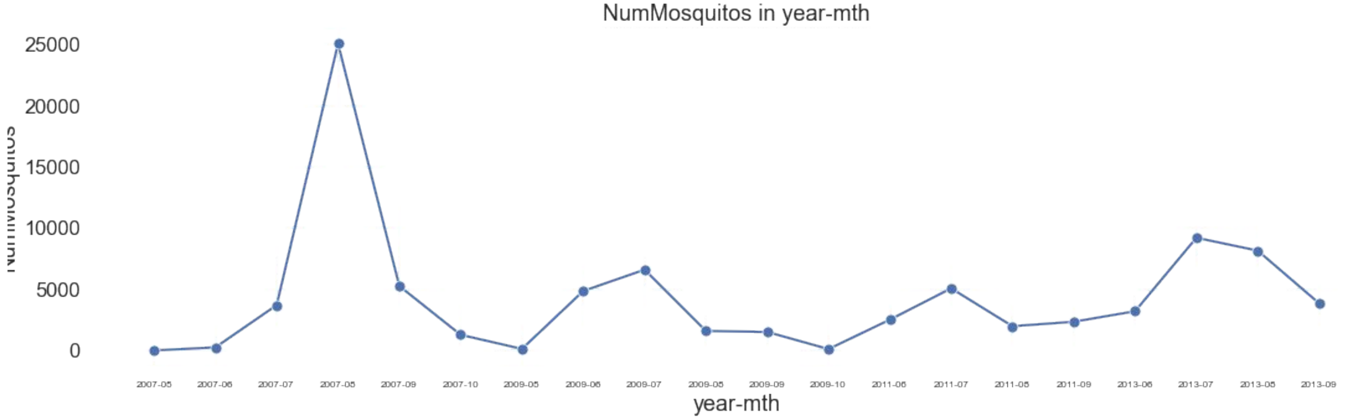
EDA - Weather Data

- Station 1 weather data is not that different from Station 2 except for Tmin, Tavg, Cool & StnPressure
- Station 2 has more null values
- For Tmin, Tavg, Cool & StnPressure, the mean of the Station 1 and Station 2 is used

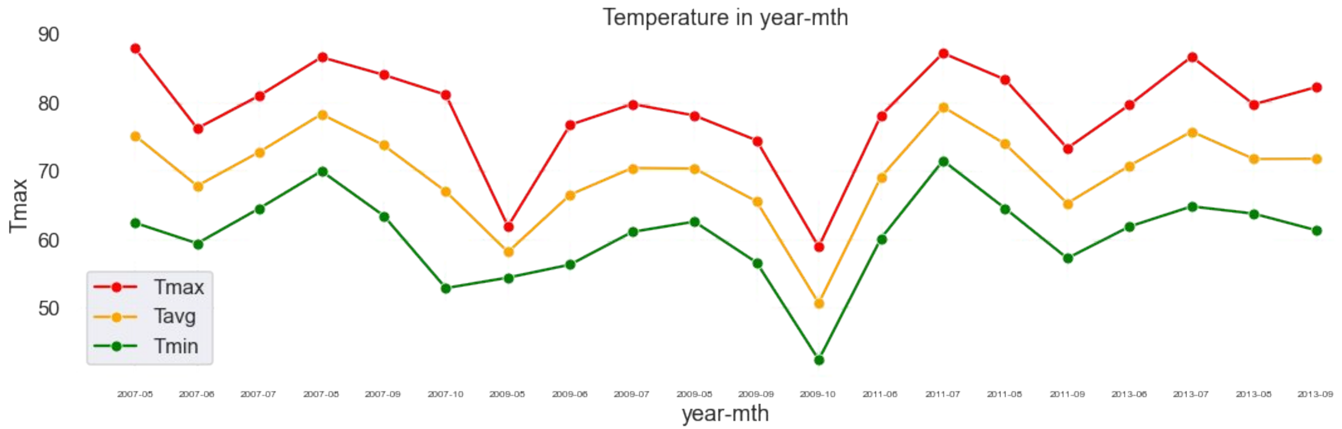


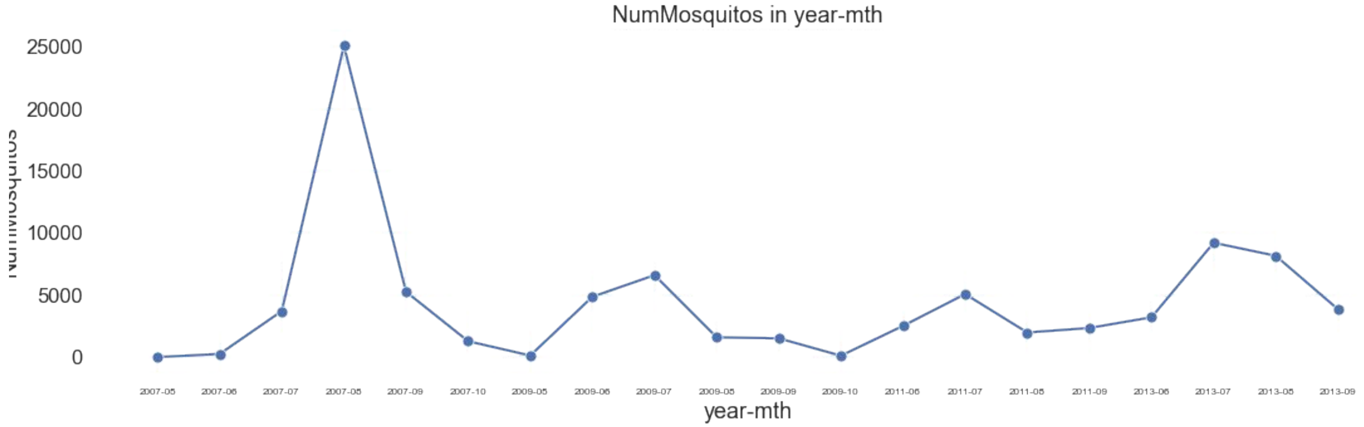
EDA - Sunhours against Time and NumMosquitos



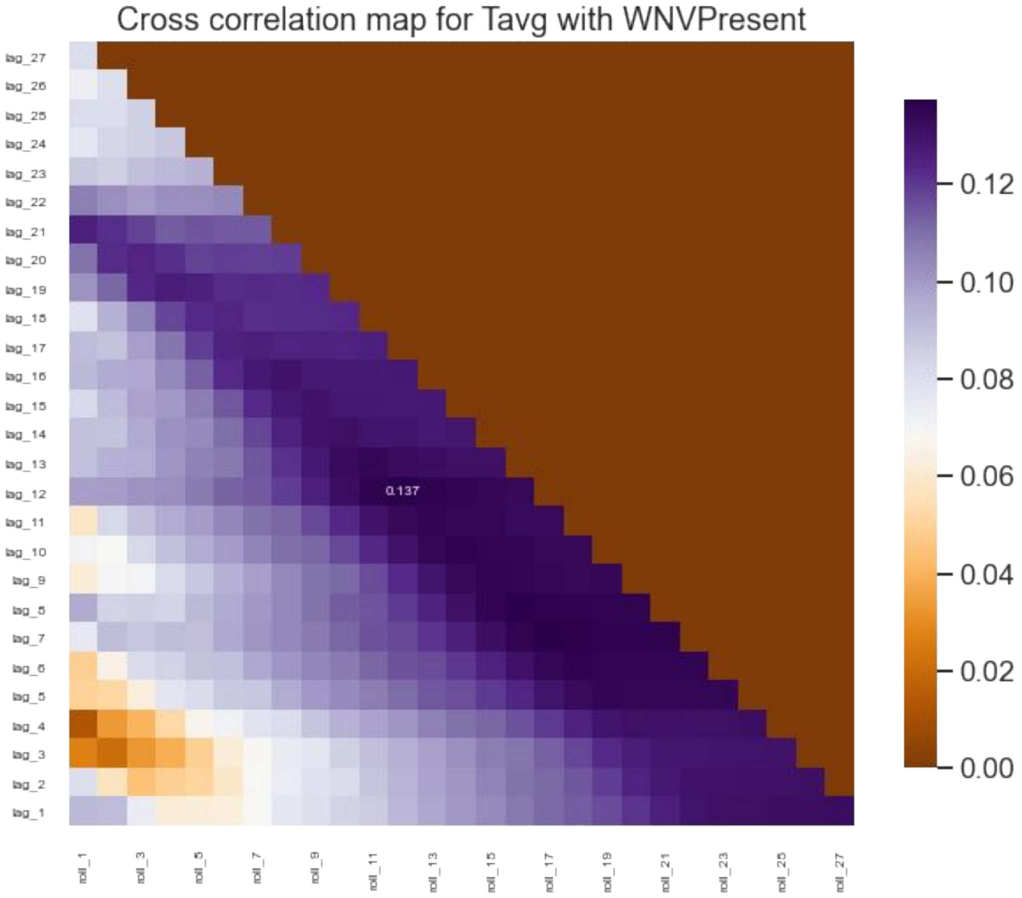


EDA - Temperature against Time and NumMosquitos



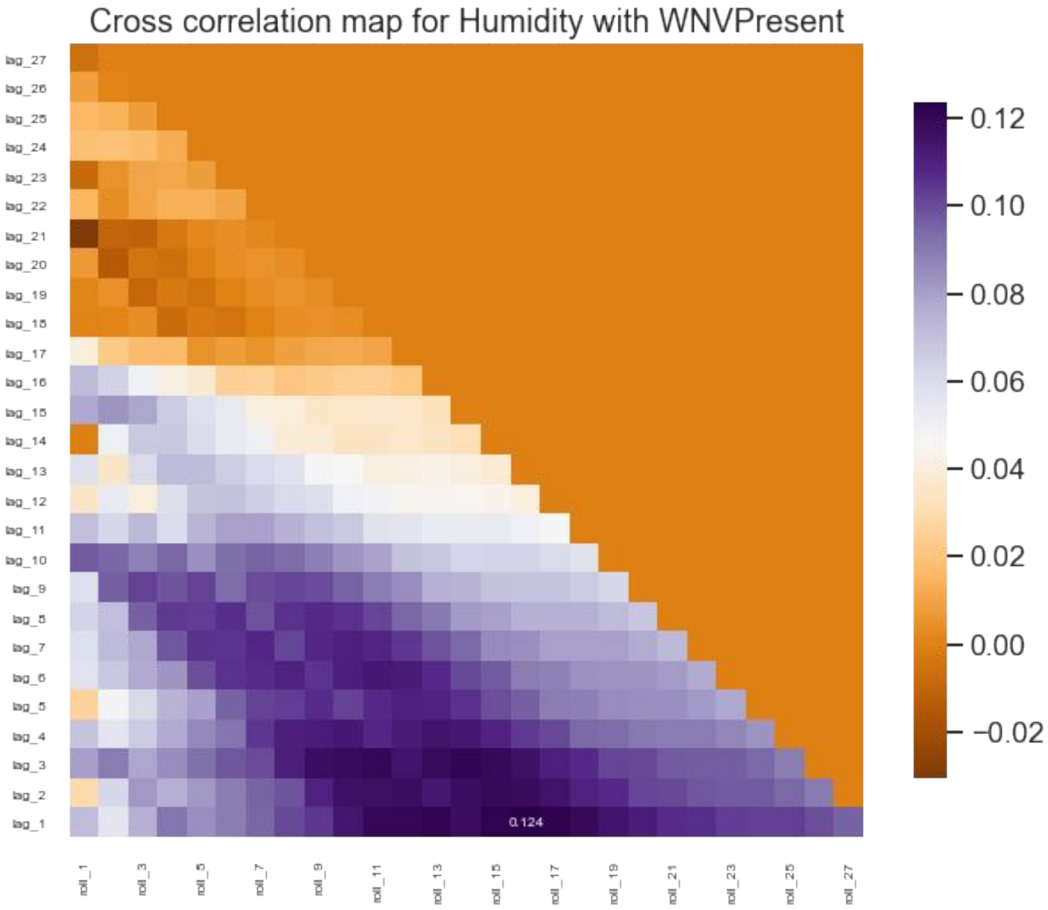


EDA - Cross correlation map for Tavg



- Identified features that showed stronger correlation against WnvPresent when they are rolled and lagged
- Strongest correlation for Tavg: Rolling mean of 12 days & Lag of 12 days

EDA- Cross correlation map for Humidity



- Humidity is calculated from Dewpoint and Tavg
- Strongest correlation for Humidity: Rolling mean of 16 days & Lag of 1 day

Model Results - We tested a few models

Linear Models

- Logistic Regression
- Stochastic Gradient Descent Classifier

Tree Models

- Random Forest
- Extra Trees

Booster Models

- ADABoost
- GradientBoost
- XGBoost
- LightGB

Model Results

Metrics used to evaluate models:

- **ROC-AUC**
  - Measure of a model’s discriminability

- **Recall, aka Sensitivity**
  - True Positive/ (True Positive + False Negative)
  - Chosen because of our focus on public health
  - False Negatives have a higher public health cost than False Positives

Model Results

Opted to use resampling techniques

- Due to imbalanced dataset, this was our result
- No WNVPresent predictions
- Recall = 0

```
Training ROC AUC: 0.833919159622l511
Training recall: 0.0
Training accuracy: 0.946107784431l377
-----------
Validation ROC AUC: 0.8367121683110329
Validation recall: 0.0
Validation accuracy: 0.9460135859849839
                      WNV present  WNV not present
Predicted WNV                   0                0
No predicted WNV              151             2646
```

Model Results

SMOTE

- Synthetic Minority Resampling

ADASYN

- Adaptive Synthetic Sampling
- Weights-adjusted approach to SMOTE

Model Results

Some of our better models:

Model	Train AUROC	Val AUROC	Val Recall	Val Accuracy
SGDClassifier with SMOTE	0.810	0.818	0.774	0.752
RandomForest with SMOTE	0.853	0.836	0.795	0.702
RandomForest with ADASYN	0.857	0.838	0.808	0.691
ExtraTrees with SMOTE	0.844	0.822	0.815	0.691
XGBoost with SMOTE	0.874	0.846	0.715	0.795
XGBoost with ADASYN	0.871	0.839	0.702	0.801

Model Results

Final Model:

Val AUROC: 0.84

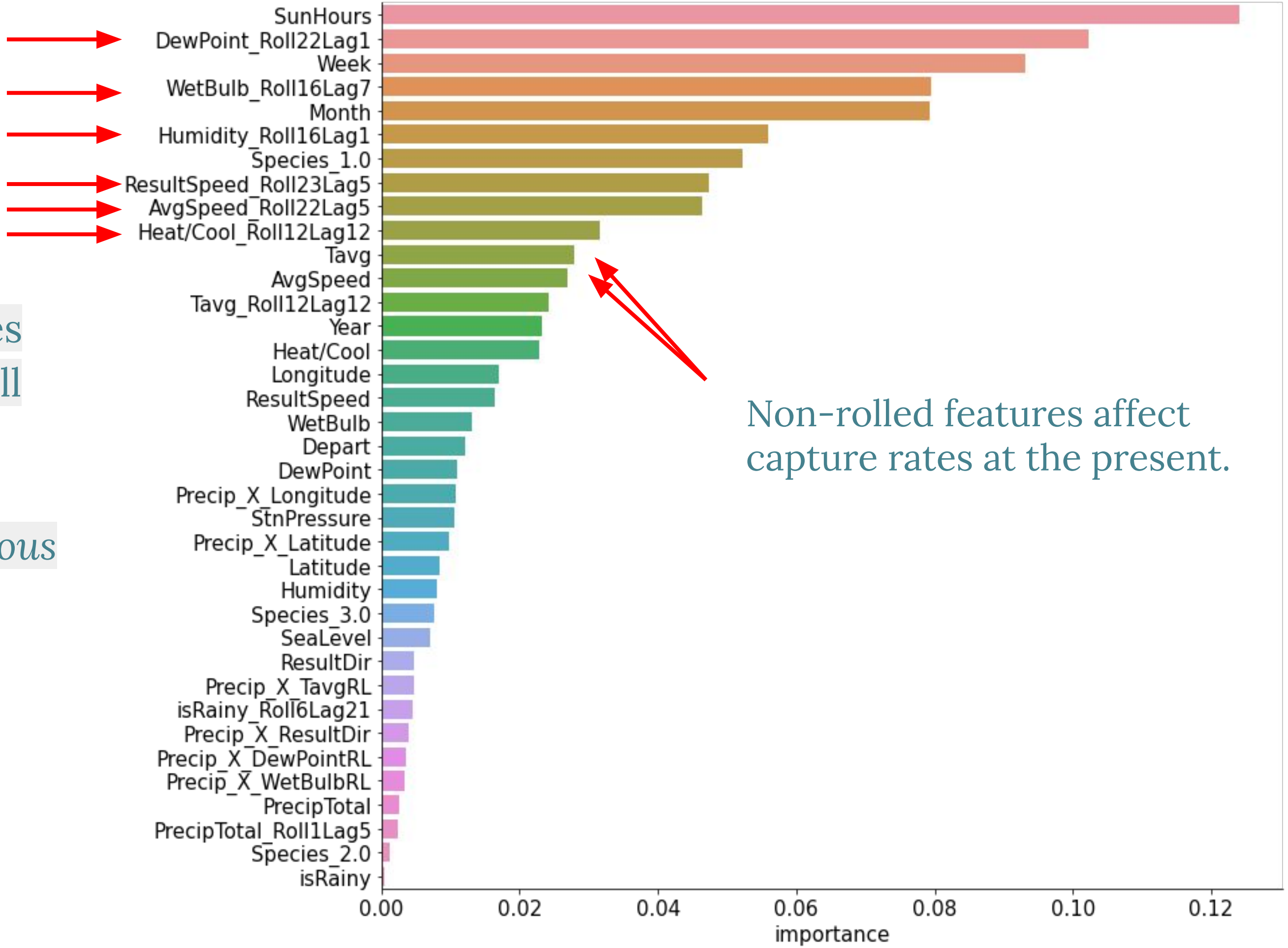
Val Recall: 0.82

```
Training ROC AUC: 0.8614897994442254
Training recall: 0.8856209150326797
Training accuracy: 0.6868615709756957
----------
Validation ROC AUC: 0.8399208101194856
Validation recall: 0.8211920529801324
Validation accuracy: 0.6907400786557025
                  WNV present  WNV not present
Predicted WNV             124              838
No predicted WNV           27             1808
```

```
pipe_params = {
    'adasyn__n_neighbors':[2],
    'rf__n_estimators':[200],
    'rf__max_depth':[5],
    'rf__min_samples_leaf':[5],
    'rf__class_weight':['balanced_subsample'],
    'rf__max_samples':[0.5]
```

Model Results

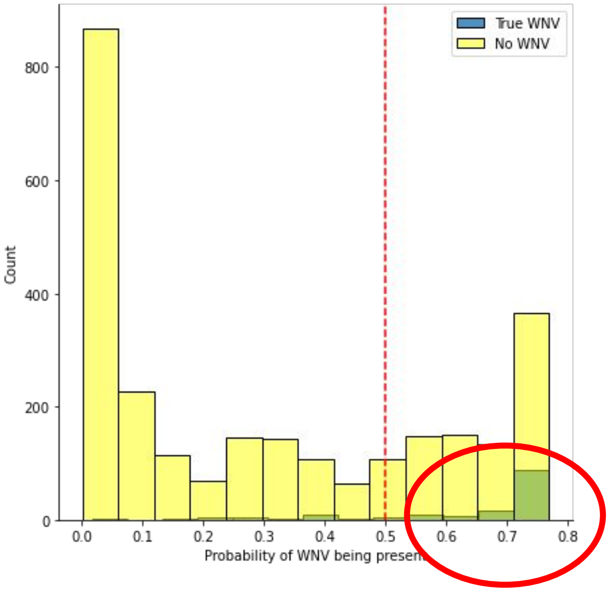
- Engineered features performed very well
- Captures weather conditions of *previous* generation

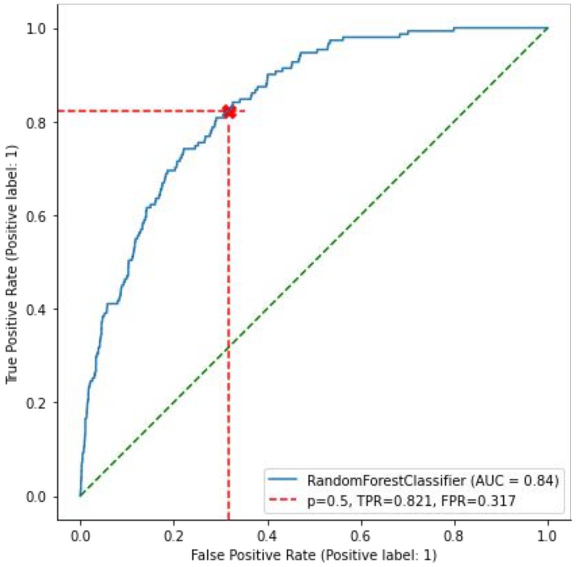


Non-rolled features affect capture rates at the present.

Model Results

Model still uncertain about our minority positive class





Conclusions

2 main types of features matter:

- **Present weather features (i.e. no roll/lag)**
  - Affects activity and trap capture rates of present generation

- **Rolled/Lagged weather features**
  - Captures conditions of previous mosquito generation that affect survivability and breeding conditions.

Cost-benefit analysis

Spraying	Healthcare
$2,000,000* a year to spray the entirety of Chicago indiscriminately (8 times)	Cost of West Nile Fever: $1,000 per person (incl. lost productivity) Cost of West Nile Neuroinvasive Disease: $72,000* per person (incl. lost productivity) Cost of death: $3,000,000*

**See appendix for details*

Cost-benefit analysis

*"According to IDPH, the first human West Nile virus death in Illinois last year was reported on September 29, 2017. In 2017, there were 90 human West Nile virus cases, **including eight deaths**."*

- CBS Chicago, Aug 29, 2018,
'West Nile Virus Death Reported In Illinois'

This is equivalent to $24,000,000 lost, or 12 years worth of sprays

Recommendations

Fund research for Genetically Modified Culex

More active on-the-ground monitoring

Timing of spray

Education

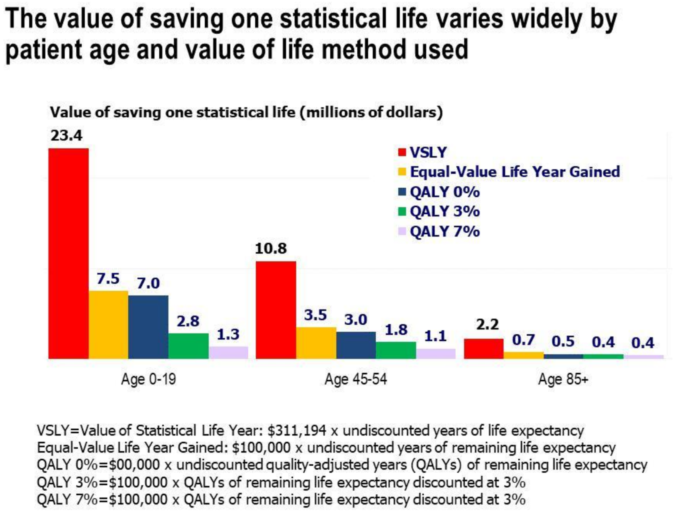
End

Appendix

Chicago uses a type of pesticide called Zenivex. A 30 gallon container of this costs $10800. For a city like Chicago, they would employ the use of trucks to spray Zenivex through the streets. From the Central Mass. Mosquito Control Project, we found out that these pesticide spraying trucks can spray betwwen 4.5-9 ounces of pesticide per minute, and travel about 10-15 mph. We can then estimate that 1 single 30-gallon container of pesticide will provide around **12-14 hours of spraying**. There are approximately 4000 miles of street in Chicago to spray, if we aim to cover all of Chicago. Given the speed of trucks and the amount of street to cover in Chicago, we estimate that these trucks can take anywhere between **260-300 manhours** to spray the entire of Chicago. Thus we will need about 25 30-gallon containers, about $250k worth, to spray Chicago once. Accounting for manpower costs at a minimum wage of around $15 per hour, it will cost Chicago an additional $4-5k. All in, the cost of spraying per instance is **$255,000**. Let's assume that for optimal benefit, we spray over a period of 8 weeks, starting from before the typical peak onset of WNV in July. It will cost Chicago **$2 mil** a year to spray the *entirety* of Chicago indiscriminately.

Appendix

A person getting infected by WNV has about a 1 in 5 chance of developing West Nile Fever, and a 1 in 150 chance of developing a more serious disease called West Nile Neurodegenerative Disease as per the CDC. A paper by R. Peterson has already done this analysis for us, so let's look at his findings. According the Peterson, the average cost for a person who develops West Nile Fever is about **$1000** in healthcare and lost productivity costs. The average cost for a person who develops the more severe WNND is **$72,000** in comparison. More recently, a research scholar named Christopher Conover published an article discussing the cost of a life lost to COVID. This graphic summarizes his findings.



While this was done in the context of COVID, we can likewise extrapolate similar 'costs' of a person's death. Conover himself opts to use what he calls the Quality Adjusted Life Year as a metric to measure the value of a person's life, which basically takes into account the quality of life. Using his metric, he estimates the value of a person's life to be about **$3 million** for someone aged 45-59 yrs.