

# West Nile Virus Detection

Data driven approach to prevention and eradication of the virus

Project for Chicago Municipality (CM) and Chicago  
Department of Public Health (CDPH)  
By Eran Schenker

**Presentation For Data Scientists**



# Background

- West Nile virus (WNV) is most commonly spread through infected mosquitos.
- 20% of people who are infected, develop symptoms ranging from fever, to serious neurological illnesses and death.
- In 2002, first human cases of **WNV** were reported in Chicago. By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program.
- Every week mosquitos in traps across the city are tested for the virus. The results of these tests influence when and where the city will spray pesticides.
- Given weather, location, testing, and spraying data, Chicago Municipality (CM) and CDPH asked (through Kaggle ) to predict when and where different species of mosquitos will test positive for West Nile virus.

# Background

- West Nile virus (WNV) is most commonly spread through infected mosquitos.
- 20% of people who are infected, develop symptoms ranging from fever, to serious neurological illnesses and death.
- In 2002, first human cases of **WNV** were reported in Chicago. By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program.
- Every week mosquitos in traps across the city are tested for the virus. The results of these tests influence when and where the city will spray pesticides.
- Given weather, location, testing, and spraying data, Chicago Municipality (CM) and CDPH asked (through Kaggle ) to predict when and where different species of mosquitos will test positive for West Nile virus.

## The goal

*“To create a more accurate method of predicting outbreaks of West Nile virus in mosquitos to help CM and CPHD to more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus.”*

Resource: <https://www.kaggle.com/c/predict-west-nile-virus>

# Work Flow

# Work Flow

- I. Is there good data to work with?
  - What are the datasets available?
  - What are the features of the datasets - Are they informative to the goal of predicting West Nile Virus (WNV) occurrences?
  - Can we predict WNV occurrences from this basic data

# Work Flow

## I. Is there good data to work with?

- What are the datasets available?
- What are the features of the datasets - Are they informative to the goal of predicting West Nile Virus (WNV) occurrences?
- Can we predict WNV occurrences from this basic data

## II. How can we modify the data so it would work for us?

- Cleaning and standardizing features across datasets
- Engineering more informative features from the basic features given
- Merging the datasets into one superior consolidated dataset

# Work Flow

- I. Is there good data to work with?
  - What are the datasets available?
  - What are the features of the datasets - Are they informative to the goal of predicting West Nile Virus (WNV) occurrences?
  - Can we predict WNV occurrences from this basic data
- II. How can we modify the data so it would work for us?
  - Cleaning and standardizing features across datasets
  - Engineering more informative features from the basic features given
  - Merging the datasets into one superior consolidated dataset
- III. Building a statistical model to meet our goal of optimal prediction of WNV
  - What is the best model to use?
  - How can we optimize the model?
  - What is the performance of the model we are offering?
  - Are the modification efforts of the data justified?

# Work Flow

- I. Is there good data to work with?
  - What are the datasets available?
  - What are the features of the datasets - Are they informative to the goal of predicting West Nile Virus (WNV) occurrences?
  - Can we predict WNV occurrences from this basic data
- II. How can we modify the data so it would work for us?
  - Cleaning and standardizing features across datasets
  - Engineering more informative features from the basic features given
  - Merging the datasets into one superior consolidated dataset
- III. Building a statistical model to meet our goal of optimal prediction of WNV
  - What is the best model to use?
  - How can we optimize the model?
  - What is the performance of the model we are offering?
  - Are the modification efforts of the data justified?
- IV. Summary of the proposed model and final recommendations
  - What the model is revealing to us about WNV factors
  - Final actionable recommendations for Chicago Municipality and Department of Public Health



# I. Is there good data to work with?

- **What are the datasets available?**

# I. Is there good data to work with?

- **What are the datasets available?**

## Train Data

- Main data (including labels of WNV response)
- **10506** trap collections (observations)
- **12** different features

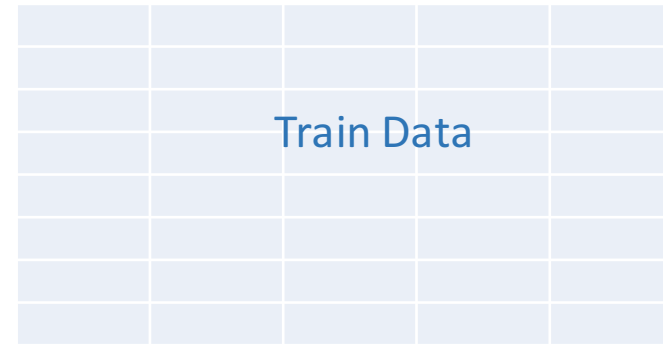
# Train Data

# I. Is there good data to work with?

- **What are the datasets available?**

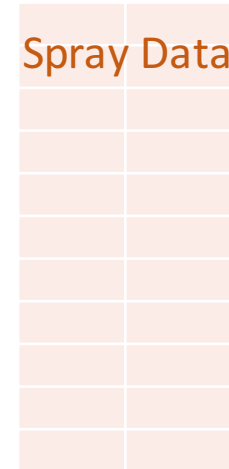
## Train Data

- Main data (including labels of WNV response)
- **10506** trap collections (observations)
- **12** different features



Train Data

## Spray Data



## Spray Data

- Time and place of spraying
- **14835** unique spraying occasions (observations)
- **4** features: Date, Time, Longitude & Latitude

# I. Is there good data to work with?

- **What are the datasets available?**

## Train Data

- Main data (including labels of WNV response)
- **10506** trap collections (observations)
- **12** different features

## Weather Data

- **2944** daily weather data, split between 2 weather stations
- **22** different features

## Spray Data

- Time and place of spraying
- **14835** unique spraying occasions (observations)
- **4** features: Date, Time, Longitude & Latitude


Train Data


Spray Data


Weather Data

# I. Is there good data to work with?

- **What are the datasets available?**

## Train Data

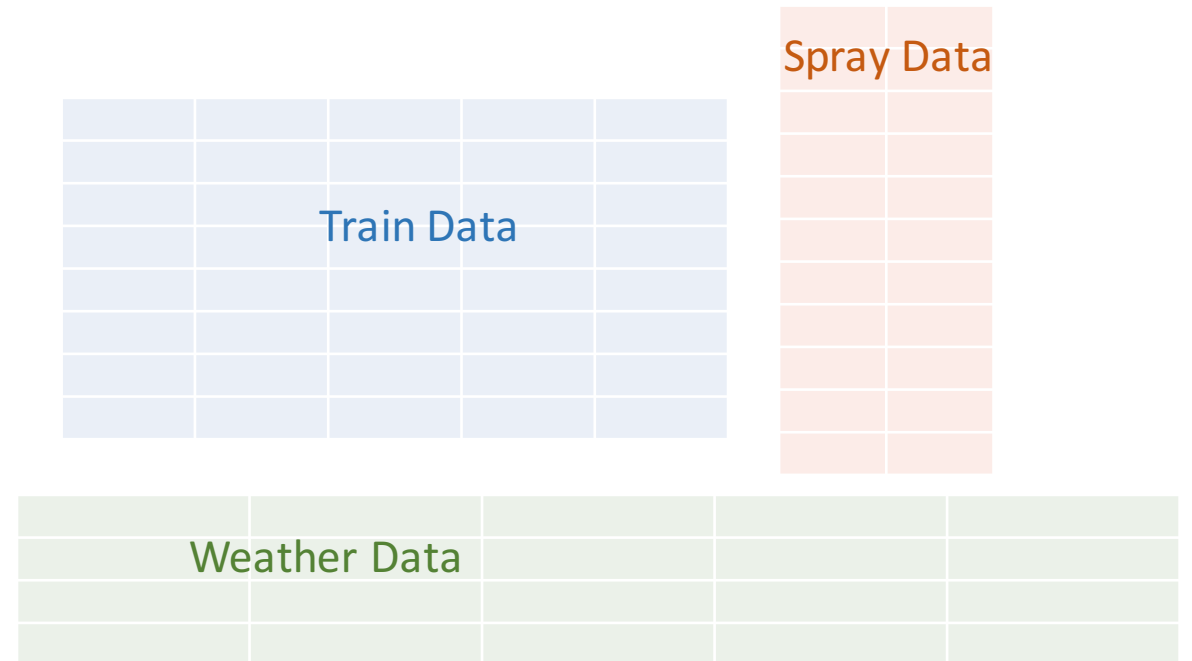
- Main data (including labels of WNV response)
- **10506** trap collections (observations)
- **12** different features

## Weather Data

- **2944** daily weather data, split between 2 weather stations
- **22** different features

## Spray Data

- Time and place of spraying
- **14835** unique spraying occasions (observations)
- **4** features: Date, Time, Longitude & Latitude



- Datasets present good potential for enriching the main Train data

# I. Is there good data to work with?

- **What are the datasets available?**

## Train Data

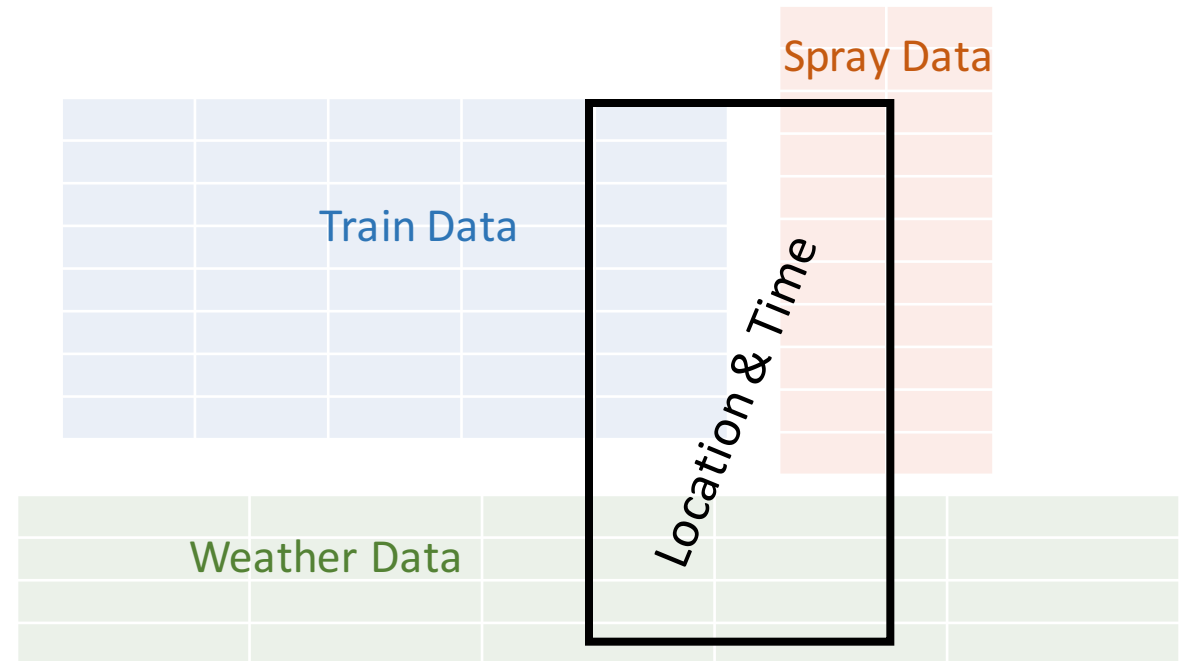
- Main data (including labels of WNV response)
- **10506** trap collections (observations)
- **12** different features

## Weather Data

- **2944** daily weather data, split between 2 weather stations
- **22** different features

## Spray Data

- Time and place of spraying
- **14835** unique spraying occasions (observations)
- **4** features: Date, Time, Longitude & Latitude



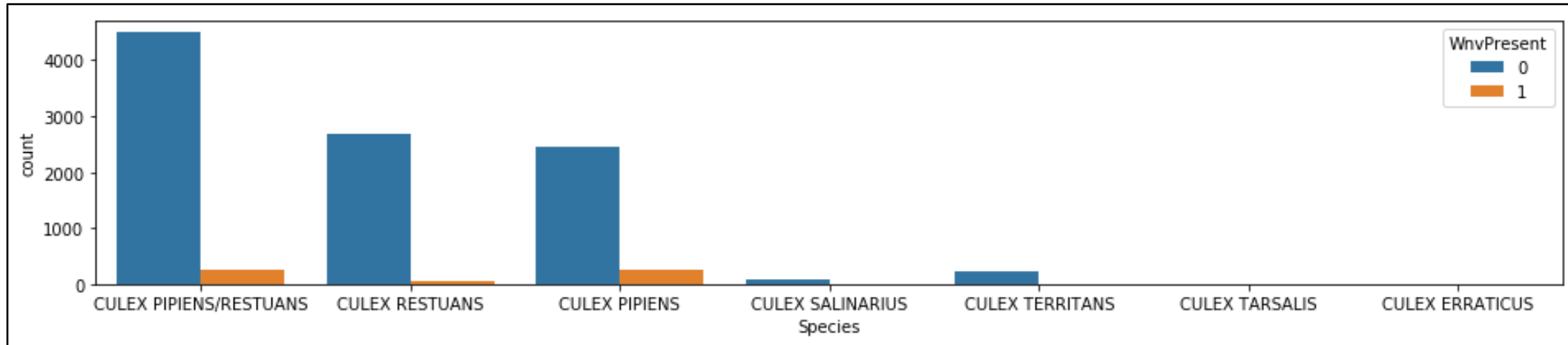
- Datasets present good potential for enriching the main Train data
- Generally speaking, location and time can be used for merging the datasets

# I. Is there good data to work with?

- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**

# I. Is there good data to work with?

- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**



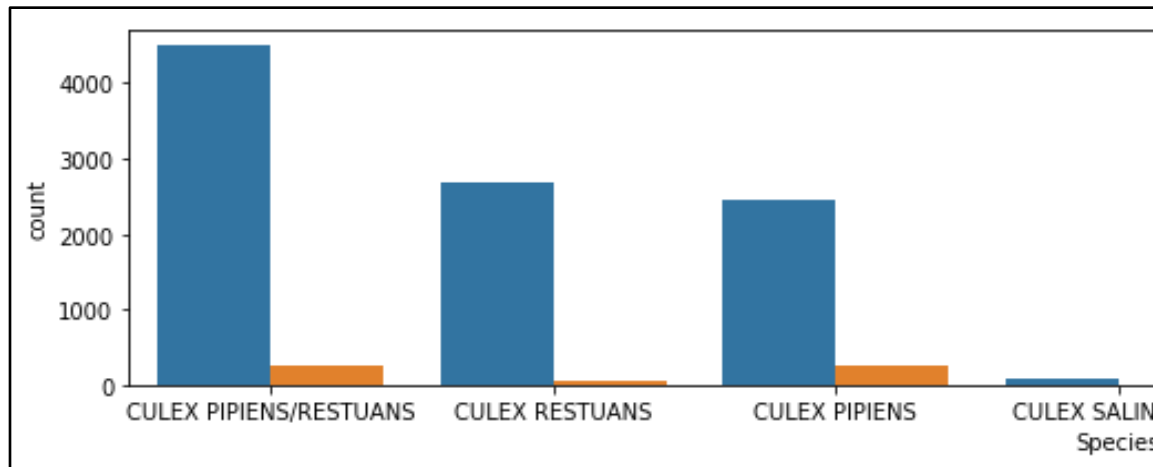
## Species

- We can see a relationship between species found in traps and the appearance of the virus. There are 6 different species (the 1<sup>st</sup> is traps with combination of 2 different species). Traps that have the virus are in orange, those who don't are in blue.
- Seems that only 2 species are common ('Culex Pipiens' and 'Culex Restuans'), and 1 specie has high number of WNV – 'Culex Pipiens' with relative portion of WNV  $\approx 10\%$



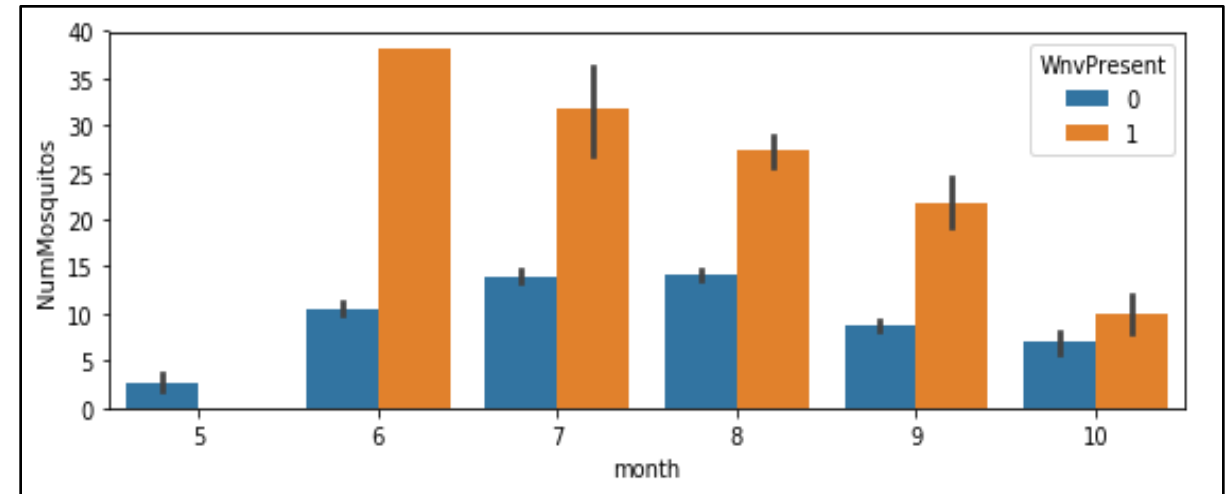
# I. Is there good data to work with?

- Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?



## Species

- We can see a relationship between species found in traps and the appearance of the virus. There are 6 different species (the 1<sup>st</sup> is traps with combination of 2 different species). Traps that have the virus are in orange, those who don't are in blue.
- Seems that only 2 species are common ('Culex Pipiens' and 'Culex Restuans'), and 1 specie has high number of WNV – 'Culex Pipiens' with relative portion of WNV  $\approx 10\%$



## Month & Number of Mosquitos

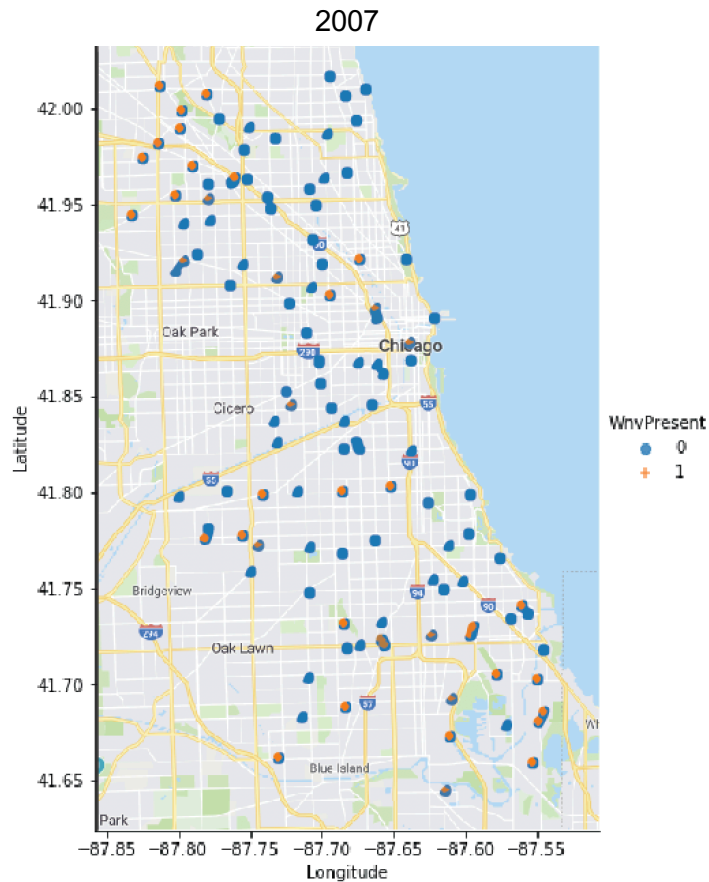
- Looking at **number of mosquitos** - number peaks on June and then reduced as the summer progresses

# I. Is there good data to work with?

- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**

# I. Is there good data to work with?

- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**

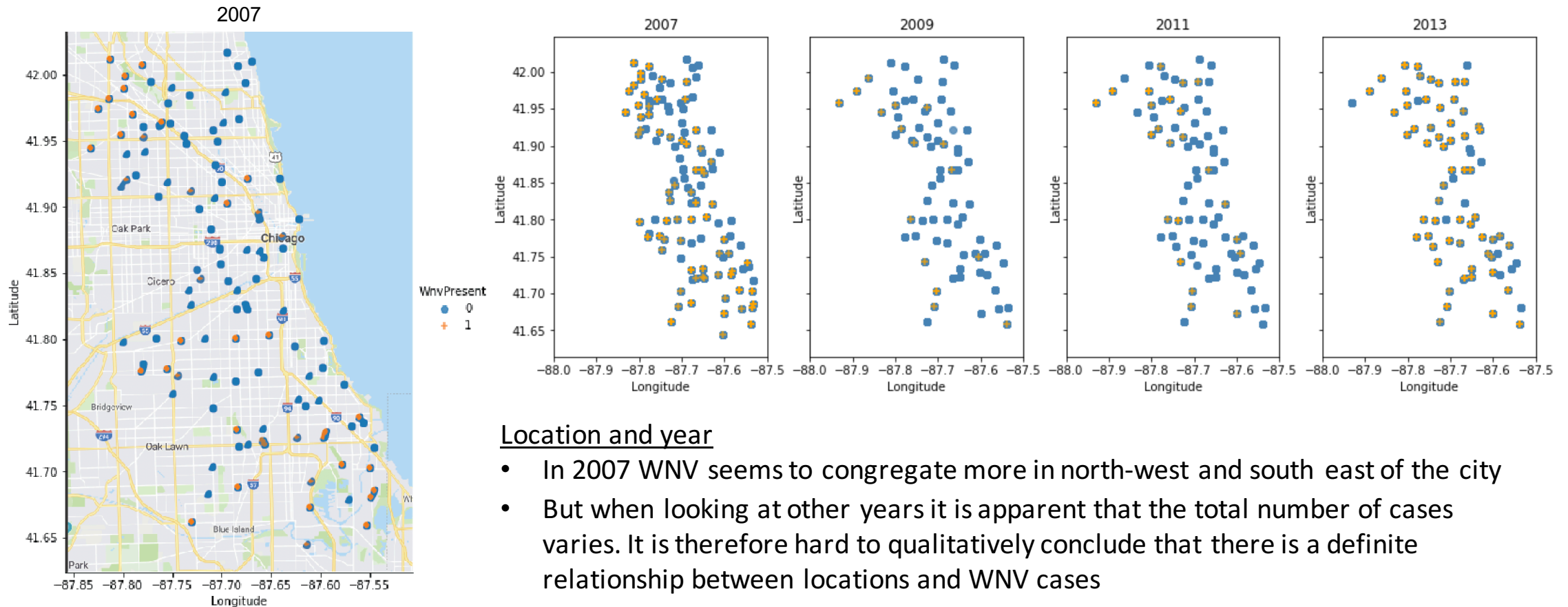


## Location and year

- In 2007 WNV seems to congregate more in north-west and south east of the city

# I. Is there good data to work with?

- Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?

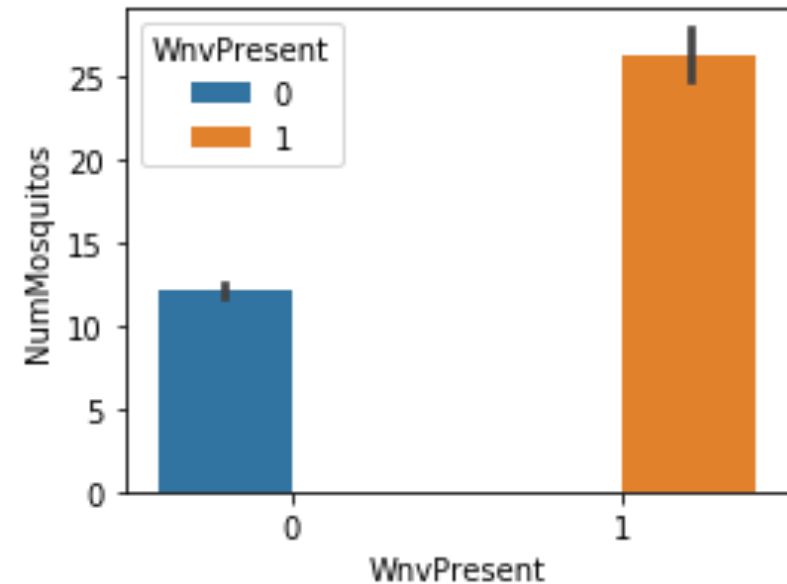


# I. Is there good data to work with?

- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**

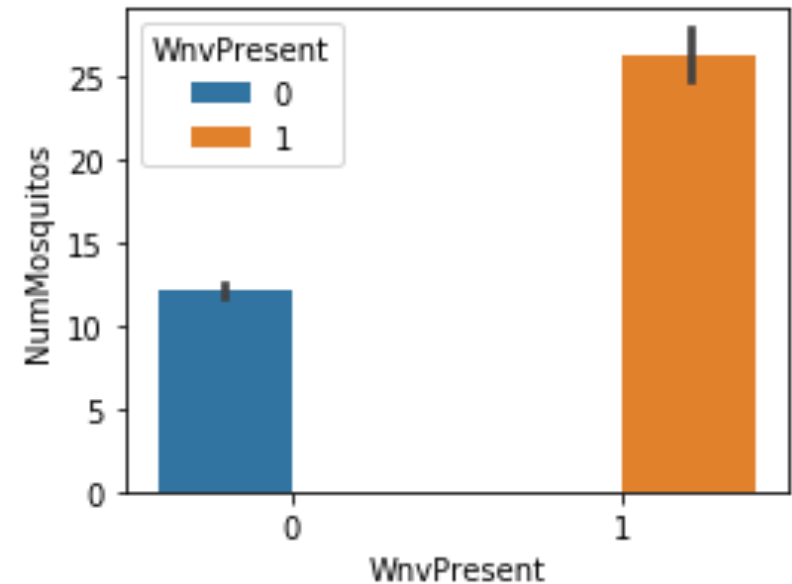
# I. Is there good data to work with?

- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**
- Expectedly, data shows that more mosquitos in the traps means more chances to find a mosquito with WNV.



# I. Is there good data to work with?

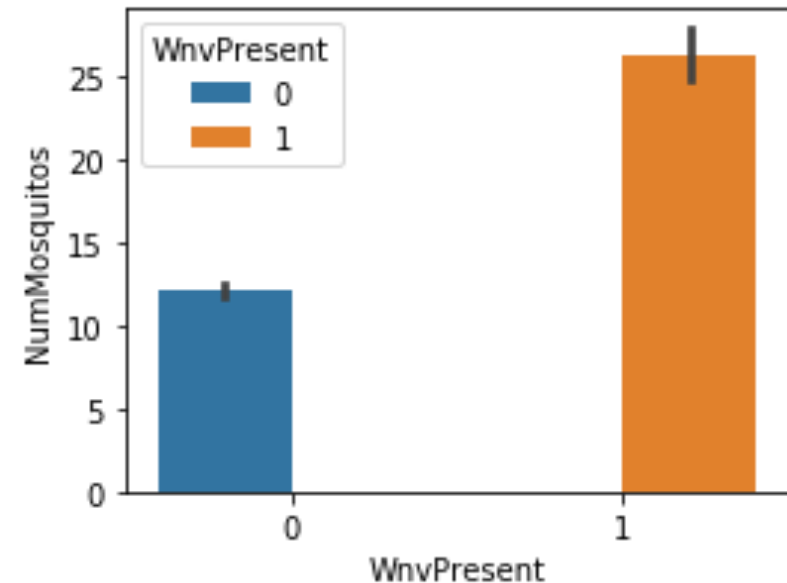
- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**
- Expectedly, data shows that more mosquitos in the traps means more chances to find a mosquito with WNV.
- Although it's a strong feature it is questionable whether we would be able to use it for prediction purposes. Feature might be "target-linked" meaning that when deploying our model to predict WNV, we wouldn't have the number of mosquitos available to us because it is determined along side the detection of the virus – hence “target link”
- in this case we might need to predict the number of mosquitos by itself
- Moreover, Chicago municipality's most straightforward mean of action against the virus might be to eradicate the mosquitos, which means that the ability to predict the mosquitos would be the real prediction question.



# I. Is there good data to work with?

- **Are the features informative to the goal of predicting West Nile Virus (WNV) occurrences?**

- Expectedly, data shows that more mosquitos in the traps means more chances to find a mosquito with WNV.
- Although it's a strong feature it is questionable whether we would be able to use it for prediction purposes. Feature might be "target-linked" meaning that when deploying our model to predict WNV, we wouldn't have the number of mosquitos available to us because it is determined along side the detection of the virus – hence “target link”
- in this case we might need to predict the number of mosquitos by itself
- Moreover, Chicago municipality's most straightforward mean of action against the virus might be to eradicate the mosquitos, which means that the ability to predict the mosquitos would be the real prediction question.



**Assumption: For this project's purposes let's assume for now that the 'NumMosquitos' feature IS NOT 'target-linked' and is an integral part of the features provided by Chicago Municipality to predict WNV occurrences.**



# I. Is there good data to work with?

- **Can we predict WNV occurrences from this basic data?**

# I. Is there good data to work with?

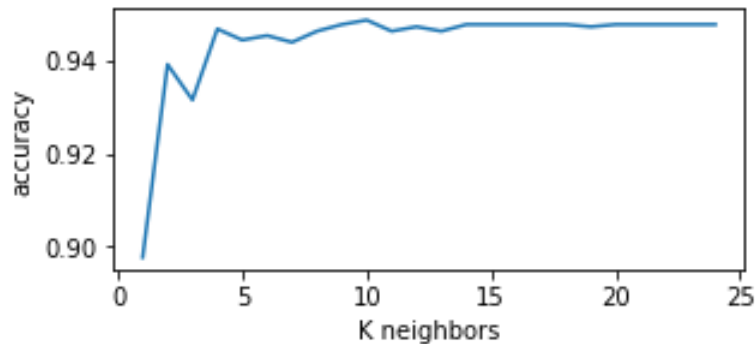
- **Can we predict WNV occurrences from this basic data?**

By running a classification model (e.g. “K Nearest Neighbors”(KNN)) we can get preliminary metrics on the ability to detect WNV successfully, with the basic “Train data” we have.

# I. Is there good data to work with?

- **Can we predict WNV occurrences from this basic data?**

By running a classification model (e.g. “K Nearest Neighbors”(KNN)) we can get preliminary metrics on the ability to detect WNV successfully, with the basic “Train data” we have.



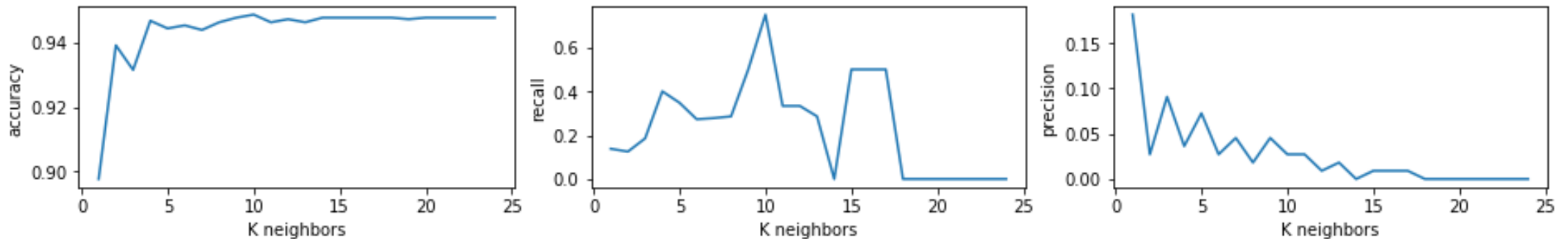
## TOP LEFT

Graph shows accuracy performance of 25 different KNN models. Accuracy starts really high (90%) and increases further when choosing “simpler” models (with higher bias by increasing K value). Accuracy is high since WNV occurrences are rare ~5%. Even a “dumb” model that predicts “No Virus” every time will reach this accuracy, therefore accuracy should not be our metric of choice

# I. Is there good data to work with?

- **Can we predict WNV occurrences from this basic data?**

By running a classification model (e.g. “K Nearest Neighbors”(KNN)) we can get preliminary metrics on the ability to detect WNV successfully, with the basic “Train data” we have.



## TOP MIDDLE & RIGHT

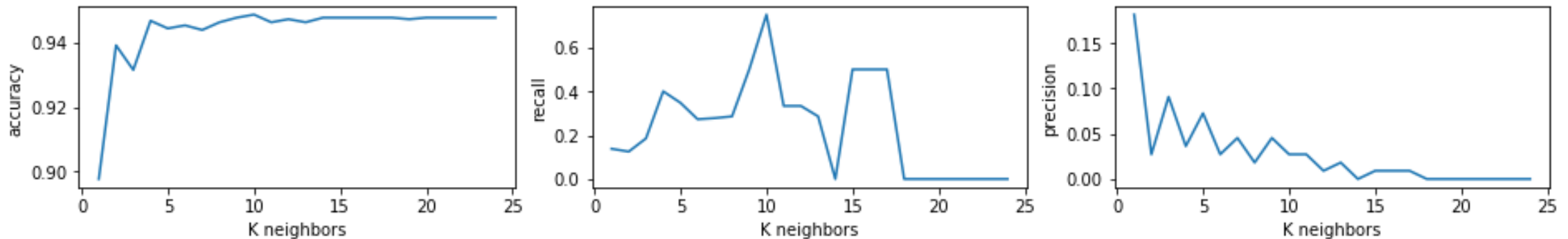
‘Sensitivity’ (Recall) and ‘Precision’ are the metrics that should be used. As we increase the bias (Ks) the sensitivity (the number of WNV that we were able to detect out of all WNV occurrences) is improving peaking at 75% at k=10 but that's at the expense of the precision which drops to around 4% at K=10. (Precision in this case is the number of cases we were able to detect correctly out of all the cases we claimed were WNV positive).

# I. Is there good data to work with?

- **Can we predict WNV occurrences from this basic data?**

By running a classification model (e.g. “K Nearest Neighbors”(KNN)) we can get preliminary metrics on the ability to detect WNV successfully, with the basic “Train data” we have.

	Predicted-negative	Predicted - positive
Actual Negative	1991	107
Actual Positive	1	3



## TABLE TOP RIGHT

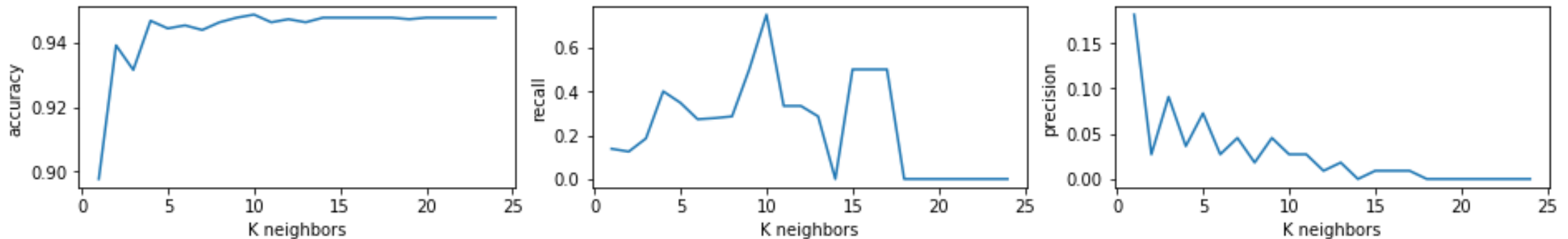
When looking at the actual numbers of actual cases and predictions, we see this tradeoff between recall and precision. When the K=10 we have 1 case of missed detections and 3 cases of correct detections which makes the recall peak at 75% but with 107 cases of Falsely predicting WNV, the precision plummets to ~4%.

# I. Is there good data to work with?

- **Can we predict WNV occurrences from this basic data?**

By running a classification model (e.g. “K Nearest Neighbors”(KNN)) we can get preliminary metrics on the ability to detect WNV successfully, with the basic “Train data” we have.

	Predicted-negative	Predicted - positive
Actual Negative	1991	107
Actual Positive	1	3



## TABLE TOP RIGHT

When looking at the actual numbers of actual cases and predictions, we see this tradeoff between recall and precision. When the K=10 we have 1 case of missed detections and 3 cases of correct detections which makes the recall peak at 75% but with 107 cases of Falsely predicting WNV, the precision plummets to ~4%.

**To conclude, with the basic dataset we can get at most, a model that is deceptively sensitive and entirely not precise.**

## II. How can we modify the data to work for us?

- **Cleaning and standardizing features across datasets**

Spray Dataset

## II. How can we modify the data to work for us?

- **Cleaning and standardizing features across datasets**

### Spray Dataset

- We aim to enrich our basic dataset with other available datasets.
- Datasets could be merged according to location and time.  
Merging the datasets (e.g. 'Spray data' with 'Train data') should allow us to assess its usefulness to our purposes



## II. How can we modify the data to work for us?

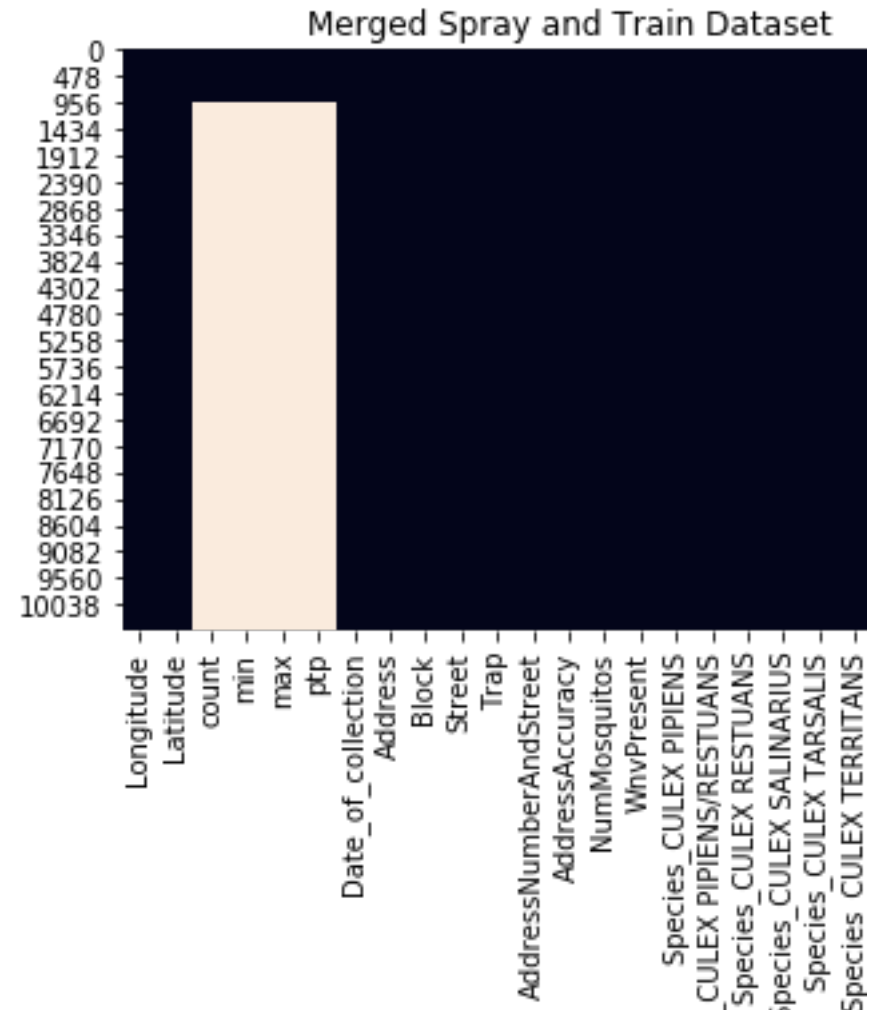
### ▪ Cleaning and standardizing features across datasets

#### Spray Dataset

- We aim to enrich our basic dataset with other available datasets.
- Datasets could be merged according to location and time. Merging the datasets (e.g. 'Spray data' with 'Train data') should allow us to assess its usefulness to our purposes

#### RIGHT

Heatmap shows that only 9% of trap observations have matching spray data. (top 962 uniformly black rows).



## II. How can we modify the data to work for us?

- **Engineering more informative features from the basic features given**

Spray Dataset

## II. How can we modify the data to work for us?

- **Engineering more informative features from the basic features given**

### Spray Dataset

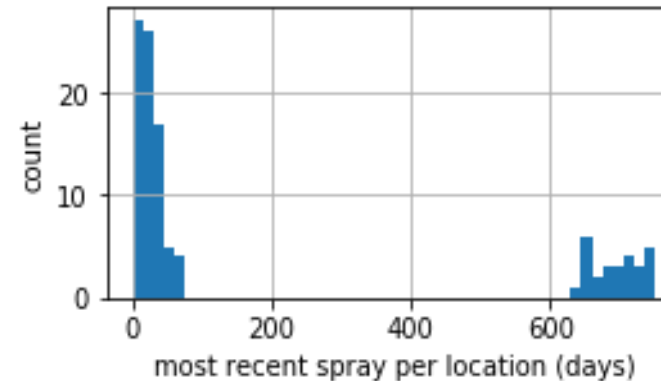
- **Assumption** – Observations with spray data (9% of data) might be useful. Our initial assumption is that recently sprayed areas would most likely reduce frequency in WNV for the same season.
- **Significance** – It is in Chicago Municipality and CDPH interest to better allocate spraying efforts, we should filter out the effect of recent spraying from our predictive model, as it is target-linked.
- **Method** - find observations with locations that have been sprayed in the past 150 days ( length of a season of collection from May to Oct), and exclude them from the dataset

## II. How can we modify the data to work for us?

### ▪ Engineering more informative features from the basic features given

#### Spray Dataset

- **Assumption** – Observations with spray data (9% of data) might be useful. Our initial assumption is that recently sprayed areas would most likely reduce frequency in WNV for the same season.
- **Significance** – It is in Chicago Municipality and CDPH interest to better allocate spraying efforts, we should filter out the effect of recent spraying from our predictive model, as it is target-linked.
- **Method** - find observations with locations that have been sprayed in the past 150 days ( length of a season of collection from May to Oct), and exclude them from the dataset



#### LEFT

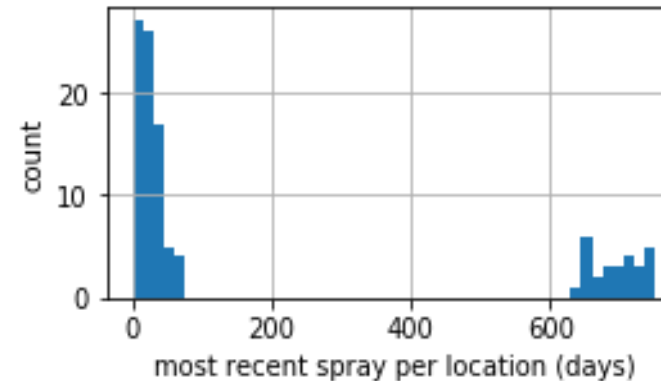
Only 0.8% of trap observations (79) experienced spraying during the same season (last 150 days)

## II. How can we modify the data to work for us?

### ▪ Engineering more informative features from the basic features given

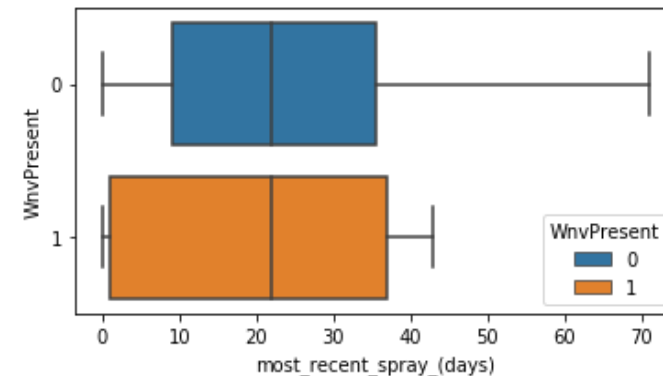
#### Spray Dataset

- **Assumption** – Observations with spray data (9% of data) might be useful. Our initial assumption is that recently sprayed areas would most likely reduce frequency in WNV for the same season.
- **Significance** – It is in Chicago Municipality and CDPH interest to better allocate spraying efforts, we should filter out the effect of recent spraying from our predictive model, as it is target-linked.
- **Method** - find observations with locations that have been sprayed in the past 150 days ( length of a season of collection from May to Oct), and exclude them from the dataset



LEFT

Only 0.8% of trap observations (79) experienced spraying during the same season (last 150 days)



LEFT

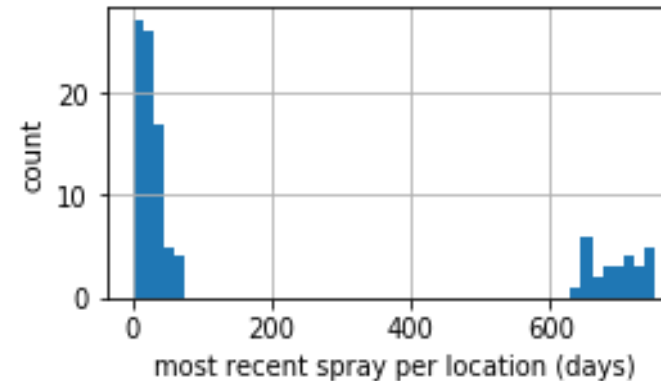
Surprisingly, we found no significant relationship (on average) between number of days since last spray and occurrences of WNV as expected

## II. How can we modify the data to work for us?

### ▪ Engineering more informative features from the basic features given

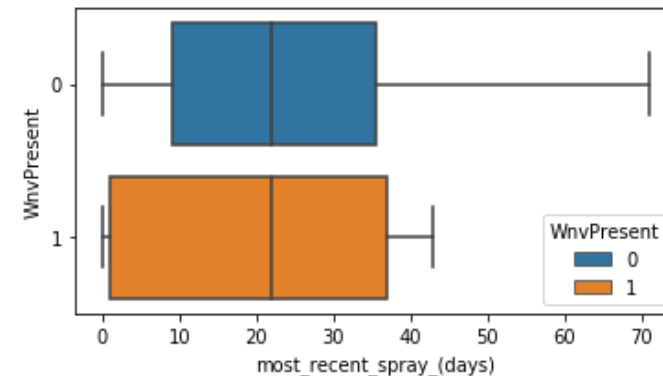
#### Spray Dataset

- **Assumption** – Observations with spray data (9% of data) might be useful. Our initial assumption is that recently sprayed areas would most likely reduce frequency in WNV for the same season.
- **Significance** – It is in Chicago Municipality and CDPH interest to better allocate spraying efforts, we should filter out the effect of recent spraying from our predictive model, as it is target-linked.
- **Method** - find observations with locations that have been sprayed in the past 150 days ( length of a season of collection from May to Oct), and exclude them from the dataset



LEFT

Only 0.8% of trap observations (79) experienced spraying during the same season (last 150 days)



LEFT

Surprisingly, we found no significant relationship (on average) between number of days since last spray and occurrences of WNV as expected

**Conclusion – The Spray dataset’s low number of merged observations and lack of significant relationship to WNV occurrences deems it un-informative to our purposes to enrich the basic Train dataset**

## II. How can we modify the data to work for us?

- **Engineering more informative features from the basic features given**

Weather Dataset

## II. How can we modify the data to work for us?

- **Engineering more informative features from the basic features given**

### Weather Dataset

- **Assumption** – Hot and dry conditions are more favorable for WNV than cold and wet. Related features should be a valuable resource for WNV prediction. How should we best use them?
- **Rational** – These conditions may need to be stable for extended periods of time (e.g. to drive Mosquito and WNV generation). We should engineer the features to reflect this
- **Method** – match every trap collection (observation) with features that summarize the weather over the previous 14 days (chosen arbitrarily)



## II. How can we modify the data to work for us?

### ▪ Engineering more informative features from the basic features given

#### Weather Dataset

- **Assumption** – Hot and dry conditions are more favorable for WNV than cold and wet. Related features should be a valuable resource for WNV prediction. How should we best use them?
  1. Related features of interest:  
['Tmax', 'Tmin', 'Tavg', 'DewPoint', 'WetBulb', 'Heat', 'Cool', 'PrecipTotal', 'StnPressure', 'ResultSpeed', 'ResultDir', 'AvgSpeed', 'weather\_type\_Norm']
- **Rational** – These conditions may need to be stable for extended periods of time (e.g. to drive Mosquito and WNV generation). We should engineer the features to reflect this
- **Method** – match every trap collection (observation) with features that summarize the weather over the previous 14 days (chosen arbitrarily)

#### TOP

- Weather feature engineering example.

## II. How can we modify the data to work for us?

### ▪ Engineering more informative features from the basic features given

#### Weather Dataset

- **Assumption** – Hot and dry conditions are more favorable for WNV than cold and wet. Related features should be a valuable resource for WNV prediction. How should we best use them?
- **Rational** – These conditions may need to be stable for extended periods of time (e.g. to drive Mosquito and WNV generation). We should engineer the features to reflect this
- **Method** – match every trap collection (observation) with features that summarize the weather over the previous 14 days (chosen arbitrarily)

#### 1. Related features of interest:

['Tmax', 'Tmin', 'Tavg', 'DewPoint', 'WetBulb', 'Heat', 'Cool', 'PrecipTotal', 'StnPressure', 'ResultSpeed', 'ResultDir', 'AvgSpeed', 'weather\_type\_Norm']

#### 2. Engineered features:

['14\_days\_AvgSpeed.mean', '14\_days\_AvgSpeed.std', '14\_days\_AvgSpeed.median', '14\_days\_AvgSpeed.outliers']

#### TOP

- Weather feature engineering example.
- The “Outliers” feature for instance should be informative on stability of conditions. Every feature of interest expands into 5 new engineered features

## II. How can we modify the data to work for us?

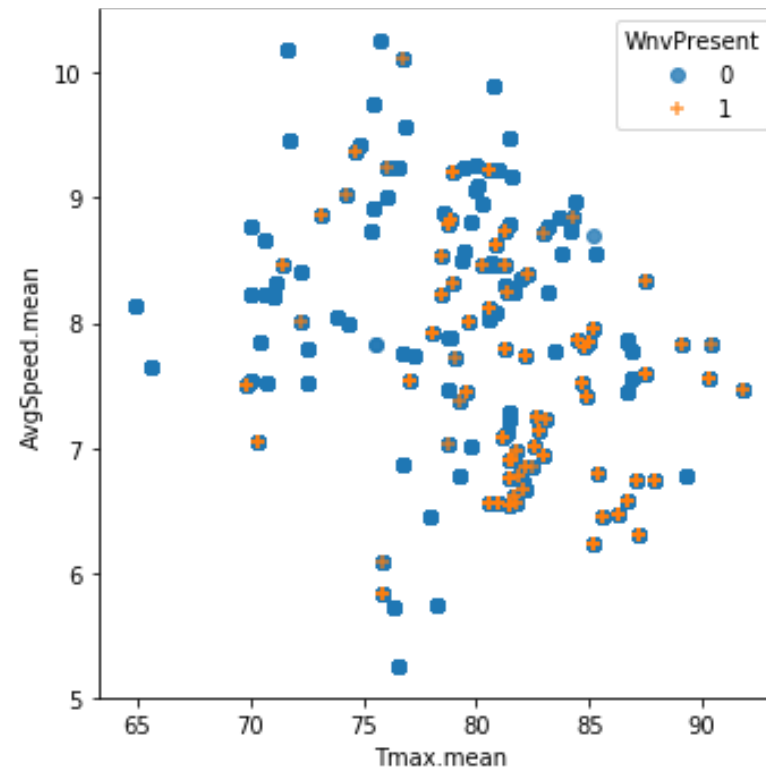
- **Engineering more informative features from the basic features given**

## II. How can we modify the data to work for us?

- **Engineering more informative features from the basic features given**

### RIGHT

Distribution of WNV(+)/(-), depended on averaged wind speed and maximum temperatures (both averaged in the 14 days prior to trap collection). We can see that the virus is more frequent in the lower right part of the figure - in higher temperatures and lower wind speeds.



## II. How can we modify the data to work for us?

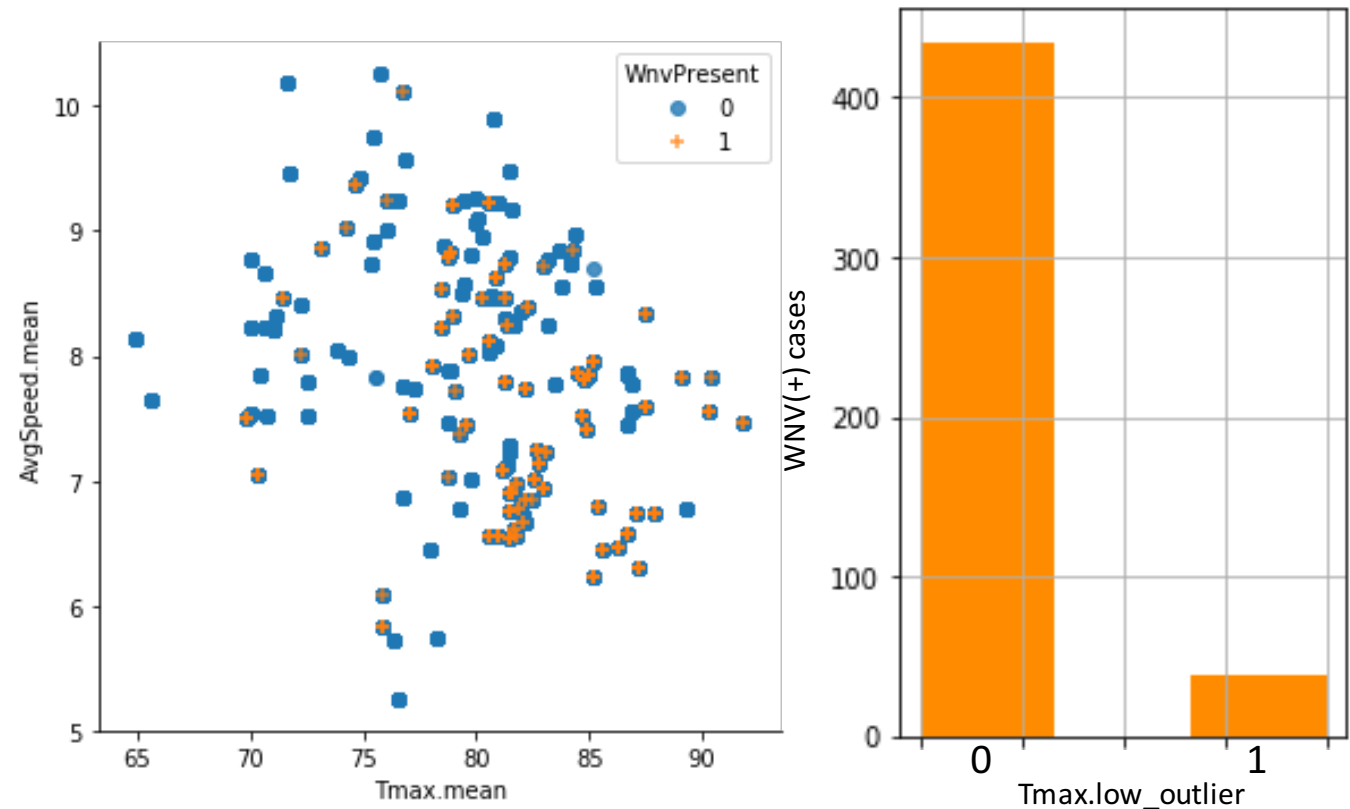
- **Engineering more informative features from the basic features given**

### RIGHT

Distribution of WNV(+)/(-), depended on averaged wind speed and maximum temperatures (both averaged in the 14 days prior to trap collection). We can see that the virus is more frequent in the lower right part of the figure - in higher temperatures and lower wind speeds.

### FAR RIGHT

WNV(+) cases, depended on whether there was a particularly low T.max in the 2 weeks prior to collection (a "low outlier"). There is a 10 X less chance for WNV(+) to appear, even if there is only one relatively cold day (as compared to when the weather is stable).



## II. How can we modify the data to work for us?

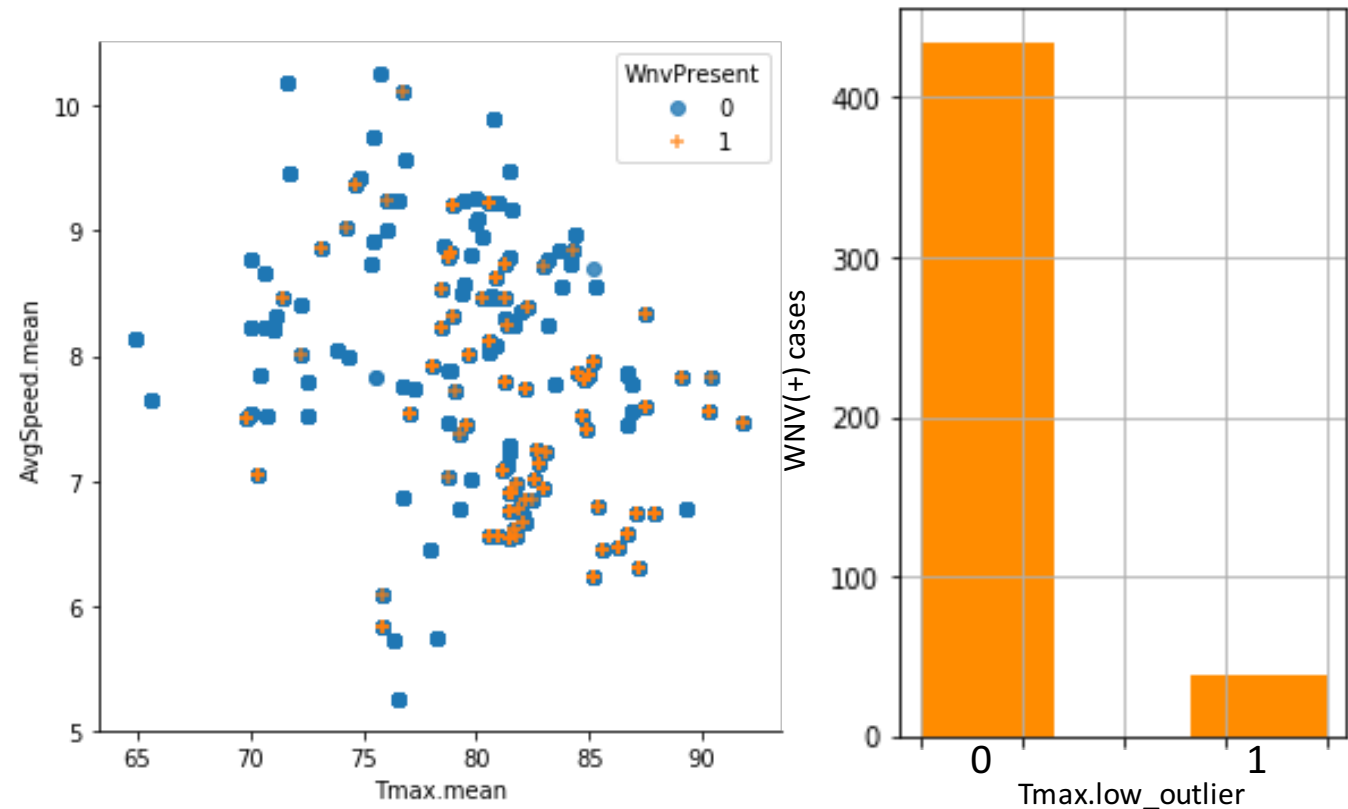
- **Engineering more informative features from the basic features given**

### RIGHT

Distribution of WNV(+)/(-), depended on averaged wind speed and maximum temperatures (both averaged in the 14 days prior to trap collection). We can see that the virus is more frequent in the lower right part of the figure - in higher temperatures and lower wind speeds.

### FAR RIGHT

WNV(+) cases, depended on whether there was a particularly low T.max in the 2 weeks prior to collection (a "low outlier"). There is a 10 X less chance for WNV(+) to appear, even if there is only one relatively cold day (as compared to when the weather is stable).



**Conclusion - Newly Engineered features are highly informative and should provide prediction power once we model**

# III. Building a statistical model to meet our goal of optimal prediction of WNV

## ■ What is the best model to use?

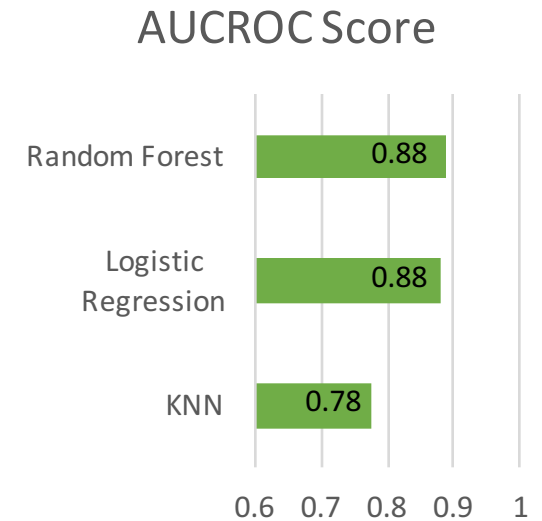
### Final step before modelling

- Building a model based on a dataset where the minority class (WNV rate)  $\sim 5\%$  is not valuable
- Any given model would not be able to learn from it due to the overpowering effect of the majority class in the dataset
- We did therefore fabricate WNV cases to balance the train Dataset (using SMOTE)

# III. Building a statistical model to meet our goal of optimal prediction of WNV

## ■ What is the best model to use?

- We ran several classification models (i.e. Logistic Regression, KNN, Random Forest (RF)) on the dataset to get initial idea about the performance of the models.
- We chose AUCROC as a metric for performance due to its robustness (which we'll discuss in next slides).
- Although it is not better than Logistic Regression, RF is generally a more robust model ( has more options to tune it ). So there is a better potential to improve its performance further

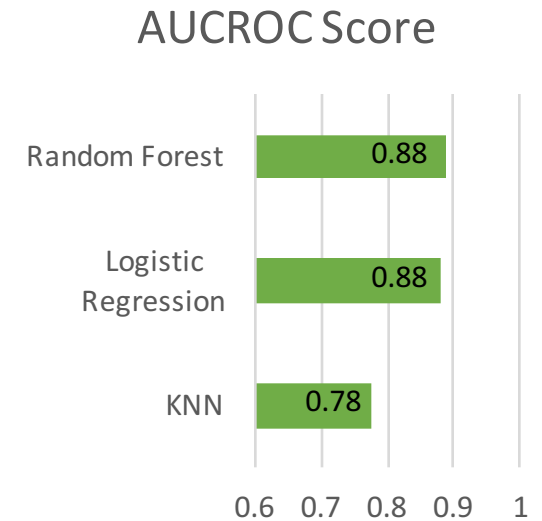




# III. Building a statistical model to meet our goal of optimal prediction of WNV

## ■ What is the best model to use?

- We ran several classification models (i.e. Logistic Regression, KNN, Random Forest (RF)) on the dataset to get initial idea about the performance of the models.
- We chose AUCROC as a metric for performance due to its robustness (which we'll discuss in next slides).
- Although it is not better than Logistic Regression, RF is generally a more robust model ( has more options to tune it ). So there is a better potential to improve its performance further



**Conclusion – Random Forest Classifier is the model of choice**

# III. Building a statistical model to meet our goal of optimal prediction of WNV

## ▪ How can we optimize the model?

- **Significance** - We chose AUCROC\*, because it allows us to tune for the best model across all sensitivities and precisions. This means that we can offer a prediction quality that would best suit our client depended on its needs
- **Assumption** – Chicago Municipality and CDPH Would like to pinpoint the highest number of WNV cases possible rather than avoid a False alarm (of detecting WNV where there isn't). This is due the high cost of an undetected WNV (due to health implications) as compared to the cost involving spraying an area due to false alarm.
- **Method** – We will iterate between several hyper-parameters of the model. Pick the ones that give us the highest AUCROC score, and then set our recommendations to to Chicago Municipality and CDPH, based on the sensitivity & precision levels we are able to provide

\* Area under the curve of ROC

# III. Building a statistical model to meet our goal of optimal prediction of WNV

- **What is the performance of the model we are offering?**

# III. Building a statistical model to meet our goal of optimal prediction of WNV

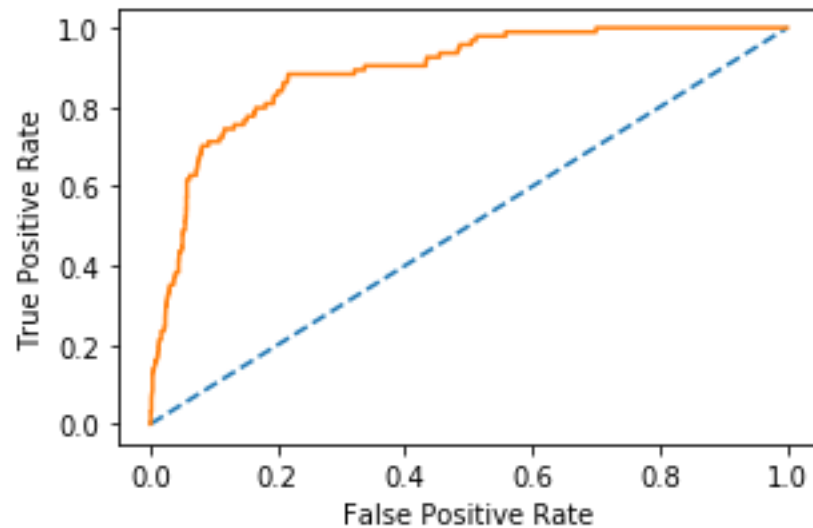
- **What is the performance of the model we are offering?**

AUCROC Score:  
0.89  
(1% higher after  
model tuning)

# III. Building a statistical model to meet our goal of optimal prediction of WNV

- **What is the performance of the model we are offering?**

AUCROC Score:  
0.89  
(1% higher after  
model tuning)



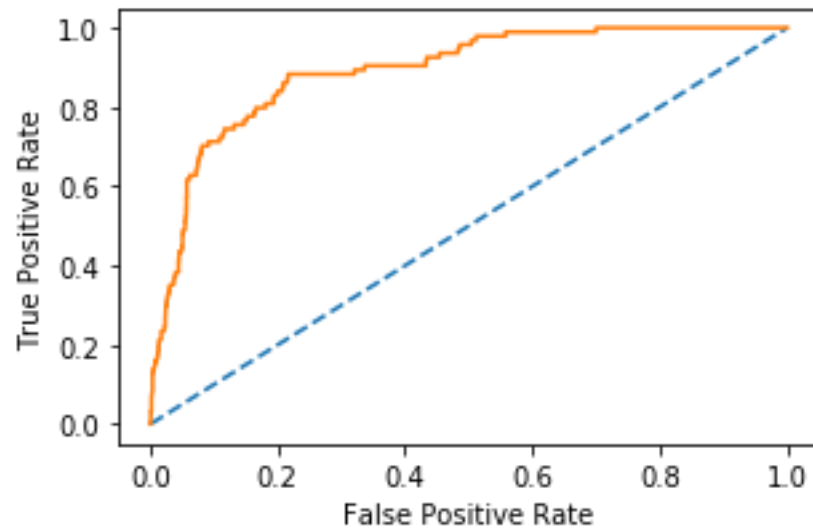
## TOP

ROC curve in orange shows the range of different sensitivities we could choose from. Since we are interested in high sensitivity, we can choose a relatively high FP rate (e.g. around 0.5 will give us more than 80% sensitivity).

# III. Building a statistical model to meet our goal of optimal prediction of WNV

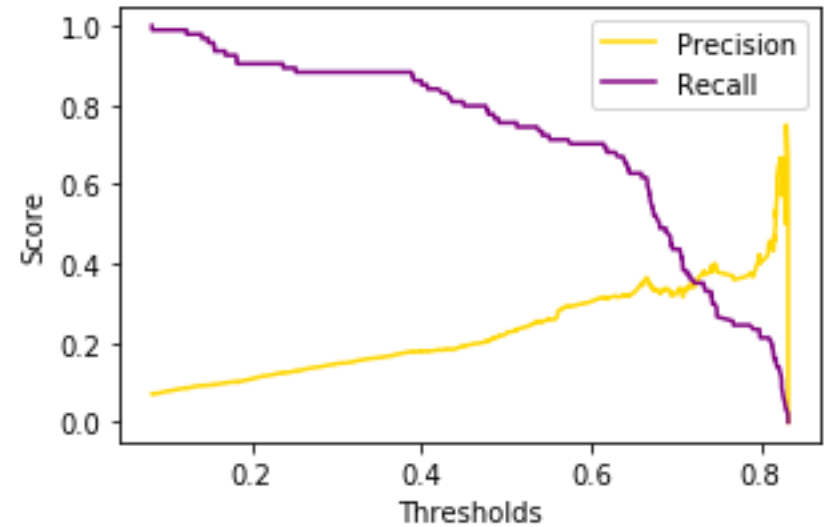
## ■ What is the performance of the model we are offering?

AUCROC Score:  
0.89  
(1% higher after  
model tuning)



### TOP

ROC curve in orange shows the range of different sensitivities we could choose from. Since we are interested in high sensitivity, we can choose a relatively high FP rate (e.g. around 0.5 will give us more than 80% sensitivity).



### TOP

By plotting recall and precision, the tradeoff is more apparent between sensitivity (recall) and precision.

# III. Building a statistical model to meet our goal of optimal prediction of WNV

- **Are the modification efforts of the data justified?**

# III. Building a statistical model to meet our goal of optimal prediction of WNV

## ■ Are the modification efforts of the data justified?

To validate that the feature engineering efforts were fruitful, we tested the ability to predict WNV performance depended on the dataset, at 4 different stages:

1. 'Baseline.imbalanced' - the clean raw dataset
2. 'Baseline' - the clean raw dataset after balancing it
3. 'sptrainW' – the dataset enriched with spray and weather data
4. 'sptrainW\_14 \_day' – the final feature engineered dataset used for the final modeling



# III. Building a statistical model to meet our goal of optimal prediction of WNV

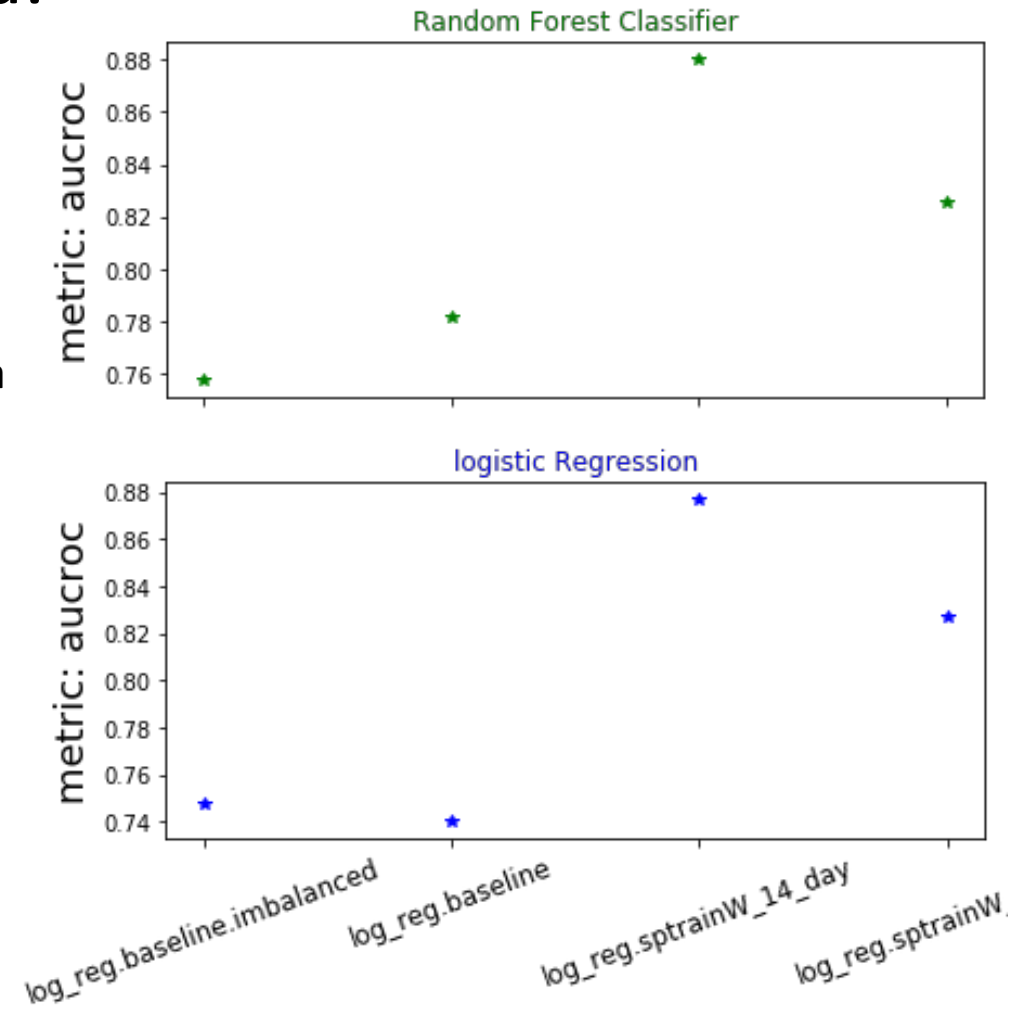
## ■ Are the modification efforts of the data justified?

To validate that the feature engineering efforts were fruitful, we tested the ability to predict WNV performance depended on the dataset, at 4 different stages:

1. 'Baseline.imbalanced' - the clean raw dataset
2. 'Baseline' - the clean raw dataset after balancing it
3. 'sptrainW' – the dataset enriched with spray and weather data
4. 'sptrainW\_14 \_day' – the final feature engineered dataset used for the final modeling

### RIGHT

the more the dataset is engineered and enriched, the higher the score which signifies the WNV prediction performance. The result is replicated in 2 different models (Logistic regression & Random Forest Classifier)



# III. Building a statistical model to meet our goal of optimal prediction of WNV

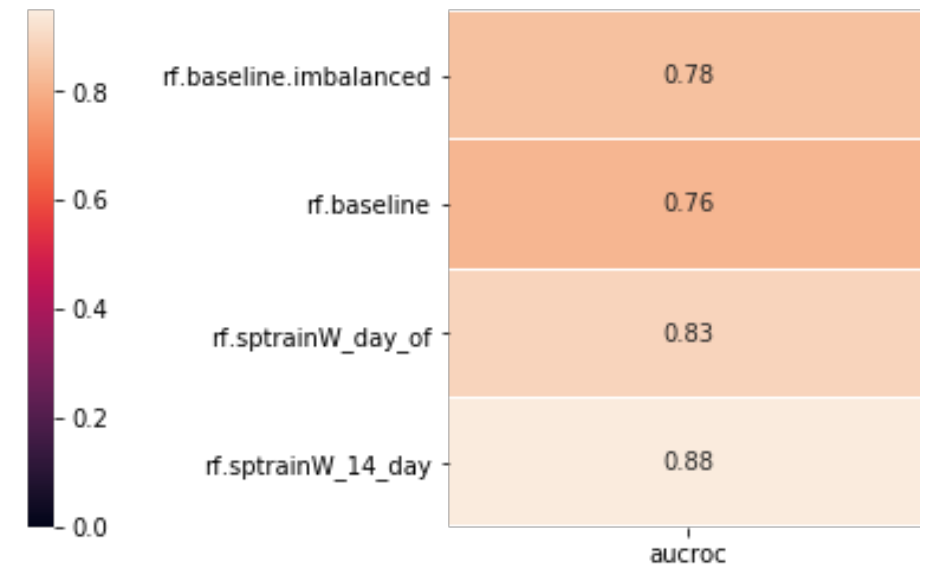
## ■ Are the modification efforts of the data justified?

To validate that the feature engineering efforts were fruitful, we tested the ability to predict WNV performance depended on the dataset, at 4 different stages:

1. 'Baseline.imbalanced' - the clean raw dataset
2. 'Baseline' - the clean raw dataset after balancing it
3. 'sptrainW' – the dataset enriched with spray and weather data
4. 'sptrainW\_14 \_day' – the final feature engineered dataset used for the final modeling

### RIGHT

the more the dataset is engineered and enriched, the higher the score which signifies the WNV prediction performance. The result is replicated in 2 different models (Logistic regression & Random Forest Classifier)



## IV. Summary of the proposed model and final recommendations

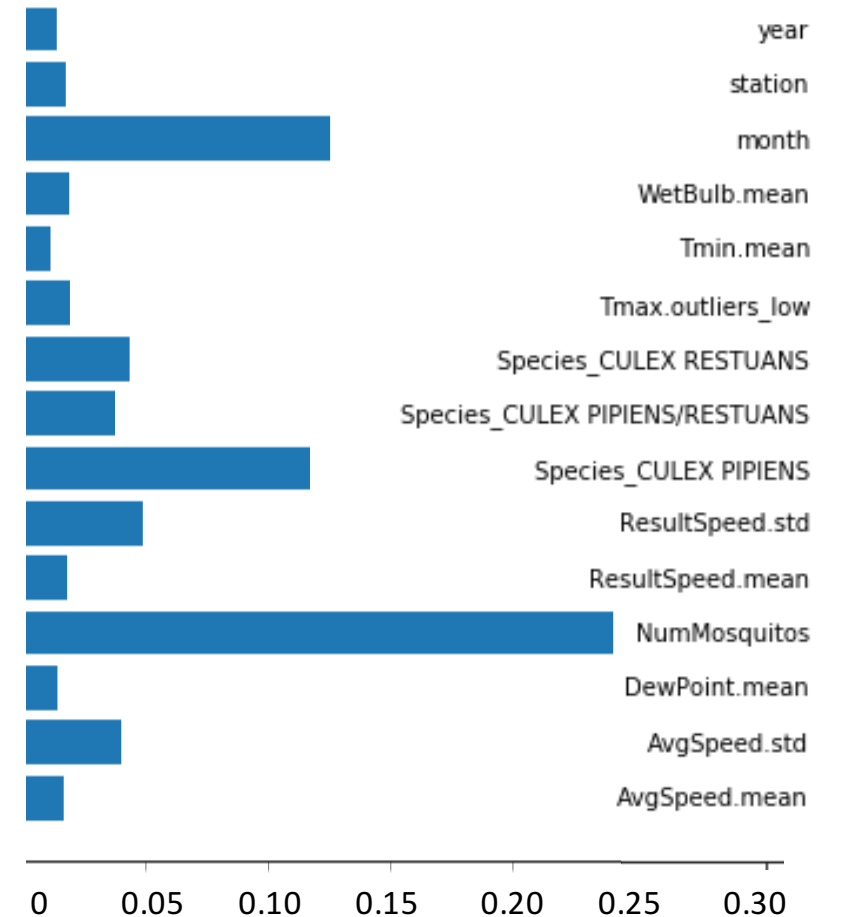
- **What the model is revealing to us about WNV factors?**

# IV. Summary of the proposed model and final recommendations

## ▪ What the model is revealing to us about WNV factors?

### RIGHT

- Not surprisingly, the most important factor is Number of Mosquitos
- Whether the specie is CULEX PIPIENS or not is important, as well as which month it is (e.g. peak of summer)
- 8 out of 13 of the most important factors are engineered weather features
- Weather factors include wind speed (most important probably due to strong winds effecting mosquito activity)
- Next in importance are humidity factors (i.e. wetbulb & dewpoint)
- Significantly cold day during otherwise stable 2 weeks is important (probably because it interrupts mosquito/virus activity)
- Some years have more occurrences then other years (probably due to specifically cold/hot years)

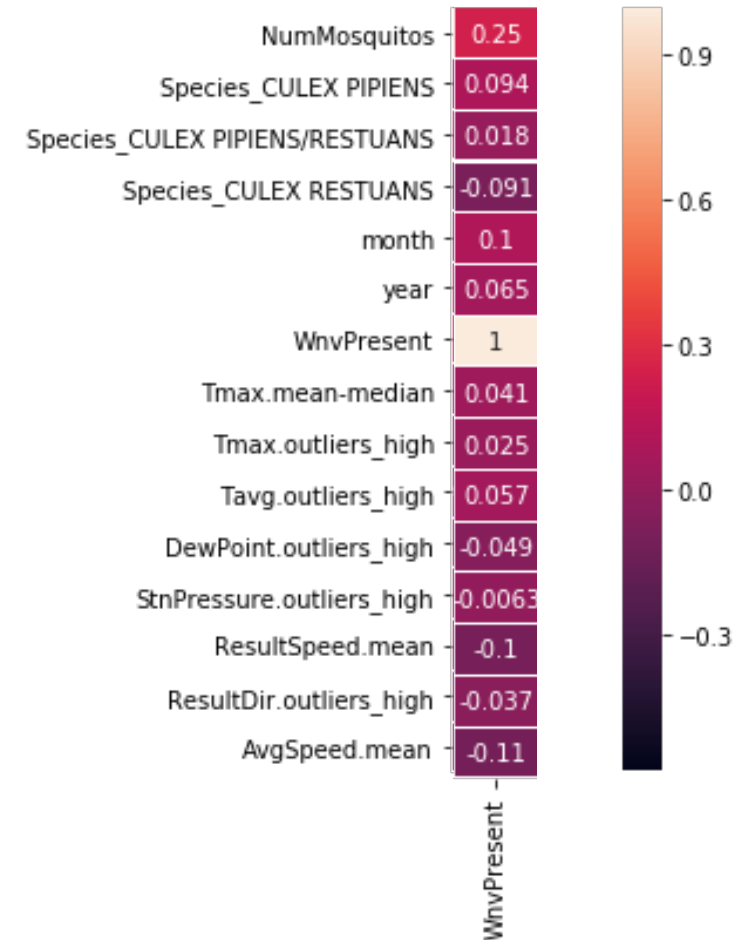


# IV. Summary of the proposed model and final recommendations

## ▪ What the model is revealing to us about WNV factors?

### RIGHT

- A heatmap of correlations can give a better sense of how these factors are linked to WNV. It allows us to see the directionality of the relationship:
- All factors involving wind have negative correlation with WNV response i.e. more WINDY -> Less WNV.
- Although a small effect, when we look at Dewpoint, we see that the more HUMID -> the Less WNV (which is aligned with known facts about WNV which prefers dry conditions)
- “Culex\_Pipiens” specie is the most indicative of WNV response, “Culex Restuans” indicative of NO-WNV



# IV. Summary of the proposed model and final recommendations

## ▪ What the model is revealing to us about WNV factors?

### Summary of factors relationship to WNV:

(+++)  
strong positive relationship

(++)  
moderate positive relationship

(-)  
mild negative relationship

Factor	Related to WNV
Number of Mosquitos	+++
CULEX_PIPPIENS (specie)	++
Month	++
High Temperatures	+
CULEX_RESTUANS (specie)	-
Wind	-
Humidity	-

# IV. Summary of the proposed model and final recommendations

- **Final actionable recommendations for Chicago Municipality and Department of Public Health**

# IV. Summary of the proposed model and final recommendations

## ▪ **Final actionable recommendations for Chicago Municipality and Department of Public Health**

- The sensitivity of the model we are offering to detect WNV could range from detecting 60% to detecting 100%.
- Depended on the cost of false alarm (e.g. spraying areas that are not infested), we could offer the right package for Chicago Municipality and CDPH purposes.
- We offer 2 possible packages:
  - i. If it's affordable to "over-spray" 4 times the number of necessary spraying (i.e. only 1 out of 5 spraying efforts are justifiable), we predict eradication of 90% of WNV of occurrences in Chicago (lowering its rate from 5% to 0.5%).
  - ii. Alternatively If it's affordable to "over-spray" only 2.5 times (i.e. 2 out of 5 spraying efforts are justifiable), then we predict eradication of 60% of WNV occurrences (lowering its rate from 5% to 2%).
- We recommend to take into consideration the inhabitants population density of the areas in order to asses the importance of eradicating the mosquitos. This could provide important clues about the right choice of over-spraying and WNV detection rate discussed previously.



# IV. Summary of the proposed model and final recommendations

## ▪ **Final actionable recommendations for Chicago Municipality and Department of Public Health**

- We suggest to take preemptive measures as well as spraying infested locations. Since June has the highest occurrences of WNV, whilst not being the hottest, driest and least windy, we could assume that the high occurrences are due to reactive measures to eradicate mosquitos been taken too late in the season (only after locations become infested).
- We propose to invest efforts in prevention, starting from increasing awareness of the general population (e.g. to not leave exposed untreated water such as ponds and pools in their yard)
- Additionally, since one specie (Culex Pipiens) stands out as a carrier with higher rates of WNV, We suggest to increase efforts in preventing the proliferation of this specie in particular. This could be done with more specific and environmentally friendly measures (e.g. "targeted biological mosquito control") which would make the process of lowering occurrences more efficient and less costly

## Supplementary Slide - Project Discussion and Final Notes

- AUC under the ROC curve is ~89% so the model is robust enough to allow flexible choice of thresholds (choosing the most fitting recall vs precision rates)
- The model is based on the assumption that the number of mosquitos is a valid info that could be collected and used for prediction. The model's performance would change if we exclude the 'Number of Mosquitos' feature. At that scenario we will use all other features to predict the number of mosquitos, and then use that prediction as a new feature in a new classification model, to make a new prediction for the WNV occurrences. We can expect the performance of this model to be lower.
- The spray data was implemented with the hypothesis that recently sprayed sites would effect the occurrences of WNV. after EDA, occurrences seem to not vary (on average) and since the number of observations taken from recently sprayed sites were low (79), these observations were not excluded from the dataset with the rational that they would not effect the model's performance.
- The data could be enriched further with publically available data, e.g. Implementing data on locations of exposed water reservoirs, lakes and other bodies of fresh water which might be correlated to increased mosquito population.

***\* All rights reserved to Eran Schenker***