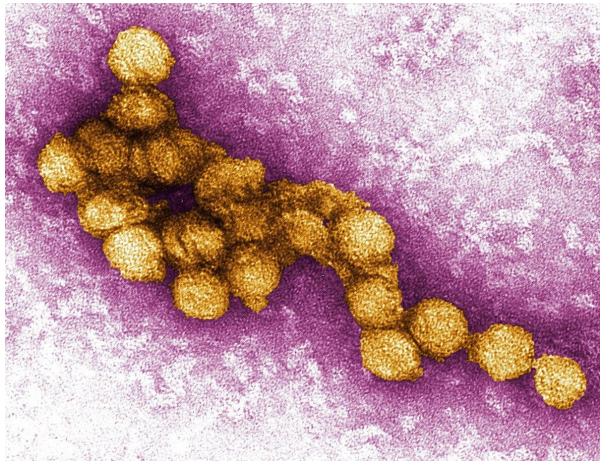
West Nile Virus Prediction

Lee Zi Xin, David Li, Xavier Rigoulet, Rachel Lim



Disease and Treatment Agency

SOCIETAL CURES IN EPIDEMIOLOGY AND NEW CREATIVE ENGINEERING

Agenda

1. Business Problem
2. Methodology
3. Exploratory Data Analysis (EDA)
4. Model & Performance
5. Limitations
6. Cost Benefit Analysis
7. Conclusion & Recommendations

Business Problem

A. Due to the recent epidemic of West Nile Virus (WNV) in the Windy City also known as Chicago, the Department of Public Health has set up a surveillance and control system to collect the samples of:

1. **mosquitos** population
2. **weather** conditions
3. **spray** information

*This information will be used to derive a plan to **curb the virus**.*

B. Disease And Treatment Agency, division of Societal Cures In Epidemiology and New Creative Engineering (DATA-SCIENCE) is tasked to

1. **Analyze** the collected data
2. **Predict** when and where the WNV could be present
3. **Provide** a recommendation of a suitable plan

West Nile Virus Information

- West Nile Virus natural hosts: **birds** and **mosquitoes**.
- West Nile Virus in a large number of mosquito species, but **the most significant** for viral transmission are *Culex* species that feed on birds, including *Culex pipiens*, *C. restuans*, *C. salinarius*, *C. quinquefasciatus*, *C. nigripalpus*, *C. erraticus* and *C. tarsalis* (Ref: https://en.wikipedia.org/wiki/West_Nile_virus)
- Most West Nile virus infections occur in warm weather, when mosquitoes are active. The incubation period* ranges from 2 to 14 days.

*Incubation period: period between the mosquito bite and the first symptoms of the illness

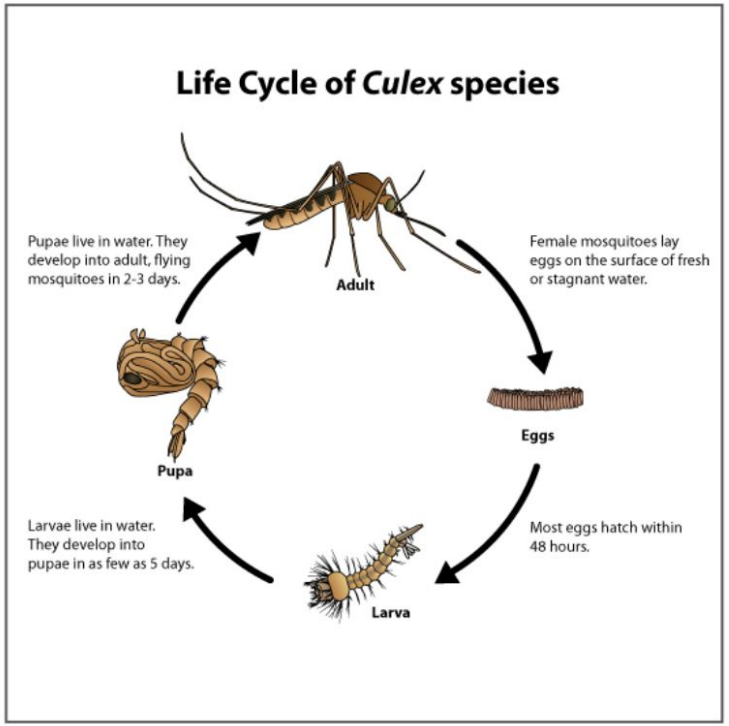
Main goal is to

1. Devise an accurate method of **predicting outbreak of West Nile virus in mosquitoes**

2. Help Chicago devise a plan to **allocate resources towards preventing transmission** more efficiently and effectively

Mosquito Life Cycle & Thriving Factors

It takes about 7-10 days for an egg to develop into an adult mosquito.



Thriving Factors

- Temperature
  - Typically thrives in about **80 Deg F**
  - Higher winter temperatures and warmer spring may lead to larger summer mosquito populations, increasing the risk of West Nile Virus outbreak.
- Precipitation
  - Rainfall may also drive mosquito replication rate and affect the seasonality and geographic variations of the virus.
- Wind
  - Likewise, wind is another environmental factor that serves as a dispersal mechanism for mosquitoes

https://www.cdc.gov/mosquitoes/about/life-cycles/index.html

Methodology

Problem	EDA	Feature Engineering	Modelling	Conclusion
Defining the problem: **Prediction of the West Nile Virus**	1. Data Cleaning 2. Imputation 3. Data Analysis	1. New Features 2. Drop Features	Loop through: 1. Feature engineering 2. Models 3. Parameters tuning 4. Metrics	1. Evaluation 2. Model Limitations 3. Cost-Benefits Analysis 4. Next Steps

Data

```
TRAIN.CSV DATA SET
**********************************************
Rows: 10506      Columns: 12
==============================================

train.csv Data Set.info()
-----------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10506 entries, 0 to 10505
Data columns (total 12 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Date                    10506 non-null  object
 1   Address                 10506 non-null  object
 2   Species                 10506 non-null  object
 3   Block                   10506 non-null  int64
 4   Street                  10506 non-null  object
 5   Trap                    10506 non-null  object
 6   AddressNumberAndStreet  10506 non-null  object
 7   Latitude                10506 non-null  float64
 8   Longitude               10506 non-null  float64
 9   AddressAccuracy         10506 non-null  int64
 10  NumMosquitos            10506 non-null  int64
 11  WnvPresent              10506 non-null  int64
dtypes: float64(2), int64(4), object(6)
```

- 12 features, 10506 observations
- 6 numerical data (Block, Latitude, Longtitude, AddressAccuracy, NumMosquitos, WnvPresent)
- 6 object data (Date, Address, Species, Street, Trap, AddressNumberAndStreet)
- no Null values observed
- There are 813 records of duplications, assuming to be due to multiple samples of 50 collected from the same trap, with same species and Wnv status detected in the sample group
- Records range May - Oct 2007, 2009, 2011 and 2013

```
WEATHER.CSV DETAILS
**********************************************
Rows: 2944       Columns: 22
==============================================

weather.csv DETAILS.info()
-----------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2944 entries, 0 to 2943
Data columns (total 22 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   Station      2944 non-null   int64
 1   Date         2944 non-null   object
 2   Tmax         2944 non-null   int64
 3   Tmin         2944 non-null   int64
 4   Tavg         2944 non-null   object
 5   Depart       2944 non-null   object
 6   DewPoint     2944 non-null   int64
 7   WetBulb      2944 non-null   object
 8   Heat         2944 non-null   object
 9   Cool         2944 non-null   object
 10  Sunrise      2944 non-null   object
 11  Sunset       2944 non-null   object
 12  CodeSum      2944 non-null   object
 13  Depth        2944 non-null   object
 14  Water1       2944 non-null   object
 15  SnowFall     2944 non-null   object
 16  PrecipTotal  2944 non-null   object
 17  StnPressure  2944 non-null   object
 18  SeaLevel     2944 non-null   object
 19  ResultSpeed  2944 non-null   float64
 20  ResultDir    2944 non-null   int64
 21  AvgSpeed     2944 non-null   object
```

- 22 features, 2944 observations
- 6 numerical data (Station Tmax, Tmin, Dewpoint, ResultSpeed, ResultDir)
- 16 object data
- Check object dtype and date dtype
- no Null values observed., no duplicated rows
- Records range from May- Oct from 2007 to 2014

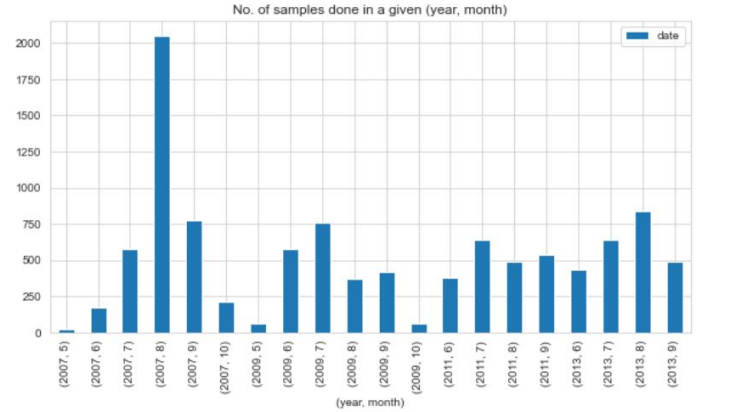
```
SPRAY.CSV DETAILS
**********************************************
Rows: 14835      Columns: 4
==============================================

spray.csv DETAILS.info()
-----------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14835 entries, 0 to 14834
Data columns (total 4 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       14835 non-null  object
 1   Time       14251 non-null  object
 2   Latitude   14835 non-null  float64
 3   Longitude  14835 non-null  float64
dtypes: float64(2), object(2)
```

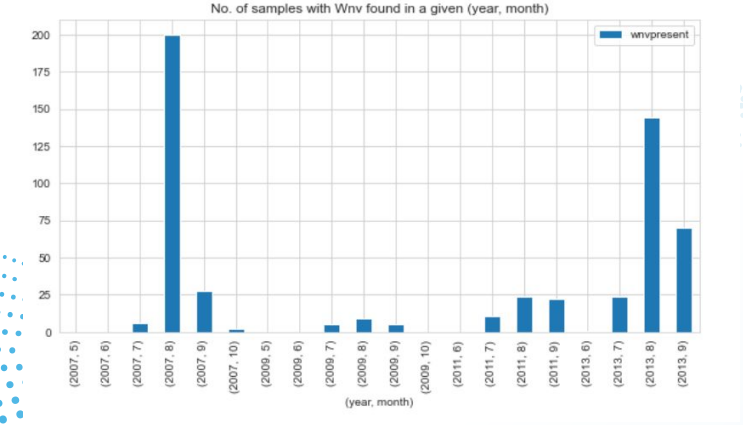
- 4 features, 14835 observations, out of which 584 are null from the Time column
- Records range from Aug, Se 2011 and Jul-Sep 2013

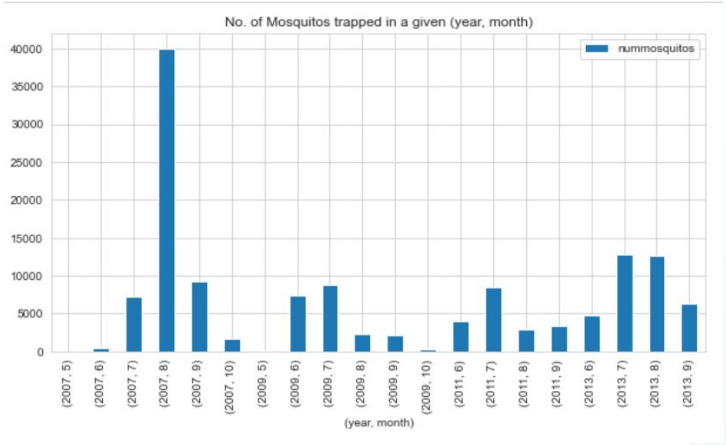
```
test.csv DETAILS.info()
-----------------------------------------------
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 116293 entries, 0 to 116292
Data columns (total 11 columns):
 #   Column                  Non-Null Count   Dtype
---  ------                  --------------   -----
 0   Id                      116293 non-null  int64
 1   Date                    116293 non-null  object
 2   Address                 116293 non-null  object
 3   Species                 116293 non-null  object
 4   Block                   116293 non-null  int64
 5   Street                  116293 non-null  object
 6   Trap                    116293 non-null  object
 7   AddressNumberAndStreet  116293 non-null  object
 8   Latitude                116293 non-null  float64
 9   Longitude               116293 non-null  float64
 10  AddressAccuracy         116293 non-null  int64
dtypes: float64(2), int64(3), object(6)
```

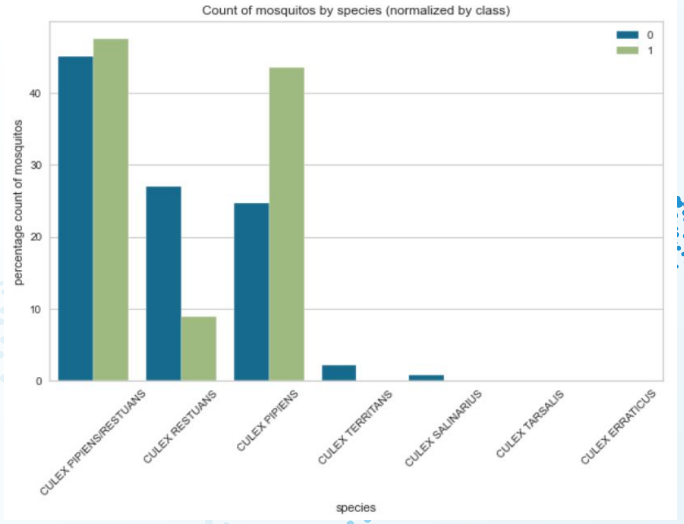
- 11 features, 116293 observations
- 5 numerical data, 6 object
- no Null values observed, no duplicated data
- Records range May - Oct 2008, 2010, 2012 and 2014

EDA - Train

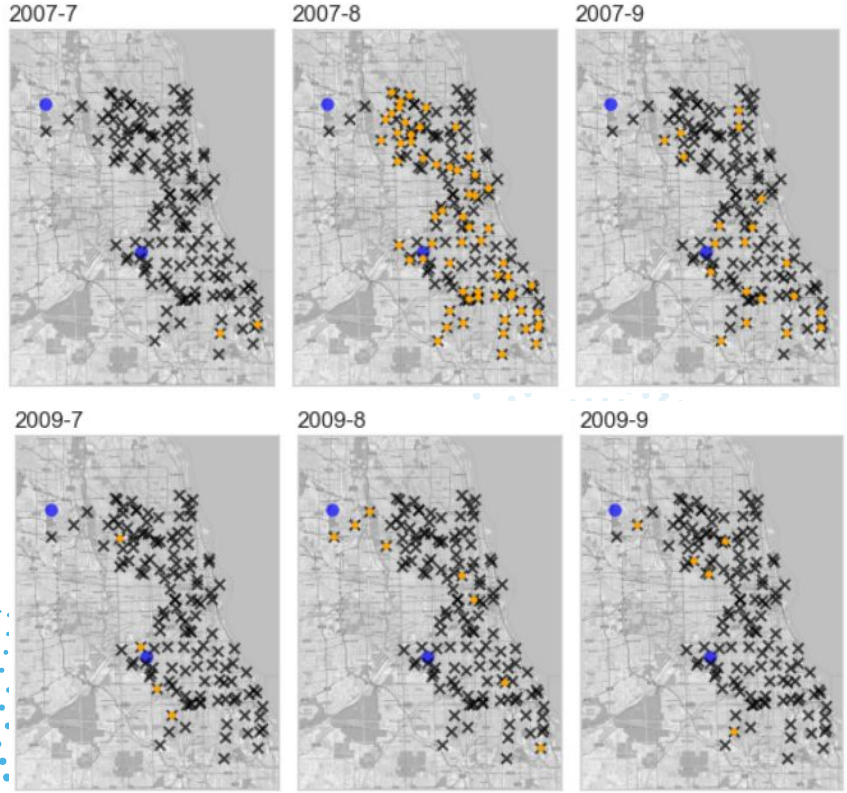

No. of samples done in a given (year, month)
date
(year, month)

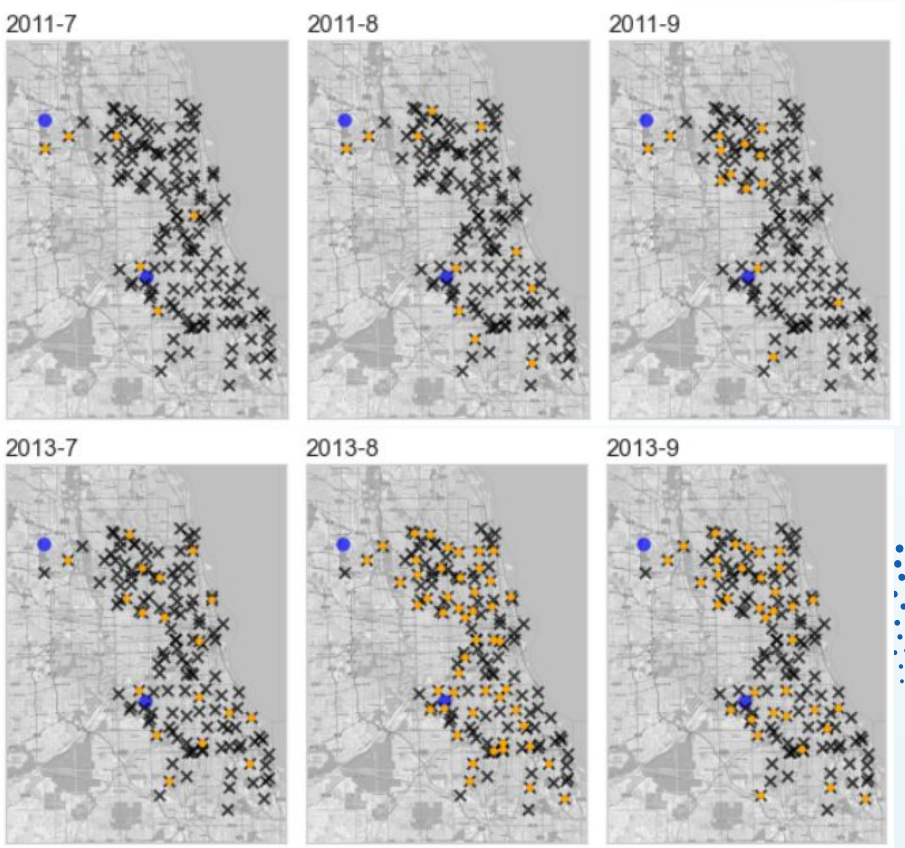

No. of samples with Wnv found in a given (year, month)
wnvpresent
(year, month)


No. of Mosquitos trapped in a given (year, month)
nummosquitos
(year, month)

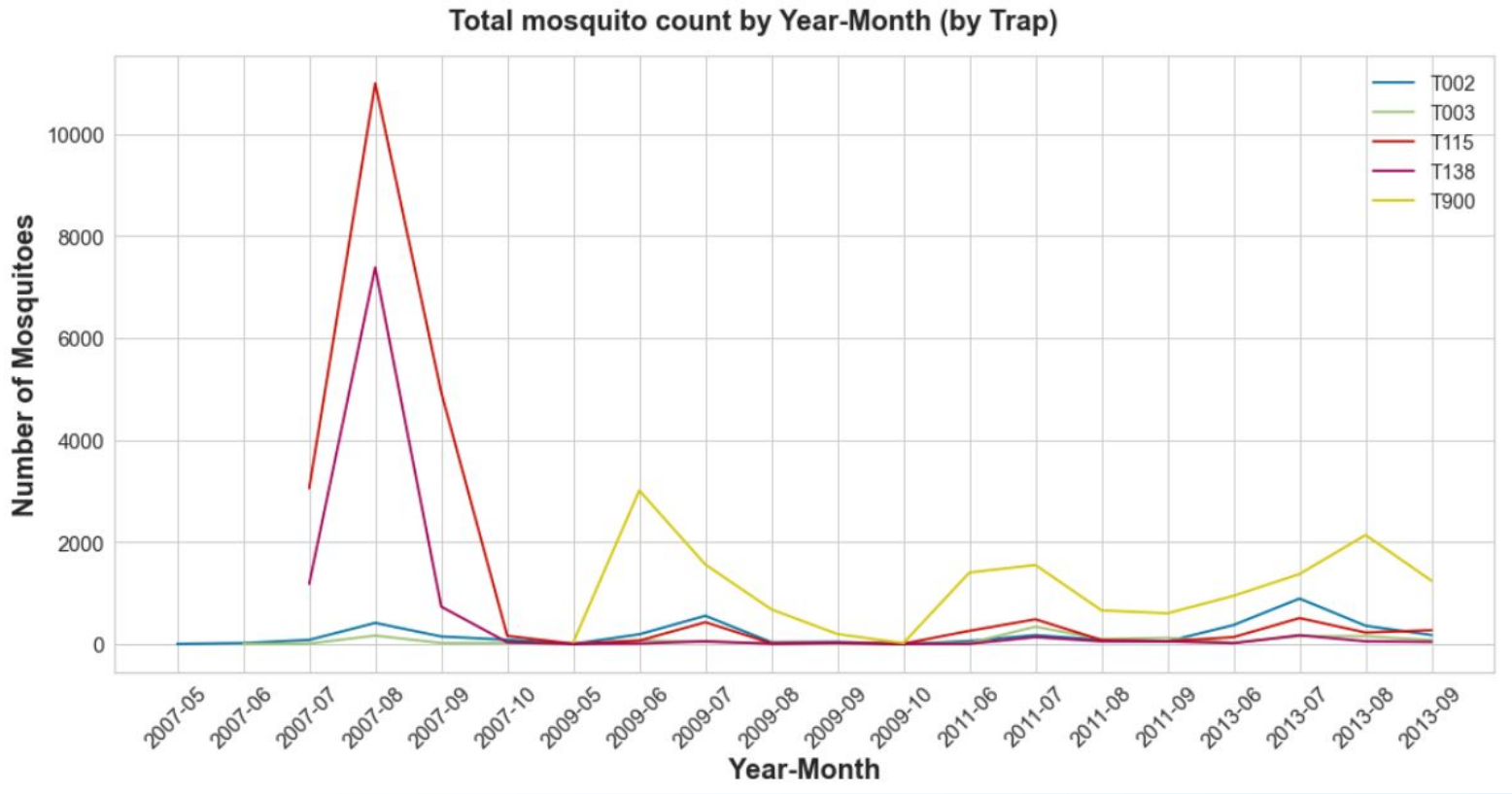

Count of mosquitos by species (normalized by class)
0
1
percentage count of mosquitos
CULEX PIPIENS/RESTUANS
CULEX RESTUANS
CULEX PIPIENS
CULEX TERRITANS
CULEX SALINARIUS
CULEX TARSALIS
CULEX ERRATICUS
species

EDA - Traps

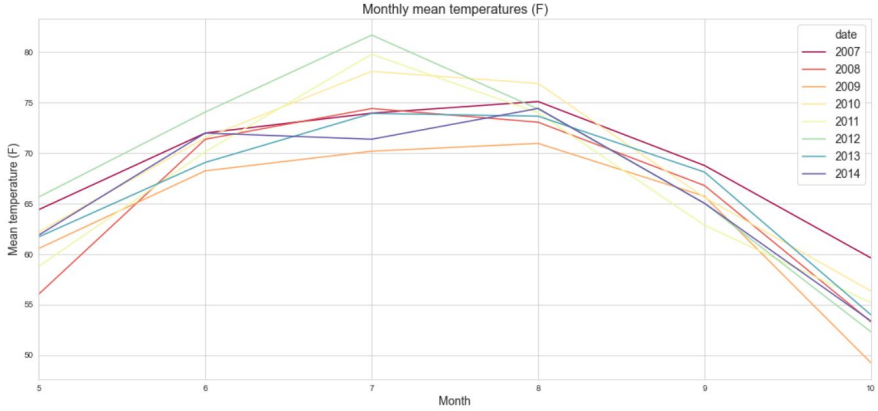

2007-7
2007-8
2007-9
2009-7
2009-8
2009-9

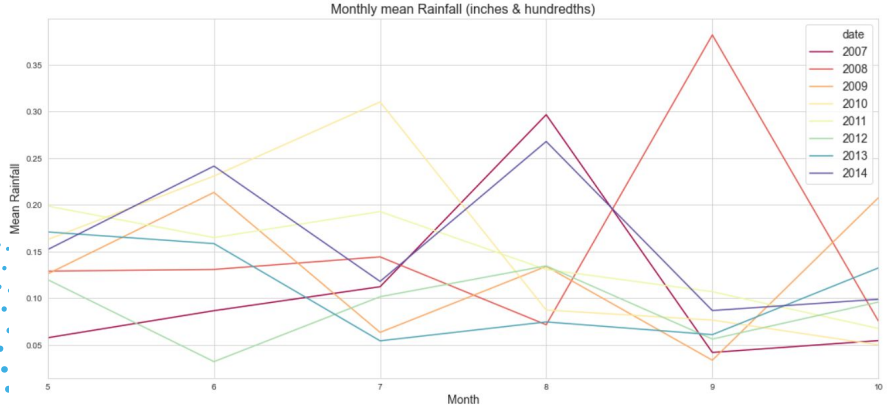

2011-7
2011-8
2011-9
2013-7
2013-8
2013-9

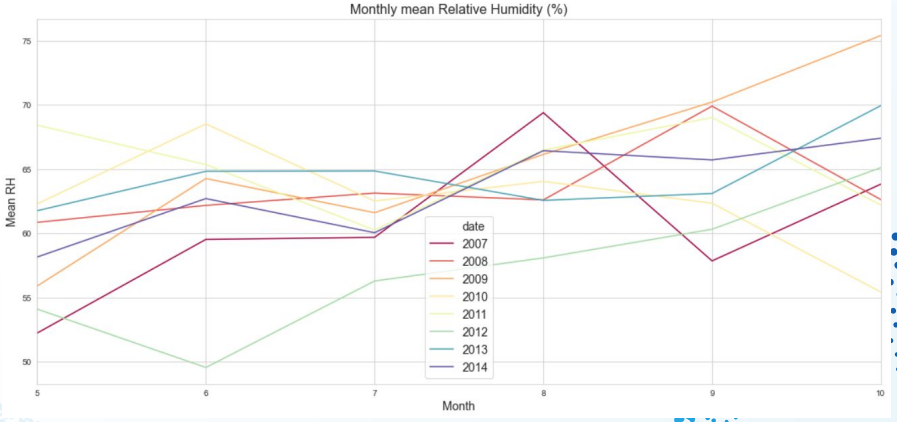
EDA - Traps


Total mosquito count by Year-Month (by Trap)
Number of Mosquitoes
Year-Month
T002
T003
T115
T138
T900
0
2000
4000
6000
8000
10000
2007-05
2007-06
2007-07
2007-08
2007-09
2007-10
2009-05
2009-06
2009-07
2009-08
2009-09
2009-10
2011-06
2011-07
2011-08
2011-09
2013-06
2013-07
2013-08
2013-09

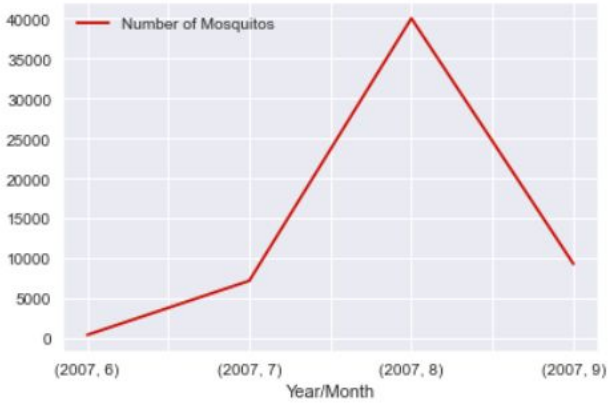
EDA - Weather

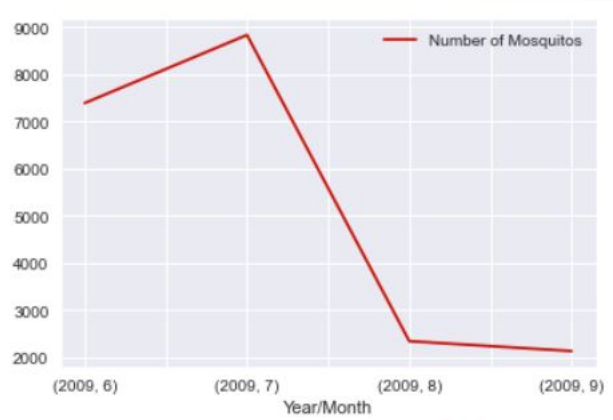

Monthly mean temperatures (F)
Mean temperature (F)
Month
date
2007
2008
2009
2010
2011
2012
2013
2014

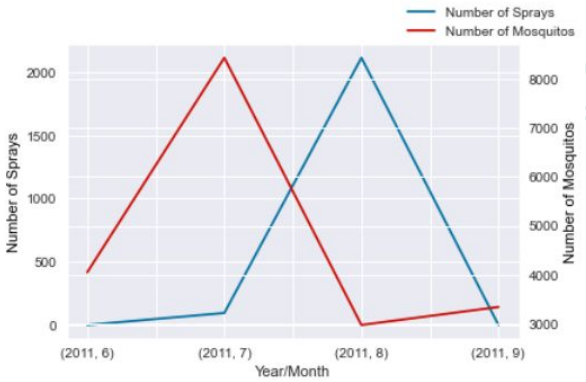

Monthly mean Rainfall (inches & hundredths)
Mean Rainfall
Month
date
2007
2008
2009
2010
2011
2012
2013
2014

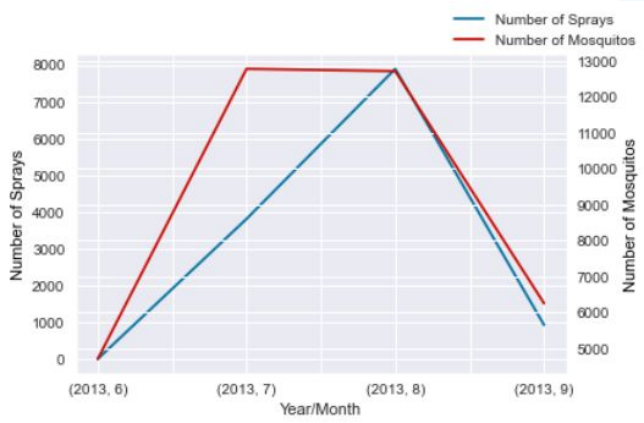

Monthly mean Relative Humidity (%)
Mean RH
Month
date
2007
2008
2009
2010
2011
2012
2013
2014

EDA - Spray vs Mosquito population









Feature Engineering

- **Train**
  - New Feature: `closest_station` - the closest weather station to observed trap (used to join train and weather datasets)
  - Species Consolidation - Consolidate all species not observed to carry virus into others category.
  - New feature: `row_count` - count the number of "duplicate" rows in order to get an estimate for nummosquitos.
  - New feature:`intensity_acc` - calculates the "intensity" for each observation in the train dataset
  - New feature: `trap_weight` -represents the relative importance of each trap based on the number of WNV cases detected at the trap, as well as the number of times it was sampled.
  - Features dropped: `'address', 'block', 'street', 'addressnumberandstreet', 'addressaccuracy'` (can be represented by longitude and latitude)

- **Weather**
  - New Feature: `daytime` – the total number of minutes from Sunrise to Sunset
  - New Feature: `rhumidity` - the relative humidity based on dewpoint and tavg
  - Rolling averages for 7, 10, 14, 30, 60, 90 days. (based on life cycle (7-14 days) of mosquitos, and monthly averages for `tavg, resultspeed, dewpoint, rhumidity, tmax, tmin, preciptotal, and rhumidity`.
  - Features dropped: `'codesum', 'depth', 'water1', 'snowfall', 'stnpressure', 'sealevel', 'heat', 'cool'`
- Joining weather dataset to train dataset by date and station.
- One hot encoding on '`species`'
- Similar functions will be done for test

Modelling: binary classification

Oversampling → Modelling → Evaluation

Imbalanced classes

- Class 0 – 94.8%
- Class 1 – 5.2%

To even out the class distribution, we oversample the minority class (WNV present) via SMOTE prior to modelling

Models used

- Logistic regression
- Support Vector Machine

Tree-based

- Bagging Classifier
- Random Forest
- Extra Trees

Boosting

- AdaBoost
- Gradient Boosting
- XGBoost

Metrics

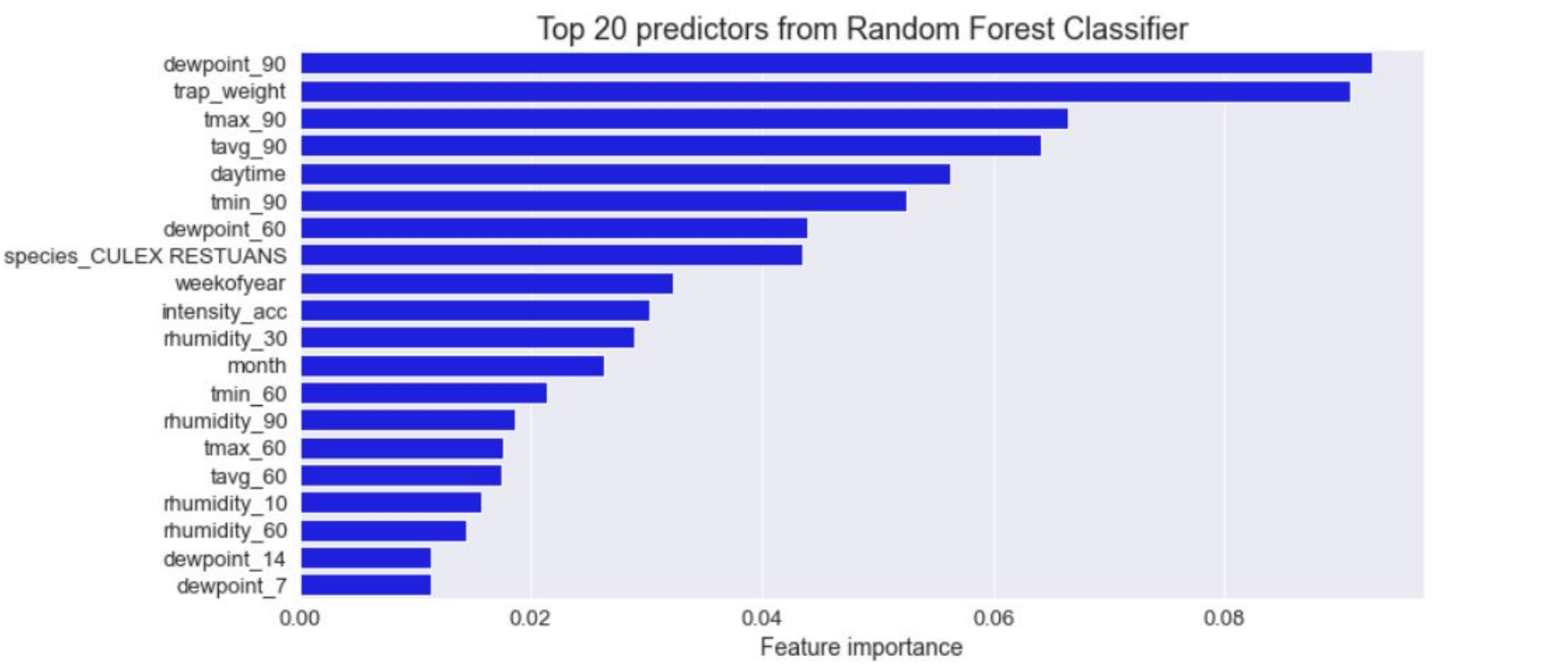
- AUC
- Sensitivity

Performance of Models

	Log Reg	SVM	Bagged DT	Random Forest	Extra Trees	AdaBoost	Gradient Boosting	XGBoost
Accuracy	0.710	0.735	0.744	0.752	0.695	0.916	0.919	0.817
Misclassification Rate	0.290	0.265	0.256	0.248	0.305	0.084	0.081	0.183
Sensitivity (Recall)	0.769	0.775	0.769	0.819	0.802	0.297	0.302	0.648
Specificity	0.707	0.733	0.742	0.748	0.689	0.950	0.953	0.827
Precision	0.127	0.139	0.142	0.153	0.125	0.249	0.263	0.172
True Positive	140.000	141.000	140.000	149.000	146.000	54.000	55.000	118.000
False Positive	964.000	877.000	846.000	827.000	1,021.000	163.000	154.000	569.000
False Negative	42.000	41.000	42.000	33.000	36.000	128.000	127.000	64.000
True Negative	2,321.000	2,408.000	2,439.000	2,458.000	2,264.000	3,122.000	3,131.000	2,716.000
AUC Score	0.820	0.831	0.811	0.846	0.818	0.806	0.806	0.844

Selected model

Selected Model - Important Features



Limitations

- Model is predicting WNV presence based on weather factors affecting Mosquito population
- Other factors affecting Wnv presence
  - Ecological landscape
  - Vegetation index
  - Birds population
  - Human factors
- Gaps in data

Cost Benefit Analysis of Spray

- Estimated cost of spraying in 2013
  - 12625 * $350 = $4,418,750 USD

- Estimated medical expenses for west nile virus treatment in 2013
  - 117 * $27,316.5 = $3,196,000 USD

- Economically not beneficial to spray, however it is the socially responsible thing to do.

- More information is needed with regards to spray for a better assessment.

Conclusion & Recommendations

Mosquito vs Wnv Prediction

- Random Forest
  - ROC ~ 0.845
- Identify high influencing factors
  - dewpoint_90, trap_weight, tavg_90, tmax_90, daytime

Limitations on prediction

- Other factors affecting Wnv presence
  - Ecological landscape
- Gaps in Data

Cost Benefits of Spray

- Not beneficial from an economical perspective, but beneficial for society at large

Recommendations

- Data Collection on Spray, Vegetation and Bird Population for further analysis
- Awareness Education (campaigns to educate on Virus prevention)