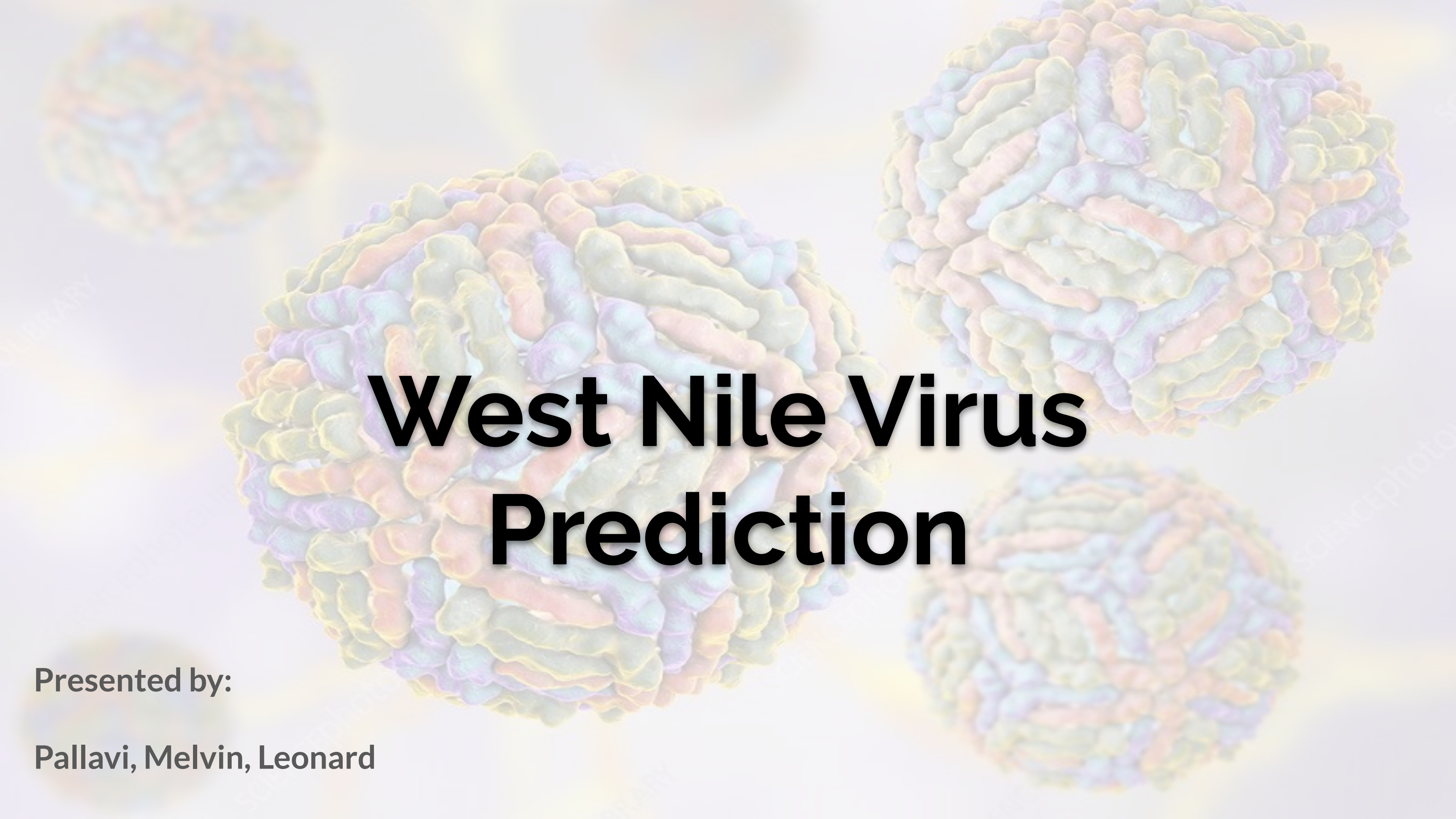


West Nile Virus Prediction

Presented by:

Pallavi, Melvin, Leonard

Problem Statement

The goal of this project is to use surveillance data to predict the probability of West Nile Virus for a given time, location and mosquito species.

Datasets

Dataset	Shape	Description	Recorded Date
Trap (Train.csv)	10506 X 12	Locations of mosquito traps	May-Oct 2007 May-Oct 2009 Jun-Sept 2011 Jun-Sept 2011
Trap (Test.csv)	11623 X 11	Locations of mosquito traps	Jun-Sept 2008 Jun-Oct 2010 Jun-Sept 2012 Jun-Oct 2014
Weather.csv	2944 X 22	Weather conditions from 2 stations	Jan-Dec 2007-14
Spray.csv	14835 X 4	GIS data for the City of Chicago's spray efforts	Jun-Sept 2011 Aug-Sept 2013

Data Cleaning & EDA

Train

813 duplicate rows for same date, trap, species and location

Why?
New entry where mosquitoes exceeded 50

Solution: Combined the data

Weather

Incomplete Data, ex missing values represented by ‘T’, ‘M’

Solution: Missing values replaced with 0

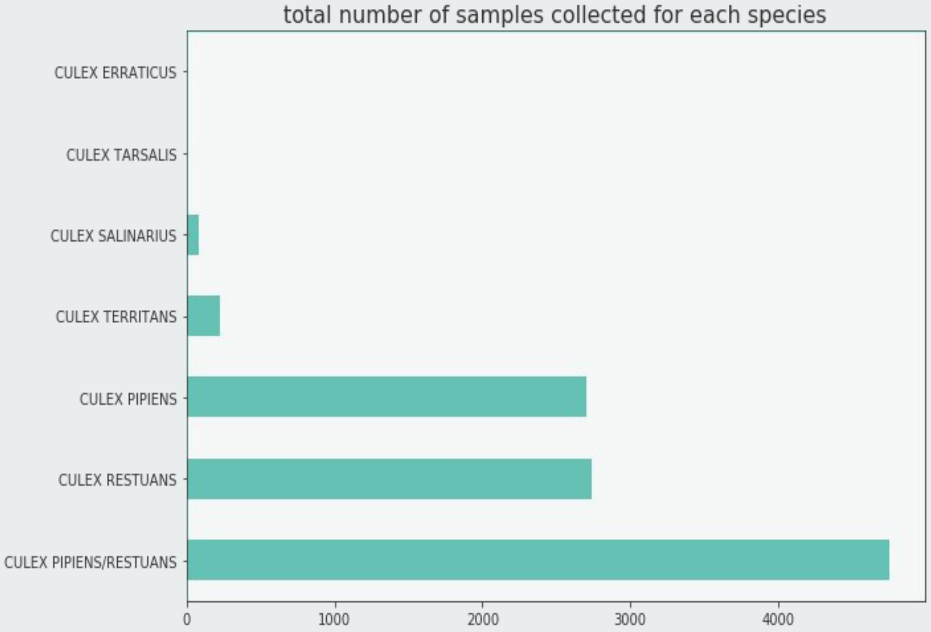
Spray

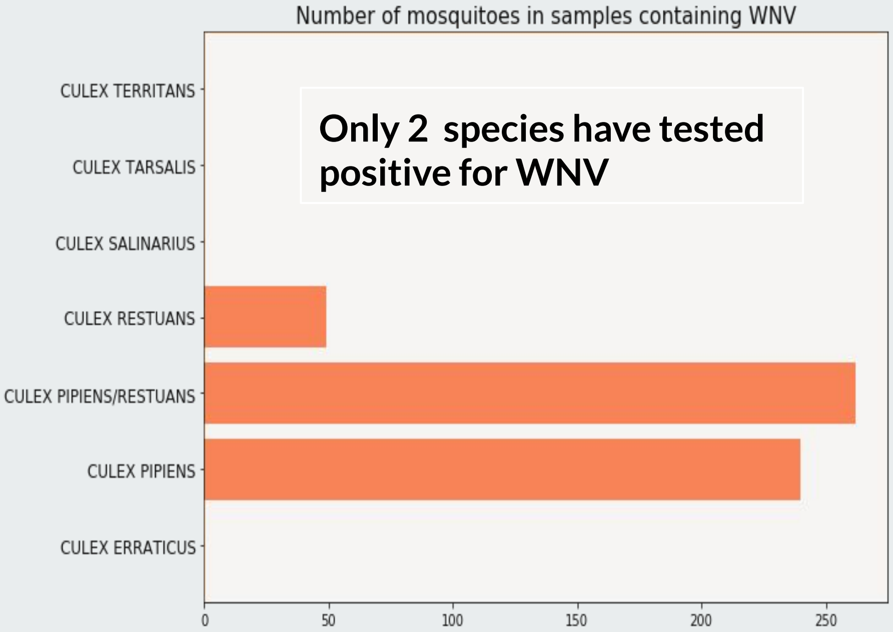
584 null values
543 duplicate rows

Why?
Null values caused by missing data in ‘Time’

Solution: Dropped the duplicate rows

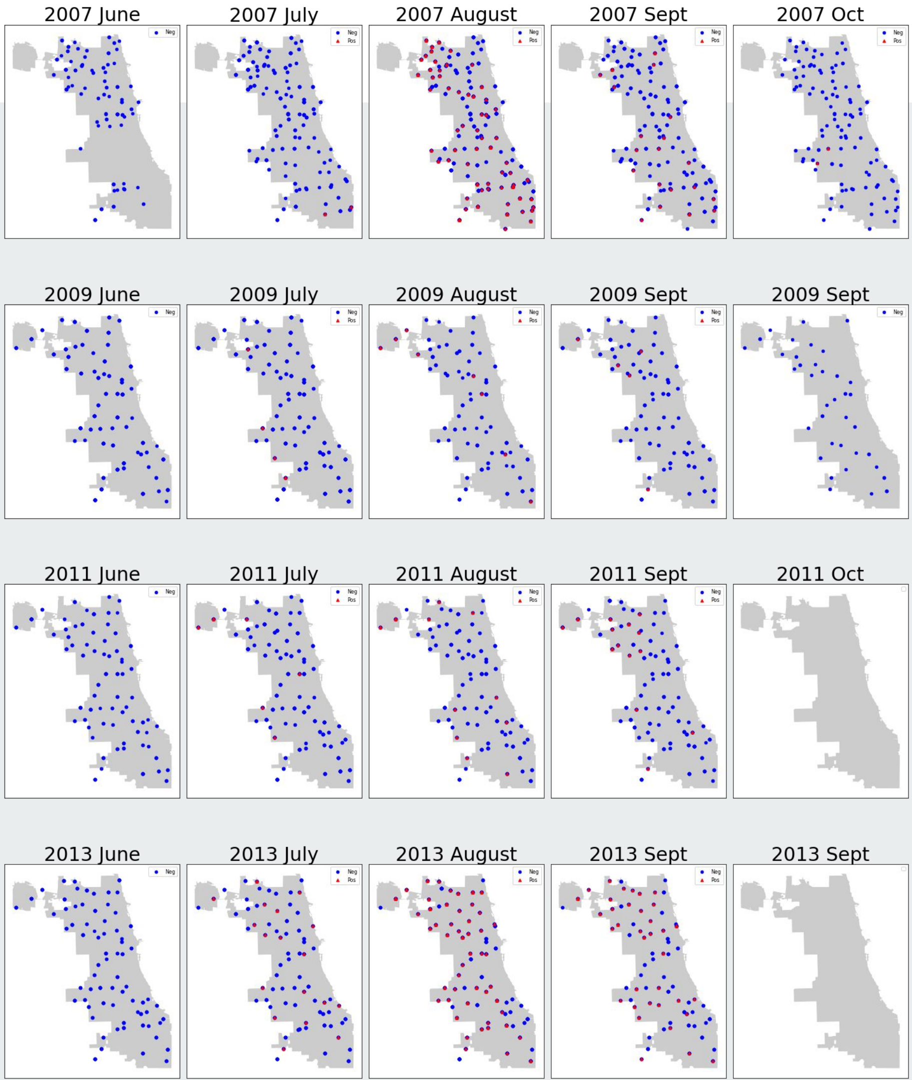
Total Mosquito Species vs WNV Mosquitoes



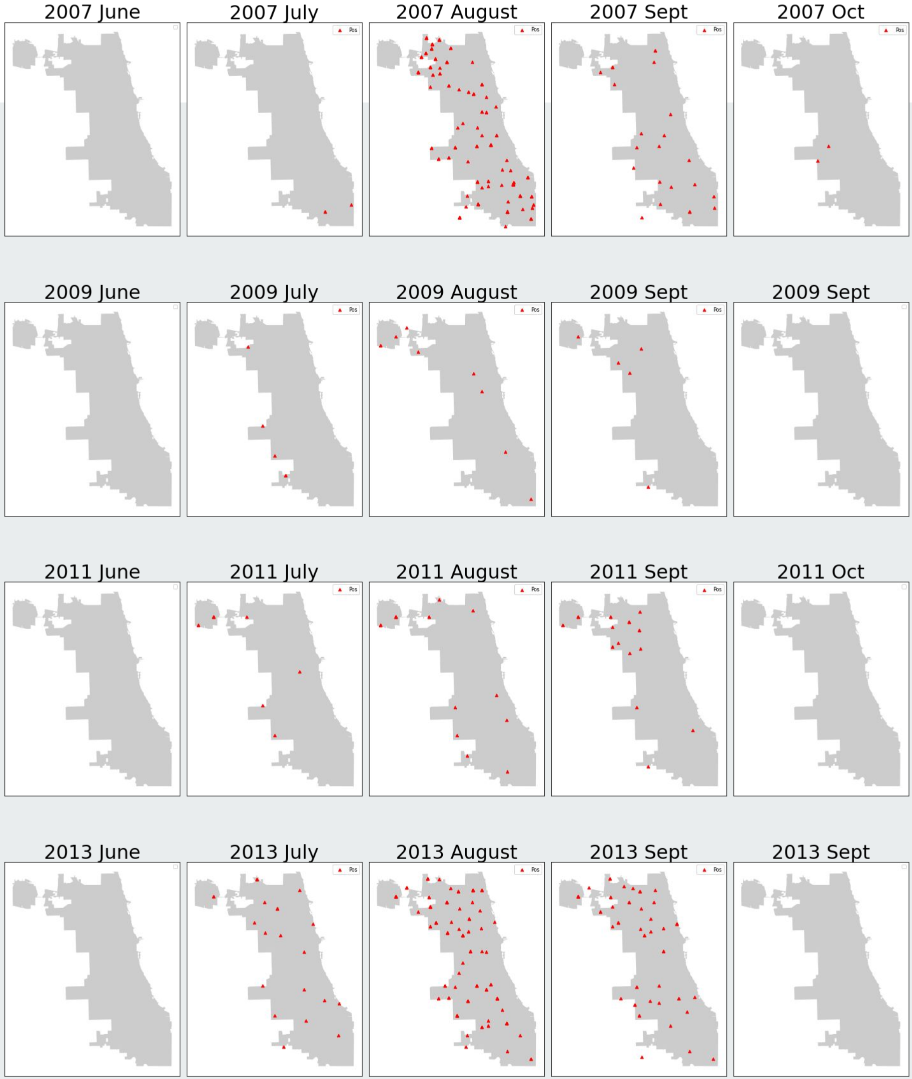


Only 2 species have tested positive for WNV

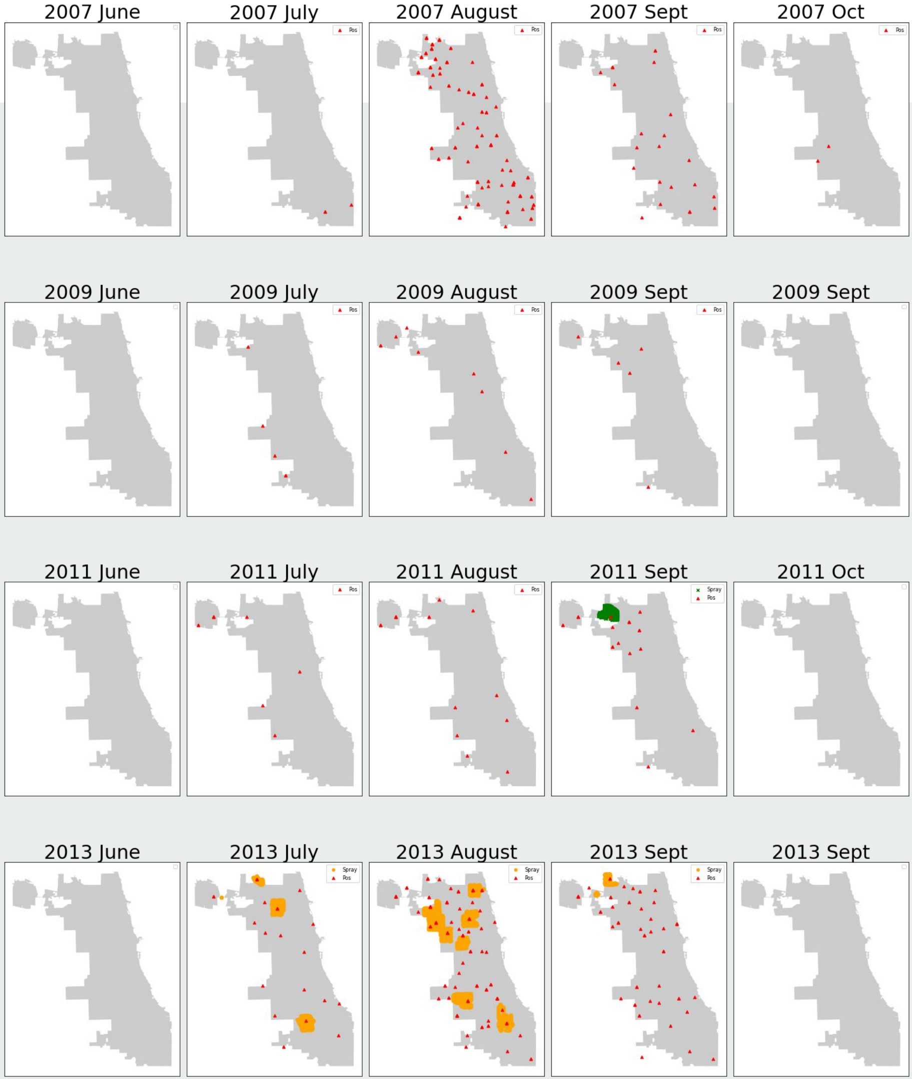
Location of traps and Mosquitos caught



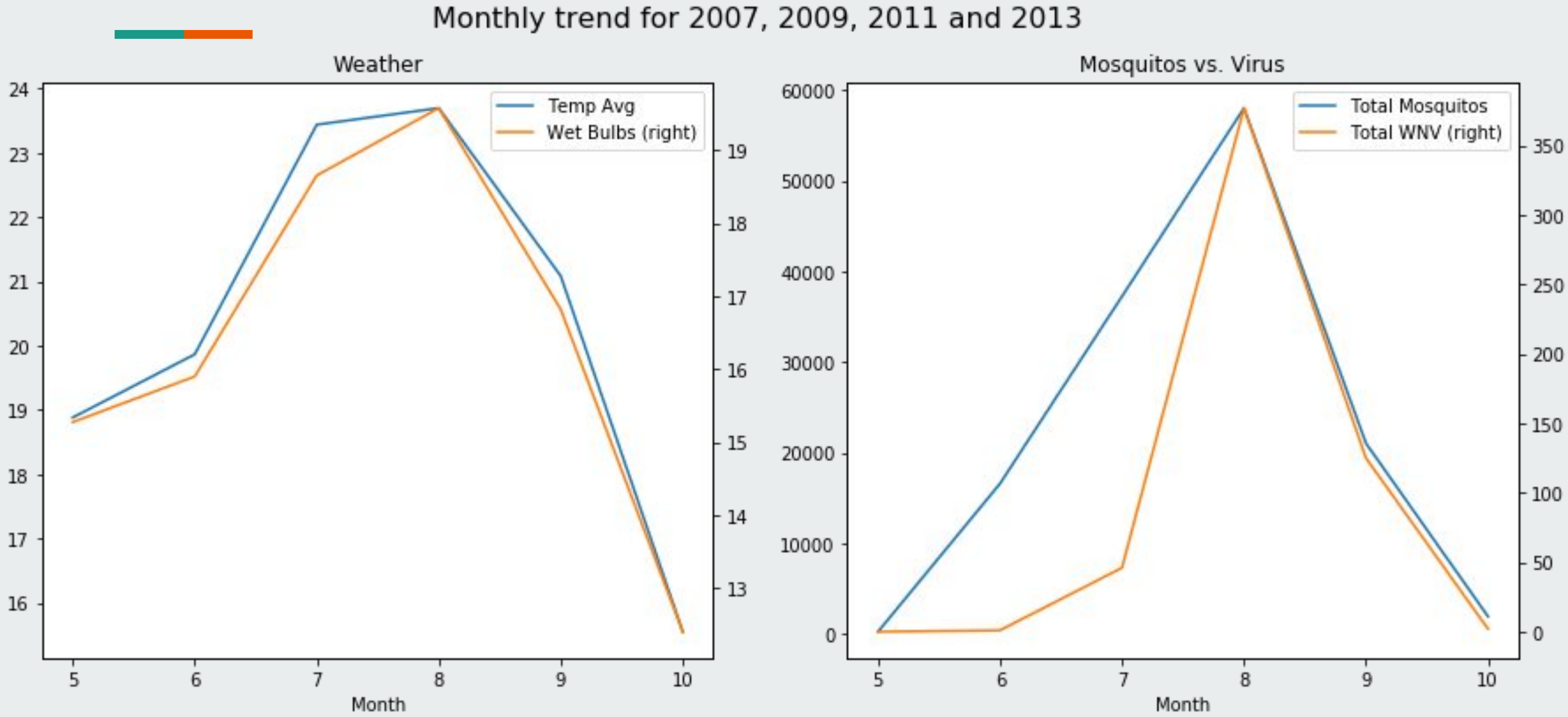
Location of WNV-positive mosquitos



Location of WNV-positive mosquitos and spray locations



Monthly Weather & Mosquito Trends

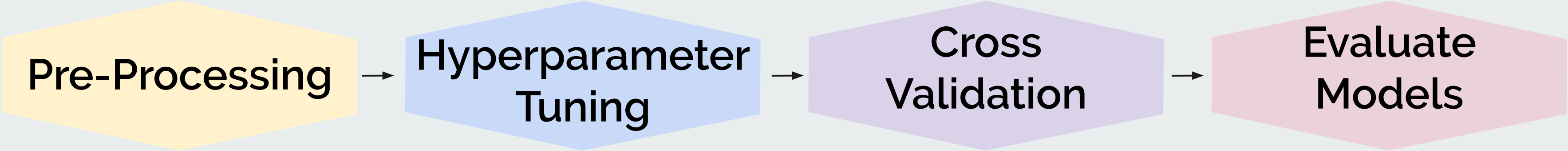


Monthly trend for 2007, 2009, 2011 and 2013

Weather

Mosquitos vs. Virus

Modeling - Workflow



Pre-Processing → Hyperparameter Tuning → Cross Validation → Evaluate Models

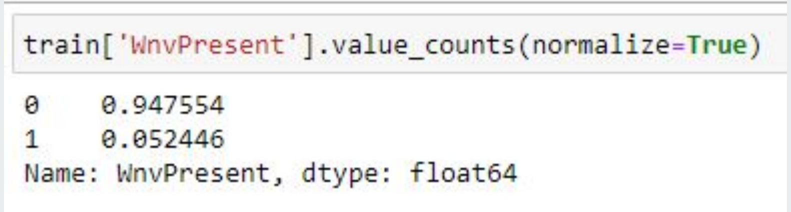
- One-Hot Encoding
- Feature Engineering
- Features Selection

- Train-test split
- GridSearchCV
- 5 Different Classifiers

- Tuned Hyperparameters
- 5 Different Classifiers

- Mean Score
- Standard Deviation

Modeling - Baseline Score



```
train['WnvPresent'].value_counts(normalize=True)
```

```
0    0.947554
1    0.052446
Name: WnvPresent, dtype: float64
```

Mitigation Methods

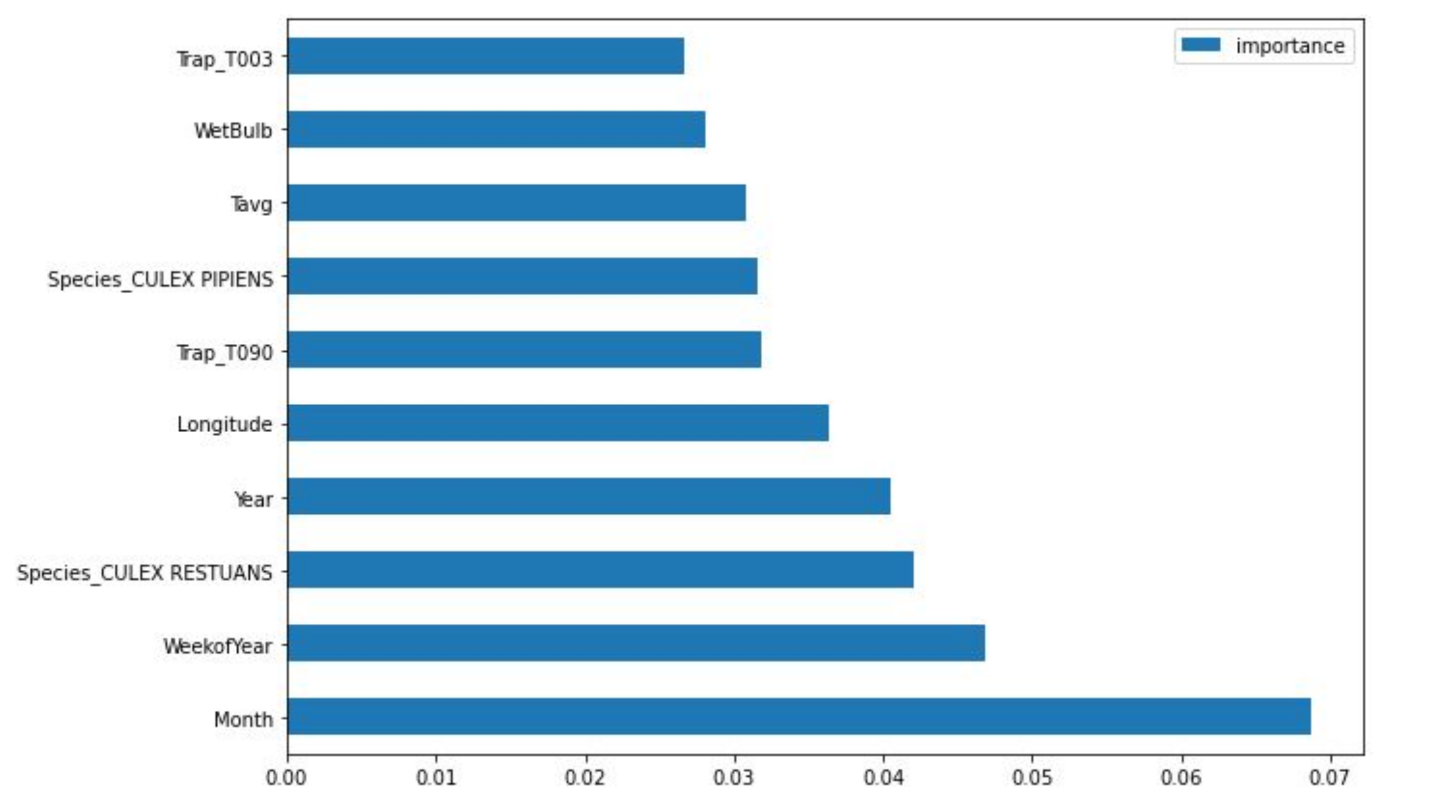
- Choosing the right evaluation metrics
- Under-sampling
- Oversampling

Modeling - Classifier

Classifier	Best Hyper-Parameters	ROC_AUC Score Train	Mean ROC_AUC CrossValScore Train
Logistic Regression	C = 0.1, penalty = ‘l1’, solver = ‘liblinear’	76.30	71.78 +/- 1.94
K Nearest Neighbor	n_neighbors = 6, leaf_size = 5, p = 1, weights = ‘distance’	68.00	73.49 +/- 3.51
Random Forest	max_features = ‘sqrt’, min_samples_leaf = 5, n_estimators = 200	84.39	84.48 +/- 1.41
Support Vector	C = 100, gamma = 0.0001, kernel = ‘rbf’	76.72	80.19 +/- 2.35
XGBoost	colsample_bytree = 0.2, gamma = 0.02, learning_rate = 0.1, max_depth = 3, reg_alpha = 0, reg_lambda = 1, subsample = 0.5	84.53	84.45 +/- 1.36

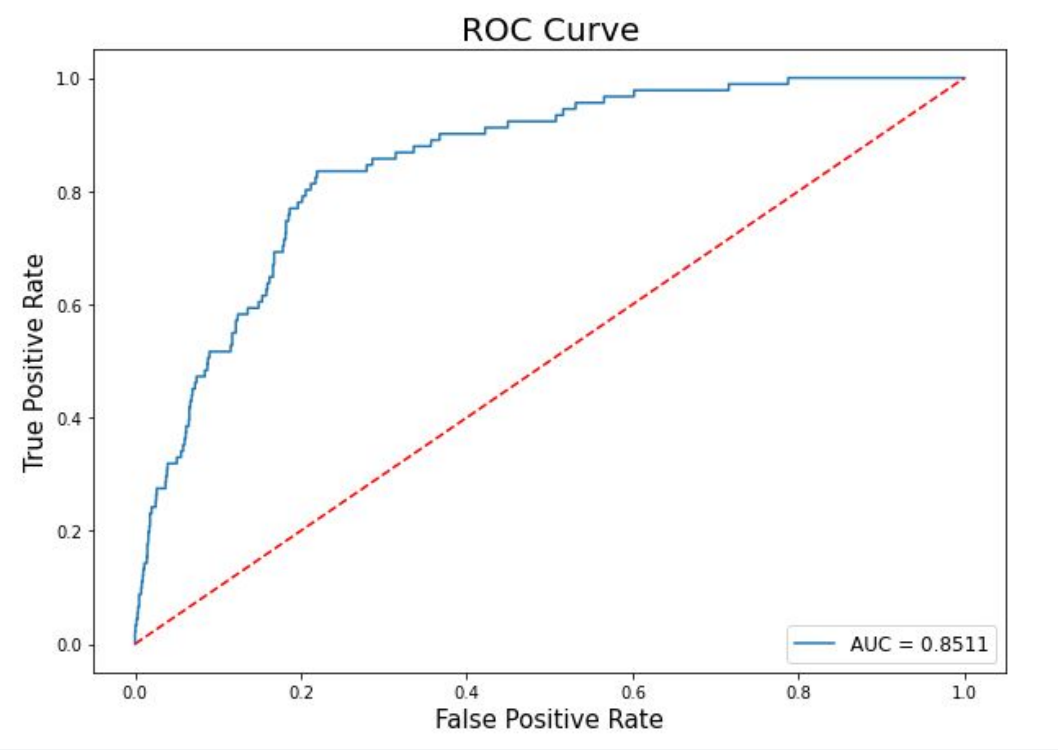
Modeling - Top 10 Features

- Weather
- Trap
- Year/Month
- Location
- Species

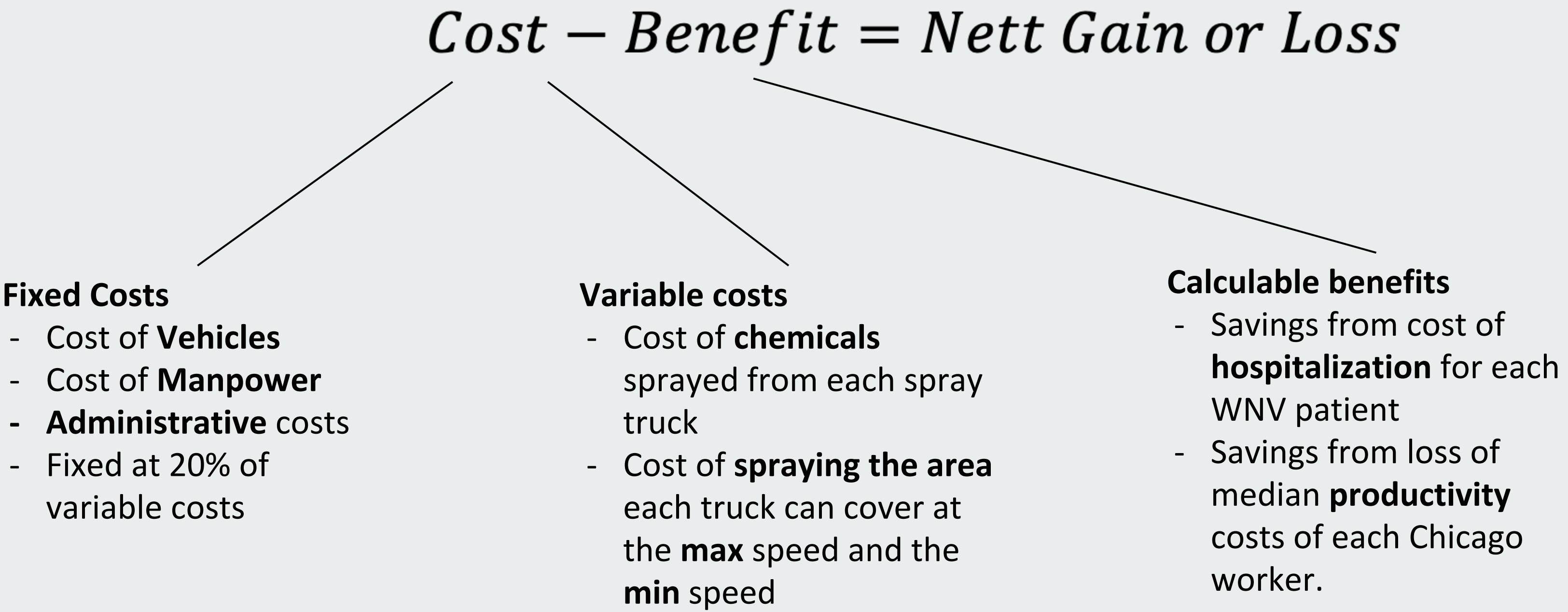


Modeling - Final Model

- XGBoost
- ROC_AUC_Score = 0.8511
- On unseen data = 0.74209



Cost Benefit Analysis



$$Cost - Benefit = Nett\ Gain\ or\ Loss$$

Fixed Costs

- Cost of **Vehicles**
- Cost of **Manpower**
- **Administrative** costs
- Fixed at 20% of variable costs

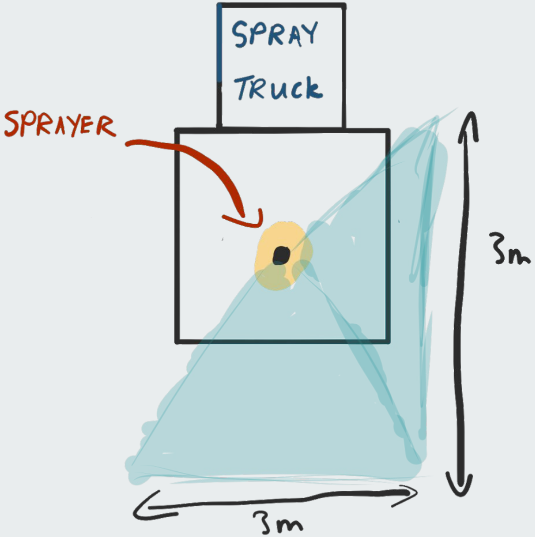
Variable costs

- Cost of **chemicals** sprayed from each spray truck
- Cost of **spraying the area** each truck can cover at the **max** speed and the **min** speed

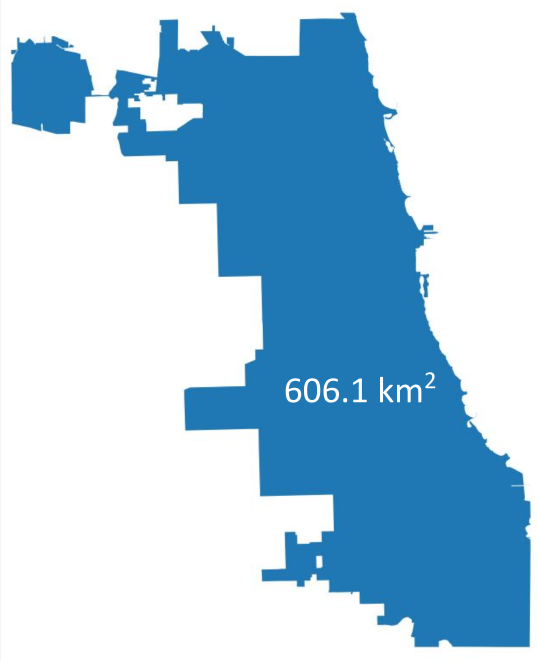
Calculable benefits

- Savings from cost of **hospitalization** for each WNV patient
- Savings from loss of median **productivity** costs of each Chicago worker.

Cost Assumptions



Coverage = 9 m^2
Speed: 16 to 24 km/h
Sprays: 0.05 gallons/min



Cost: USD $80 / US Gallon

Benefits Assumptions

Median income:

USD $55,295

(2017)

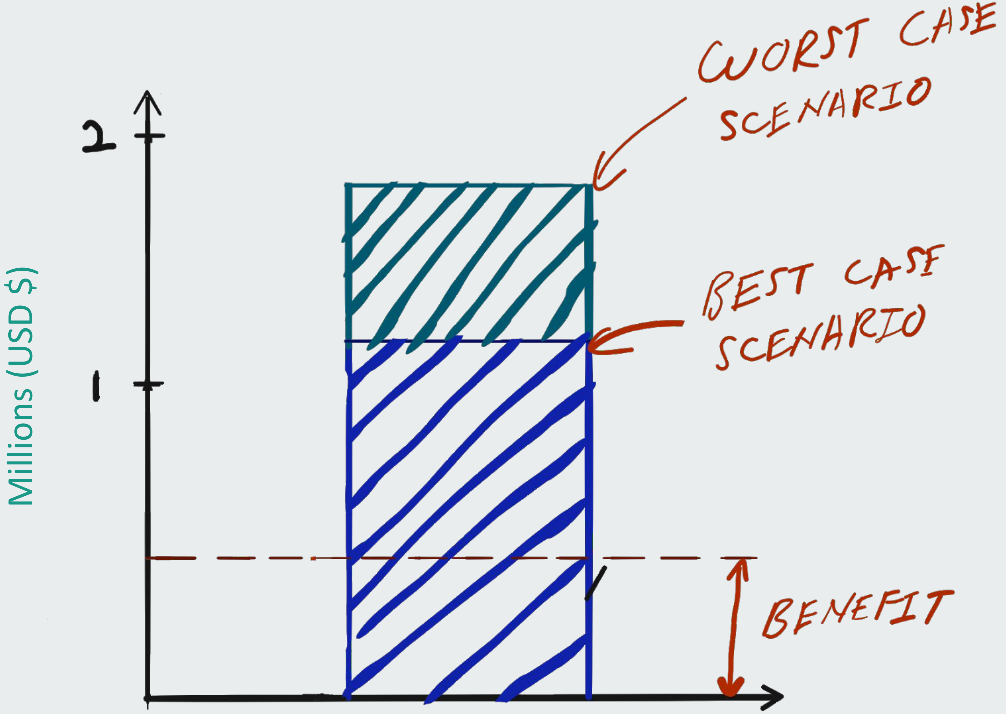
Hospitalisation costs:

USD $25,000 per

hospital stay

(2017)

Cost Benefit Analysis



Conclusion & Recommendation

After studying the effectiveness of spraying and given its high cost:

- Re-examine the effectiveness of spraying Zenivex™.
- Develop mosquito spraying regimes in a more organised and evidence-driven manner.
- Examine new ways of controlling the mosquito population that may cost less than spraying the whole of Chicago.

Recommendations based on model

Our model achieved a 0.74209 ROC_AUC score :

- Recommend getting more data over the years to improve the score

- Once optimised (achieves a better score than the baseline of 0.94) will be able to use the model to predict WNV hotspots.

Questions?