West Nile Virus Prediction Analysis in Chicago

Matthew | Esther | Zhi Cong

Agenda

Background	• Background information • Project objective
Data analysis	• Overview of datasets • Exploratory Data Analysis
Modelling	• Feature Engineering • Model selection & tuning
Conclusion	• Model insights • Cost-Benefit Analysis & Recommendations

Understanding the problem

West Nile Virus

Spread to humans through infected mosquitos.

Symptoms:

- persistent fever
- serious neurological illnesses
- death

Chicago

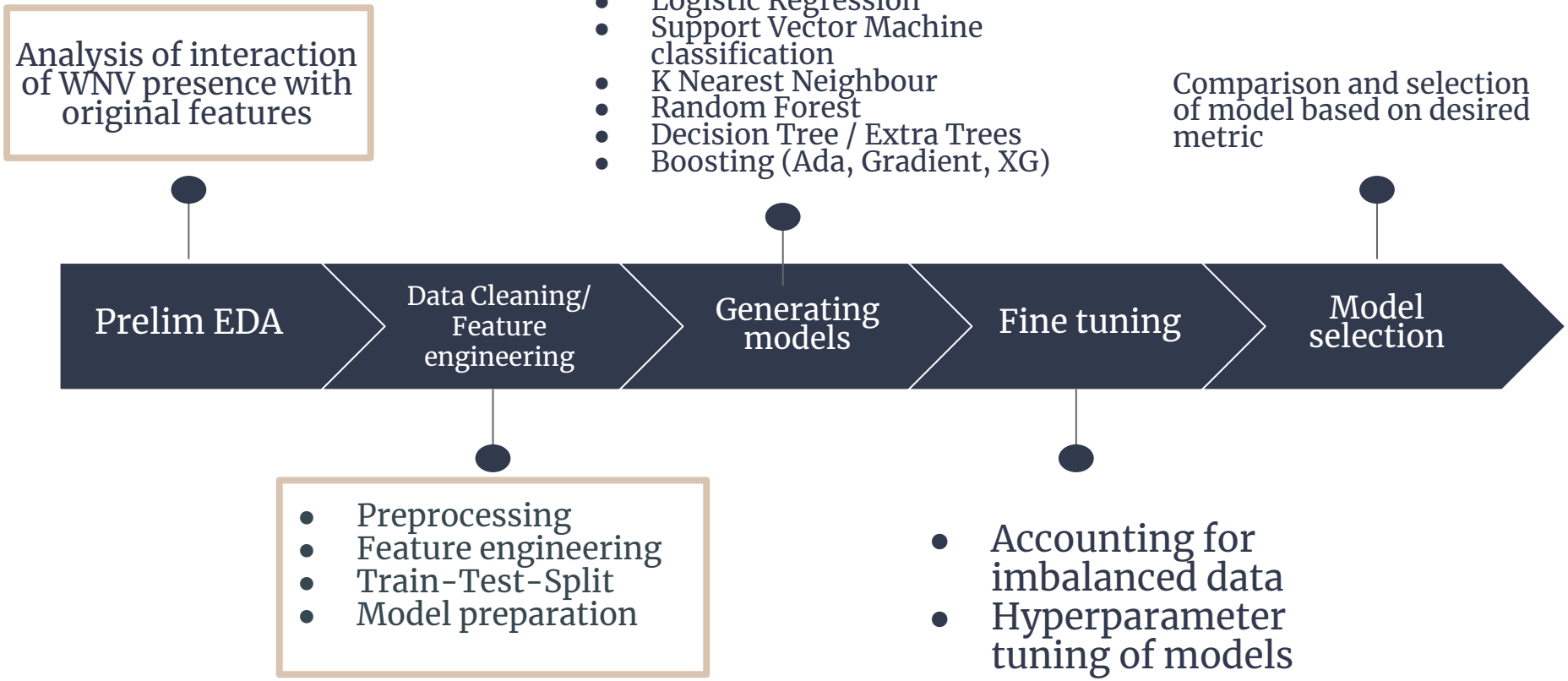
- 2002: First human case reported
- 2004: Established surveillance & control program

Testing results influence location and timing of airborne pesticide sprays

Project objective:
To predict the probability of the presence of West Nile Virus

- For a specific location
- On a specific date

Data Modeling Workflow



Data Analysis

Datasets

Train

Trap records:

- 2007, 2009, 2011, 2013
- Location
- Date
- Number of mosquitos
- Presence of WNV

Weather

Weather locations:

- 2007 -2014
- 2 weather stations
  - O'Hare airport
  - Midway airport

Spray

Spraying efforts:

- 2011 and 2013
- 2 spray date in 2011
- 8 spray dates in 2013
- Date
- Location

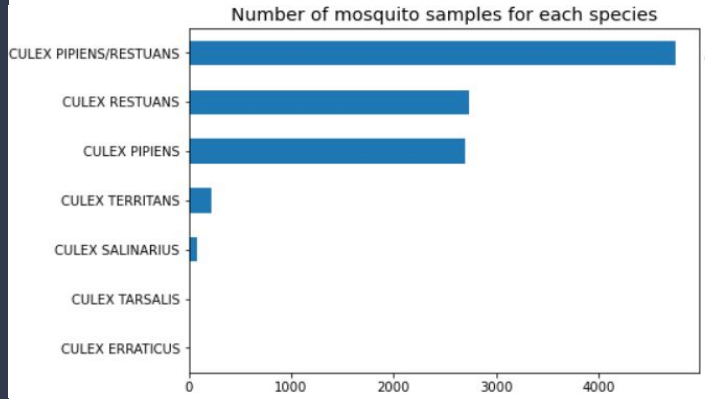
*Missing values and duplicates were dropped for spray dataset

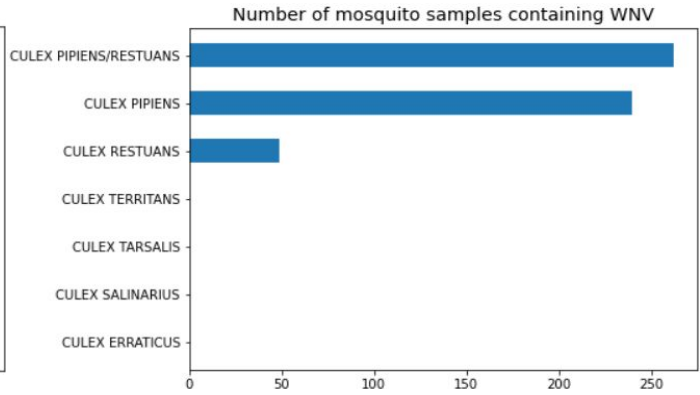
Data Analysis: Train dataset

Different mosquito species

- 2 most common species: Culex Pipiens and Culex Restuans
- Also the only 2 species that carry WNV
- Much more Culex Pipien samples that are WNV-present
- Highest proportion of WNV-present samples: Culex Pipiens could be the more dangerous species

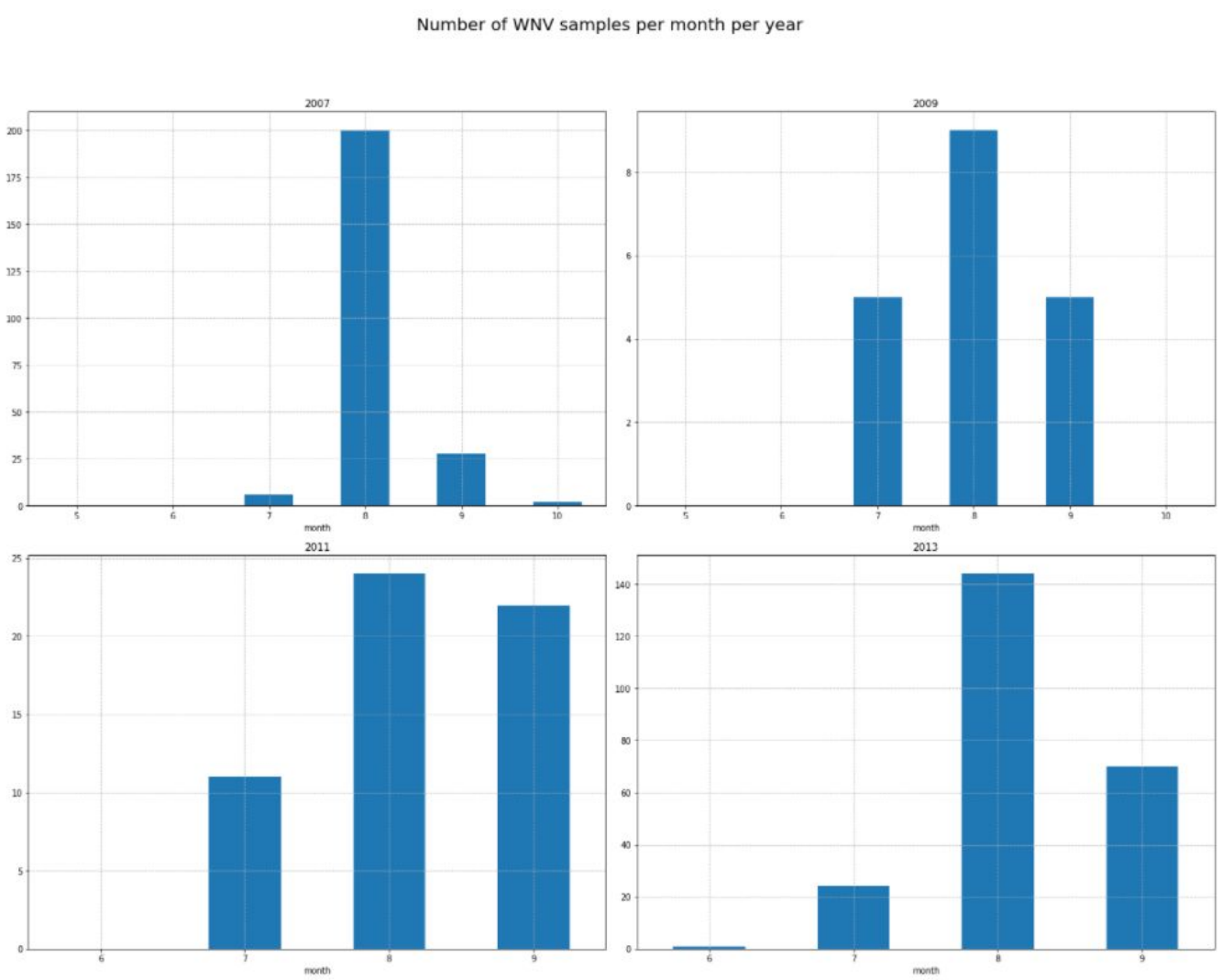
	species	nummosquitos	wnvpresent	sample_count
0	CULEX PIPIENS/RESTUANS	66268	262	4752
1	CULEX PIPIENS	44671	240	2699
2	CULEX RESTUANS	23431	49	2740
3	CULEX ERRATICUS	7	0	1
4	CULEX SALINARIUS	145	0	86
5	CULEX TARSALIS	7	0	6
6	CULEX TERRITANS	510	0	222





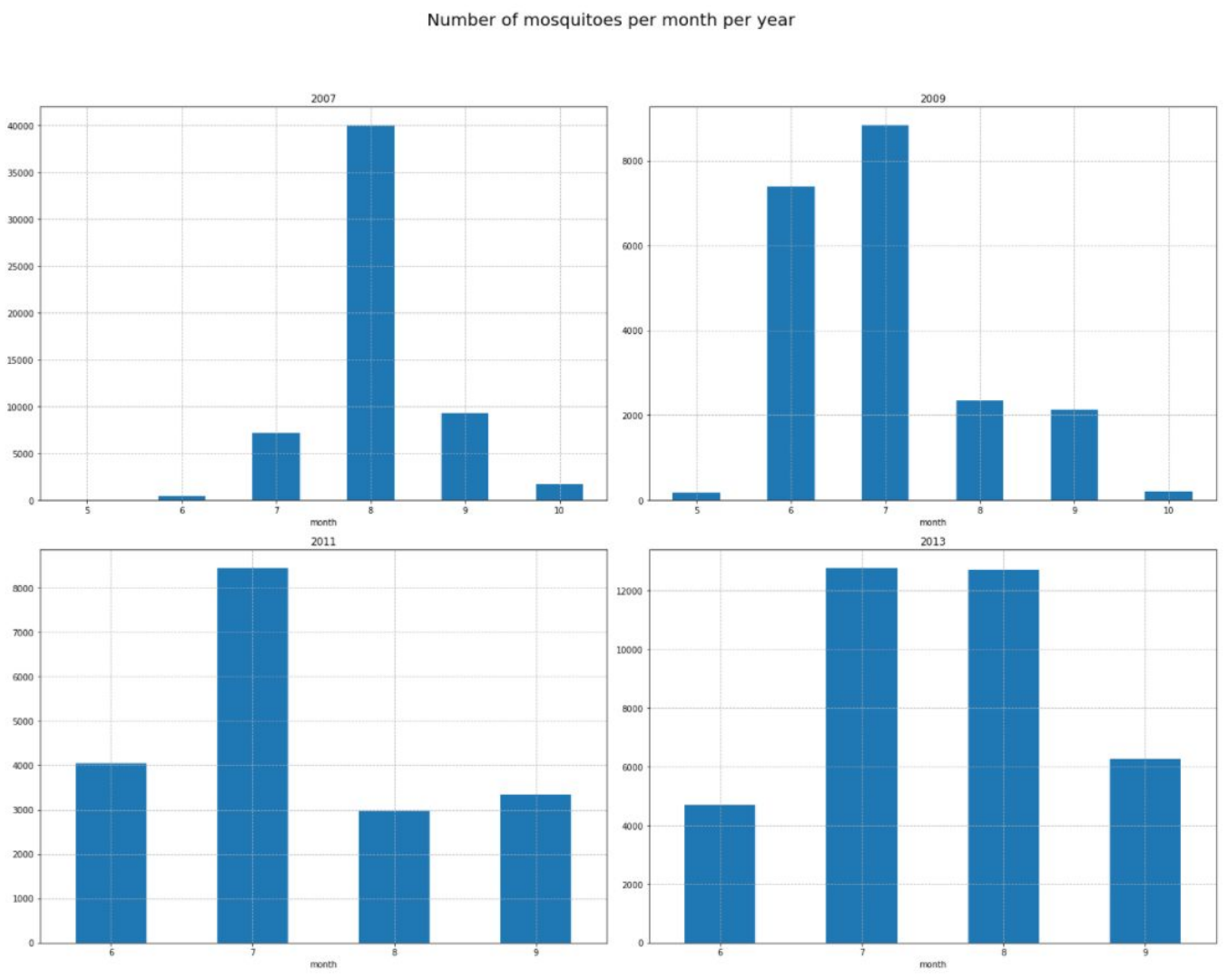
Number of WNV cases per month for each year

- Year with highest WNV: 2013
- WNV starts around June, rise rapidly from July
- Peak month: August
- Reduce from Sept, cease from Oct
- Summer months in Chicago: June to August



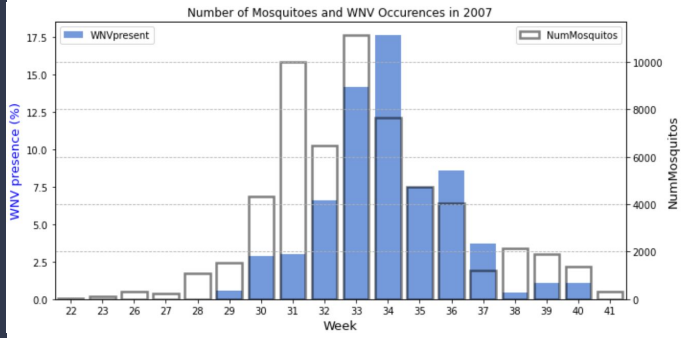
Number of mosquitoes trapped per month for each year

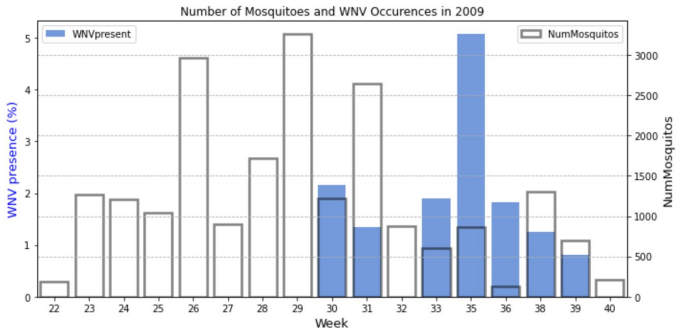
- Month with highest number of mosquitoes: July or August
- Highest number of mosquitoes precede highest number of WNV occurrences

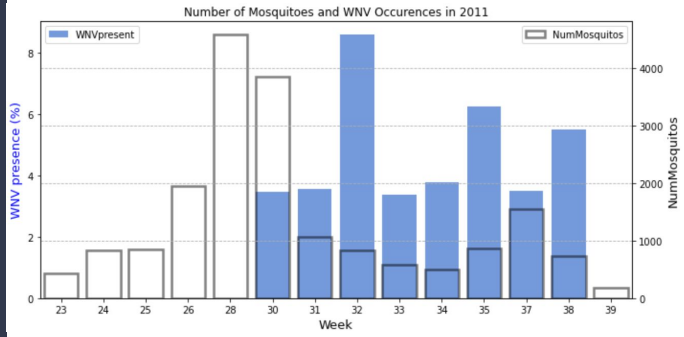


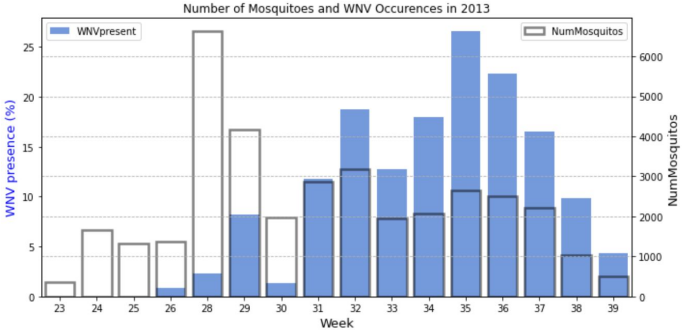
Number of WNV cases VS Number of mosquitoes

- High number of mosquitoes precede high number of WNV
- High number of mosquitoes: week 26 to 33
- High WNV cases: week 30 to 39







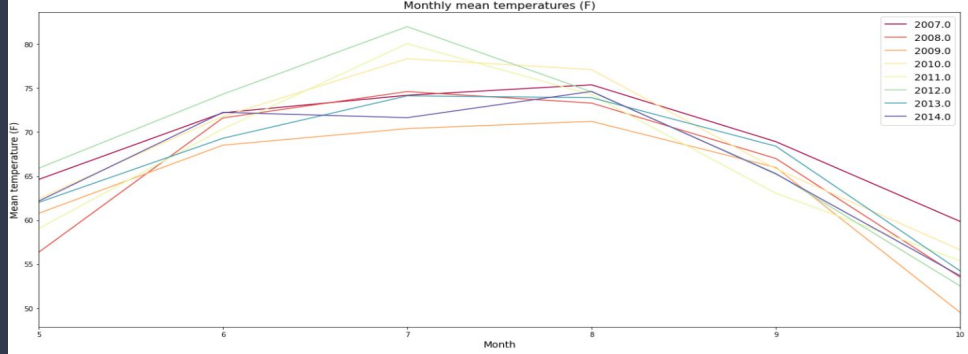


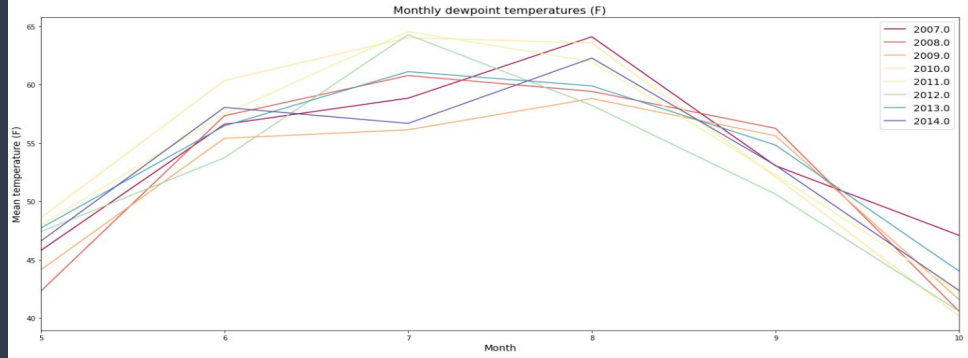
Data Analysis: Weather dataset

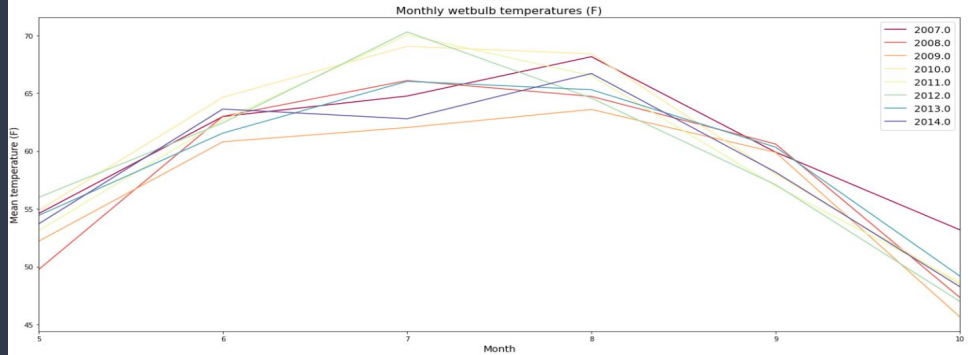
Weather Dataset: Temperature

- Monthly mean temp
- Dew-point temp
- Wet-bulb temp

- Summer months: Similar trends with high number of mosquitoes and high WNV over the months of June, July and August
- High humidity - High mosquito activity
- High precipitation - High WNV







Data Analysis: Combined dataset

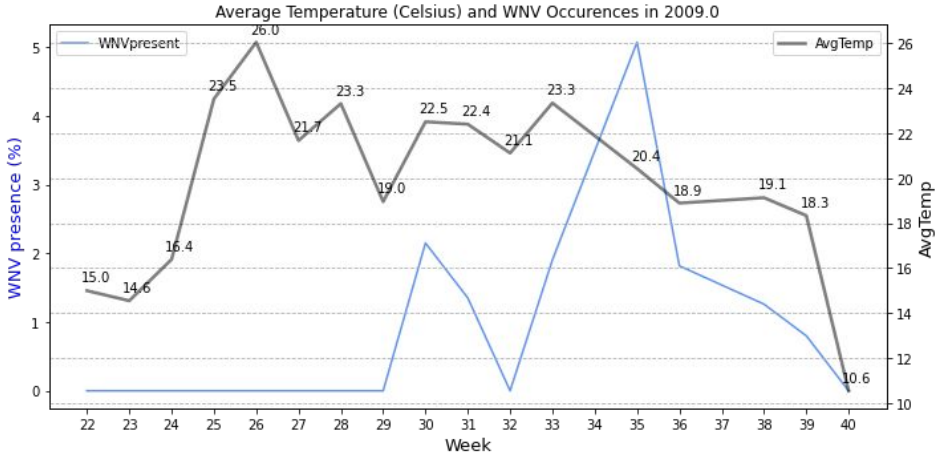
WNV
Average temperature
Time

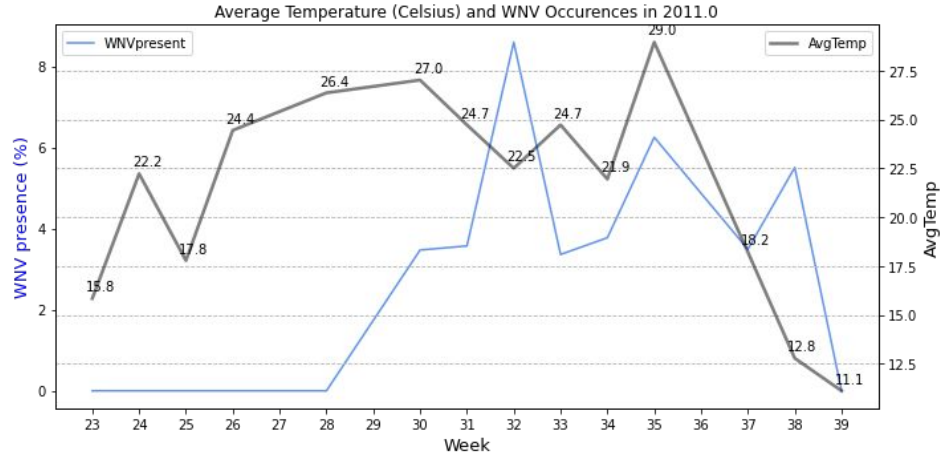
- Potential lead indicator
  - Lead indicator: shows change in direction before corresponding change in target

- 2009:
  - Week 28-30
  - Week 33-35

- 2011:
  - Week 30-32
  - Week 33-35
  - Week 35-38

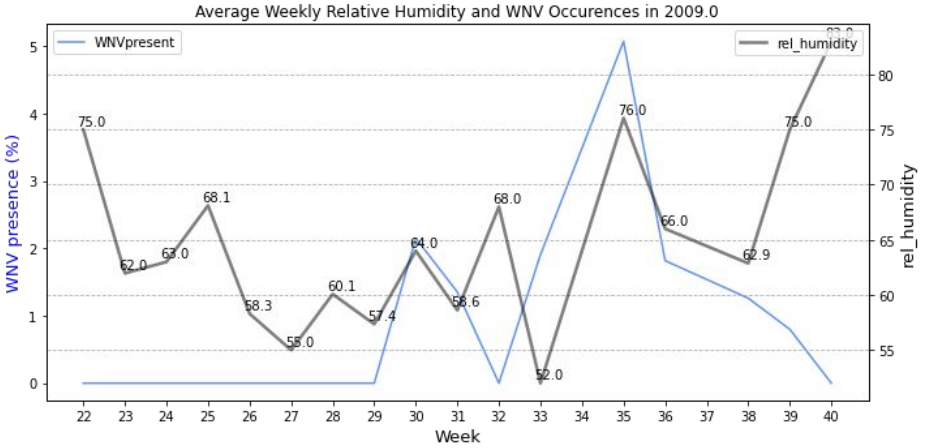
- Feature engineer time lag columns

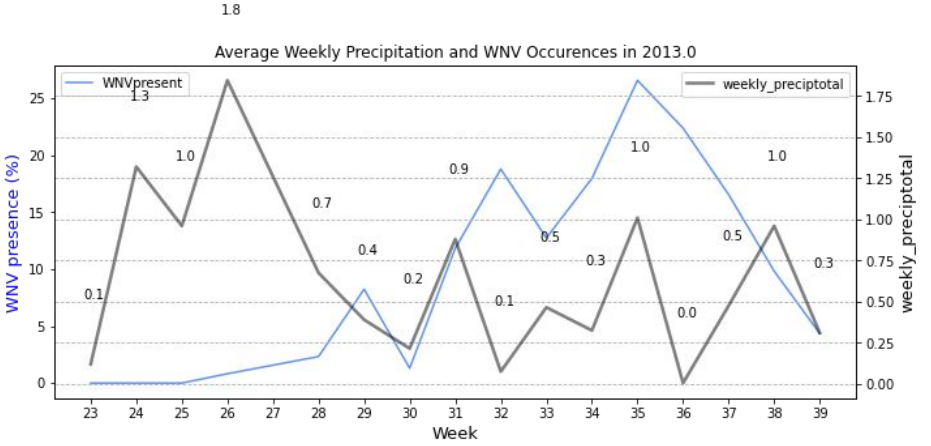




WNV
Average relative humidity/ Total precipitation
Time

- Potential lead indicators as well
- 2009 (humidity):
  - Week 28-30
  - Week 32-35
- 2013 (precipitation):
  - Week 31-32
  - Week 33-35
- Feature engineer time lag columns

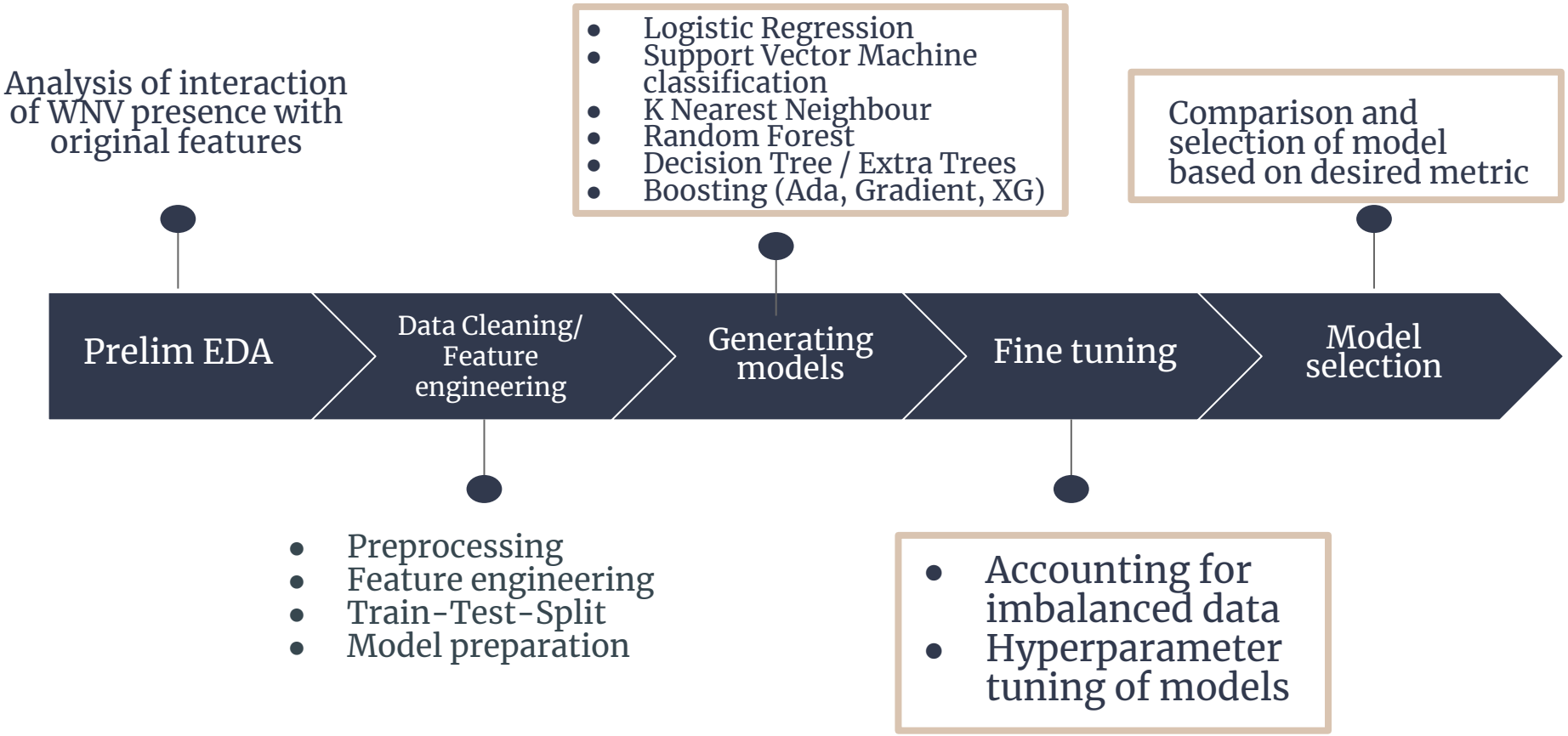




Feature engineering

- '*species*' converted to ordinal numeric
- '*codesum*' converted to binary
- '*tavg*' & '*dewpoint*' to celsius
- Relative humidity
  - Temperature
  - Dewpoint
- Weekly average temperature and total precipitation
- Time lag columns
  - Precipitation, avg_temp and relative humidity
- Polynomial Features

Data Modeling Workflow



Data Modelling

Baseline score

```
# Baseline score
combined_train['wnvpresent'].value_counts(normalize = True)
```

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

Models used:

1. Logistic Regression
2. K Nearest Neighbor
3. SVM
4. Decision Tree
5. Extra Trees
6. Random Forest
7. AdaBoost
8. Gradient Boost
9. XGBoost

Models with SMOTE

Model (SMOTE)	Train AUC	Test AUC	Precision	Specificity	Recall	F_score
AdaBoost	0.867	0.802	0.162	0.842	0.550	0.250
XGBoost	0.914	0.800	0.116	0.689	0.739	0.201
Logistic Regression	0.843	0.822	0.132	0.723	0.760	0.225
Random Forest	0.905	0.817	0.150	0.805	0.623	0.242
Extra Trees	0.885	0.815	0.144	0.787	0.644	0.235
GradientBoost	0.957	0.802	0.285	0.959	0.289	0.287
Decision Trees	0.826	0.776	0.129	0.781	0.586	0.212
SVC	0.878	0.799	0.136	0.773	0.644	0.225
KNearestNeighbor	0.961	0.706	0.158	0.855	0.492	0.240

Models w/o SMOTE

Model (Class weights)	Train AUC	Test AUC	Precision	Specificity	Recall	F_score
AdaBoost	0.891	0.834	0.750	0.999	0.021	0.042
XGBoost	0.942	0.827	0.149	0.806	0.615	0.241
Logistic Regression	0.844	0.827	0.134	0.722	0.775	0.228
Random Forest	0.937	0.822	0.162	0.840	0.557	0.252
Extra Trees	0.917	0.817	0.142	0.778	0.666	0.235
GradientBoost	0.992	0.799	0.301	0.979	0.159	0.208
Decision Trees	0.855	0.771	0.143	0.793	0.623	0.233
SVC	0.874	0.759	0.000	1.000	0.000	0.000
KNearestNeighbor	0.933	0.714	0.333	0.990	0.086	0.137

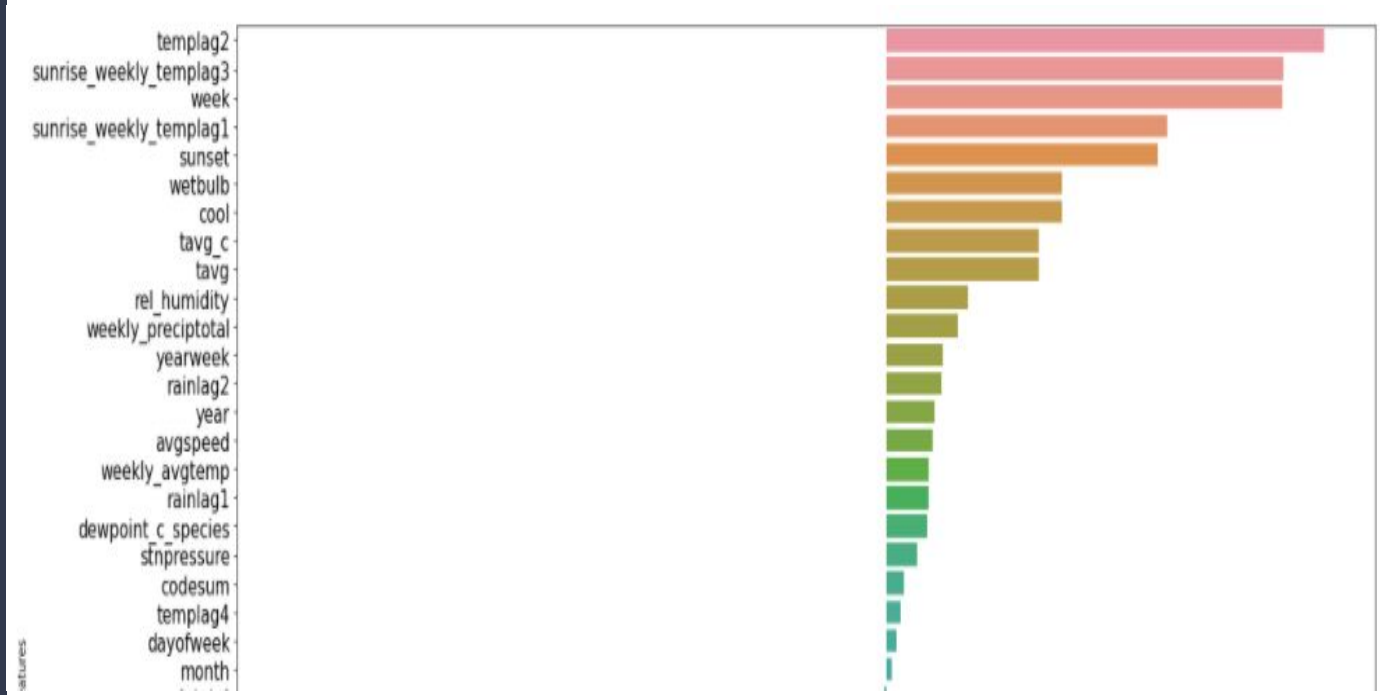
Model Evaluation

Final Model Selected: **Logistic Regression (no SMOTE)**

```
Fitting 5 folds for each of 8 candidates, totalling 40 fits
### BEST PARAMS ###

{'lr__C': 10, 'lr__class_weight': 'balanced', 'lr__solver': 'newton-cg'}

### METRICS ###

{'model': 'lr',
 'train_auc': 0.8446820096138815,
 'test_auc': 0.8272369440028882,
 'precision': 0.1342534504391468,
 'specificity': 0.7227802330253114,
 'recall': 0.7753623188405797,
 'f_score': 0.22887700534759356}

True Negatives: 1799
False Positives: 690
False Negatives: 31
True Positives: 107
```

Top Features

- Strong predictors are related to temp lag and temperature features
- Corroborates our earlier findings



At what cost?

Cost of Spraying

Sprayer trucks using Zenivex E4

- Spray duration of 5 hours
  - $844 - $1688 per truck

- Total area of Chicago
  - 606.1km2
- Total cost of spraying
  - $844k - $1.69mil

Medical Costs

Inpatient & Outpatient costs

- ≈ $33k per inpatient WNV case
- ≈ $6k per outpatient WNV case
- ≈ $18k per WNV patient (nursing home)

Productivity costs

- ≈ $11k per patient < 60 years old
- ≈ $8k per patient > 60 years old

Total cost

- Medical + productivity loss
  - ≈ $3.7mil

73%*

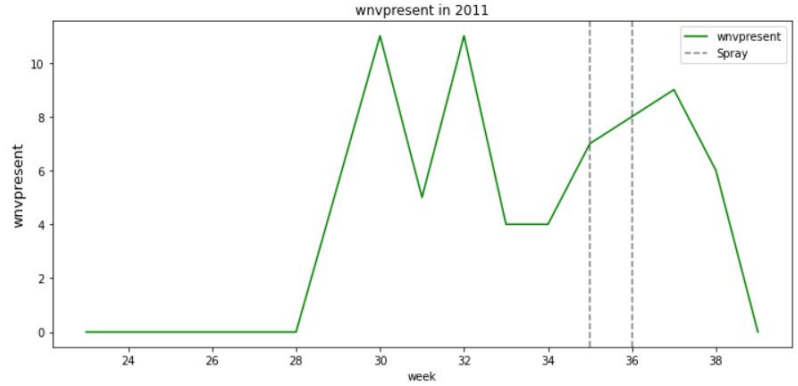
- Number of cases before spray program (2002)
  - 225 cases
  - 22 fatalities
- Number of cases after spray program (2012)
  - 60 cases
  - 4 fatalities

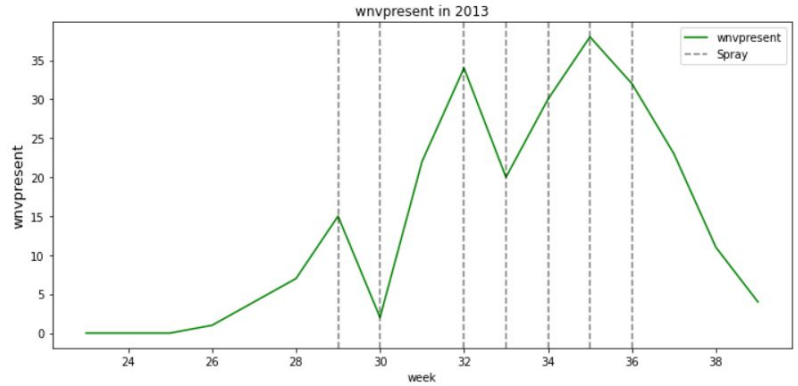
* Based on rough estimates of cases in 2002 vs 2012

https://www.chicago.gov/content/dam/city/depts/cdph/statistics_and_reports/CDInfo_2013_JULY_WNV.pdf

Cost Benefit Analysis:

Spray Efficacy





2013:

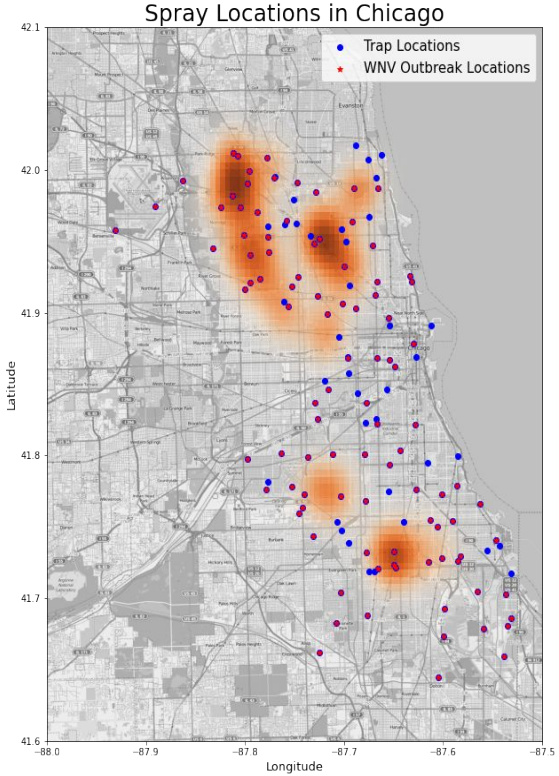
- Sprayings almost every week (week 29 to 36, from 17 July 2013 to 5 Sept 2013)

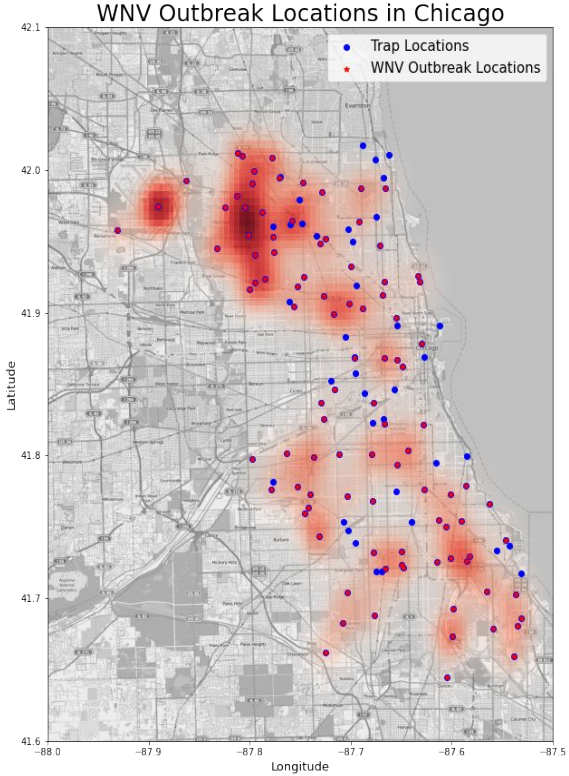
- WNV occurrences peak (weeks 29, 32, 35)
- Spraying done with reduction in WNV on next immediate weeks (weeks 30, 33, 36 and beyond)

- Although WNV increased again subsequently in-between the weeks aforementioned, pattern shows spraying curbs WNV

Cost Benefit Analysis:

Spray and WNV Locations





Accurate spraying locations on WNV locations:

- North-West from city (-87.65 , 41.95)
- Further North-West (-87.8 , 41.9 to 42)
- South, small spot (-87.65 , 41.7)

WNV locations inadequately or not covered by spraying:

- Furthest North-West (-87.9 , 42)
- South of city, huge scattered area (-87.5 to -87.8 , 41.6 to 41.8)
- South -West small spot inadequately/inaccurately covered (-87.7 , 41.8)

Recommendations

Recommendations

From cost benefit analysis

- Increase sprays to cover more WNV occurrence areas
- Reduce inaccurate sprayings
- Increase accurate sprays with the aid of prediction model and visualisation

Increase sprays

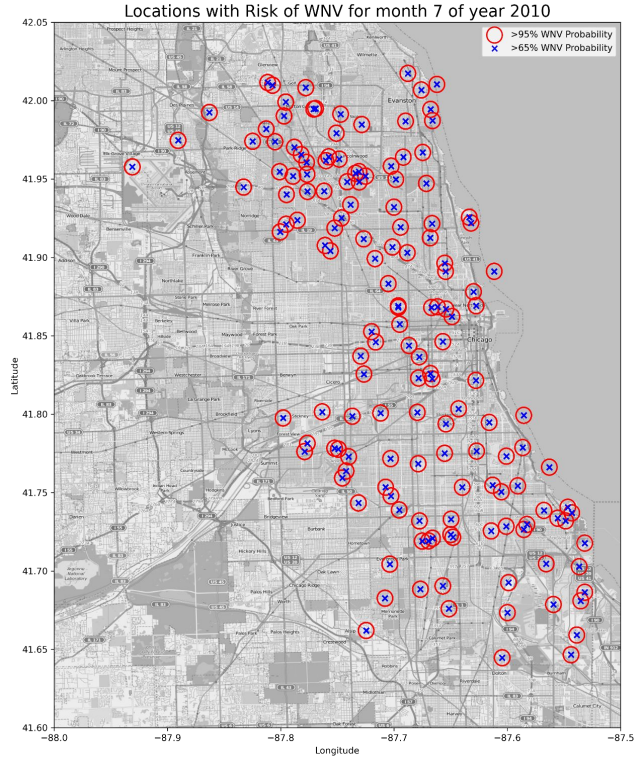
Increase spray spendage

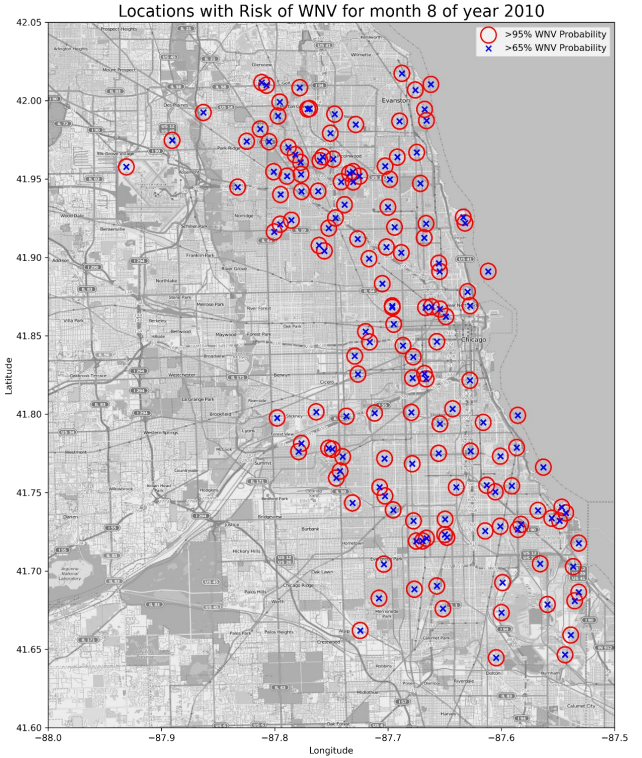
Decrease WNV cases

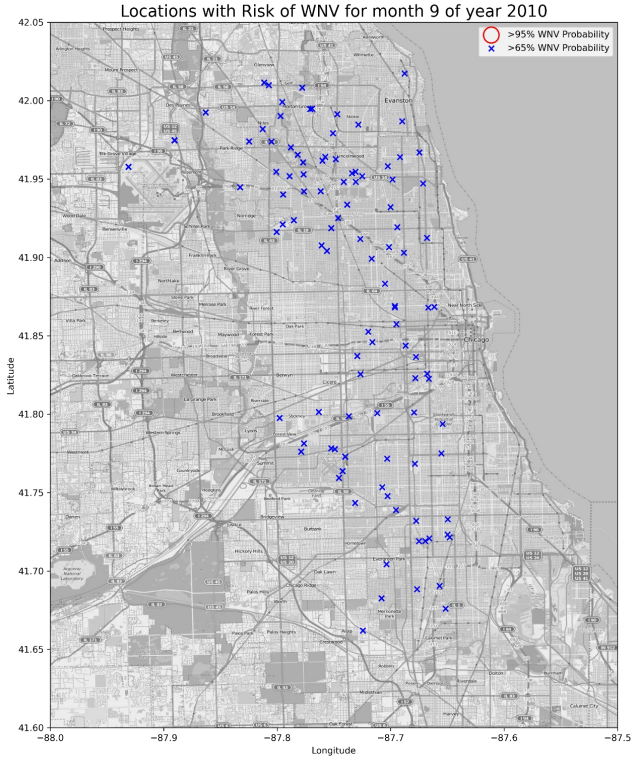
Decrease medical costs

Decrease overall costs

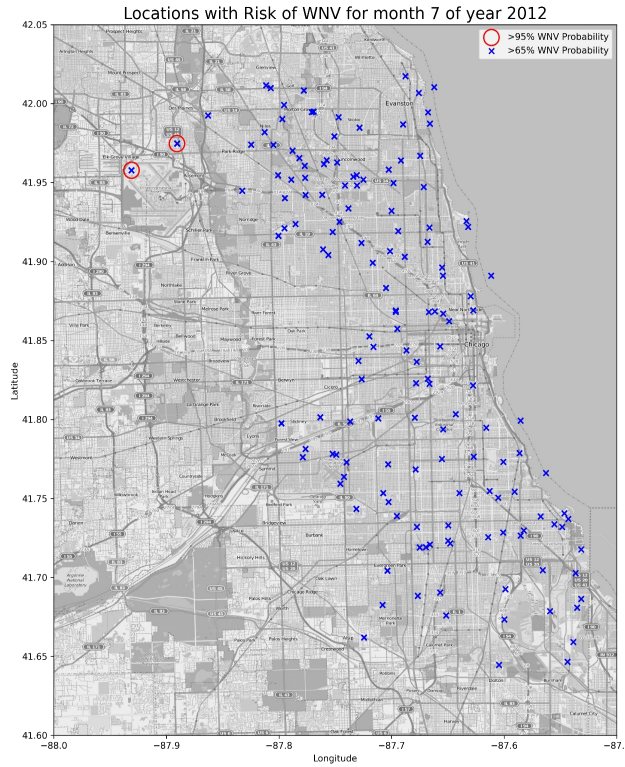
Recommendations: Where to spray? Year 2010

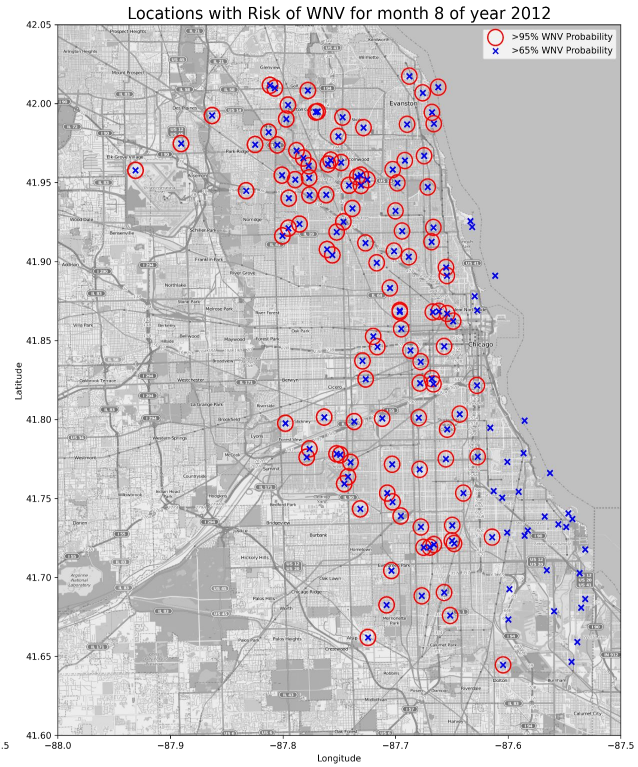


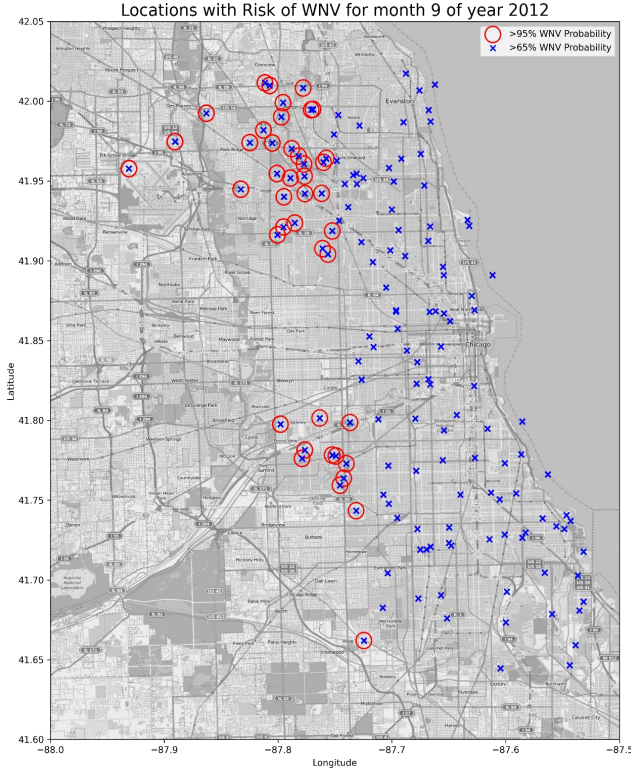




Recommendations: Where to spray? Year 2012







Model Limitations

Geographical limitations:

- Limited to only the city of Chicago

Dataset limitations:

- More years of data
- Consecutive years of data
- Spray dataset limitations

Thank you!