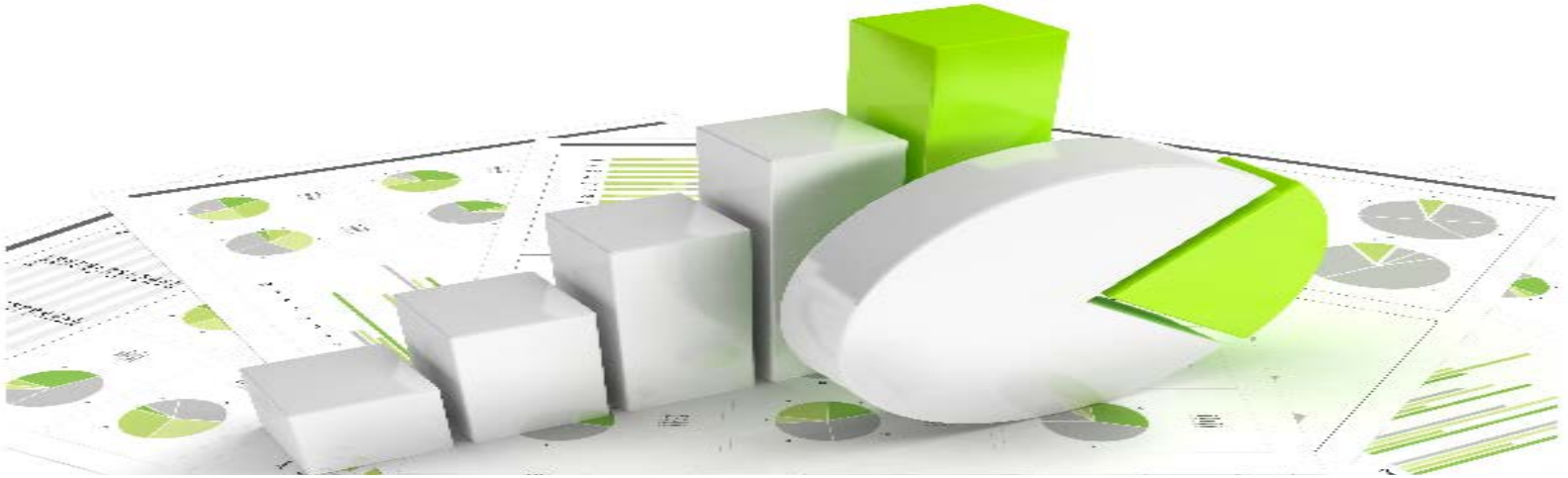


West Nile Virus Prediction

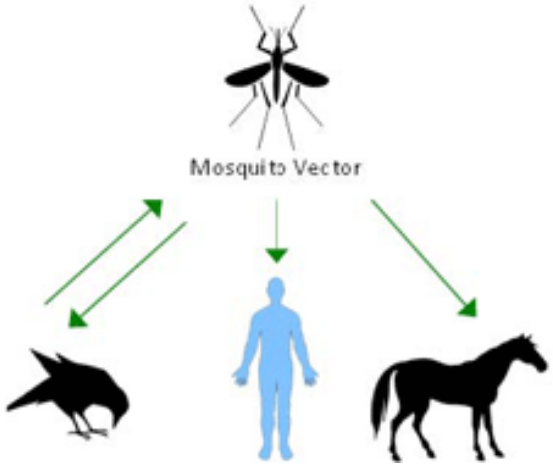
Machine Learning to predict out break of West Nile Virus in mosquitoes in the City of Chicago

Qiaolin Chen, Ph.D. 15 Aug 2017

West Nile Virus and Mosquitoes

To improve the treatment of patients suffering from Parkinson's Disease

- West Nile Virus (WNV):
  - Potentially deadly virus; cause seasonal epidemic summer to fall
- WNV symptoms:
  - High fever, headache, neck stiffness, stupor, disorientation, coma, tremors, convulsions, muscle weakness, vision loss, numbness
- Spread of WNV – bit of infected mosquitoes
- Chicago: surveillance and control
  - Set up mosquito traps and test for the virus
  - Spray airborne pesticides - when and where



Project Overview

Predictive Models for WNV Outbreak

- Goal: Given weather, location and testing, predict when and where different species of mosquitoes will test positive for WNV

1. Data wrangling
2. Feature generation and selection
3. Exploratory analysis
4. Model building and optimization
5. Insights and suggestions for the City of Chicago

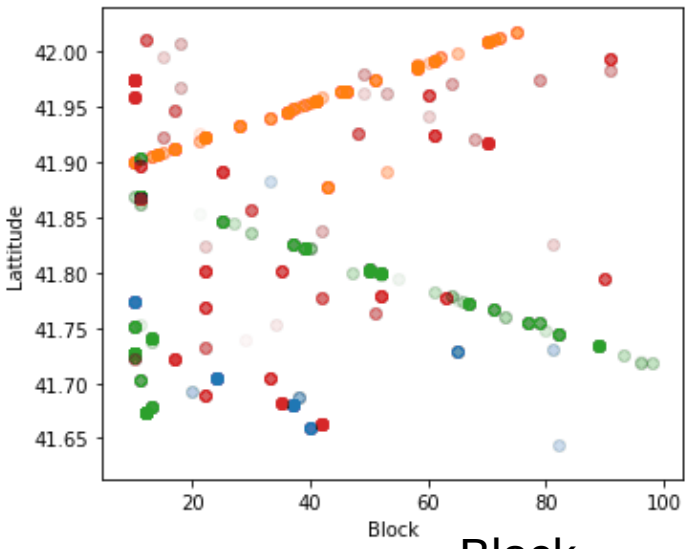
Data Description

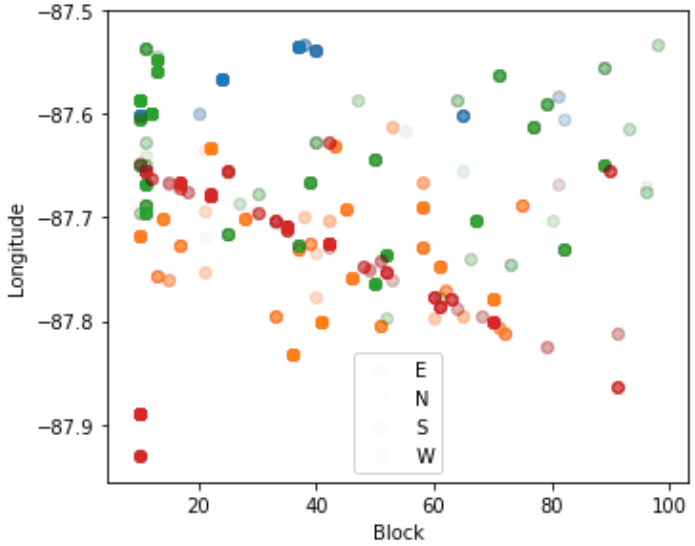
- Testing Data (2007, 2009, 2011 and 2013 May - Oct)
  - Label: WnvPresent - Whether WNV was present in mosquitoes
  - Predictors: Test date, the species of mosquitoes, number of mosquitoes caught, ID of the trap, and trap location variables (latitude and longitude , approximate address, block number, street name).
- Weather Data:
  - Hot and dry conditions are more favorable for WNV
  - NOAA weather conditions of 2007 to 2014
- Map Data: primarily provided for use in visualizations.
- Spraying Data: Date and location of spraying. Not used

Data Cleaning – Testing Data

- Testing data
  - Summarize categorical and numeric variables for abnormal values
  - Predictors: Test date, number of mosquitoes caught, trap latitude and longitude, street name
  - New features by trap/street: Total num of mosquitoes, percent of 3 main species

Lattitude

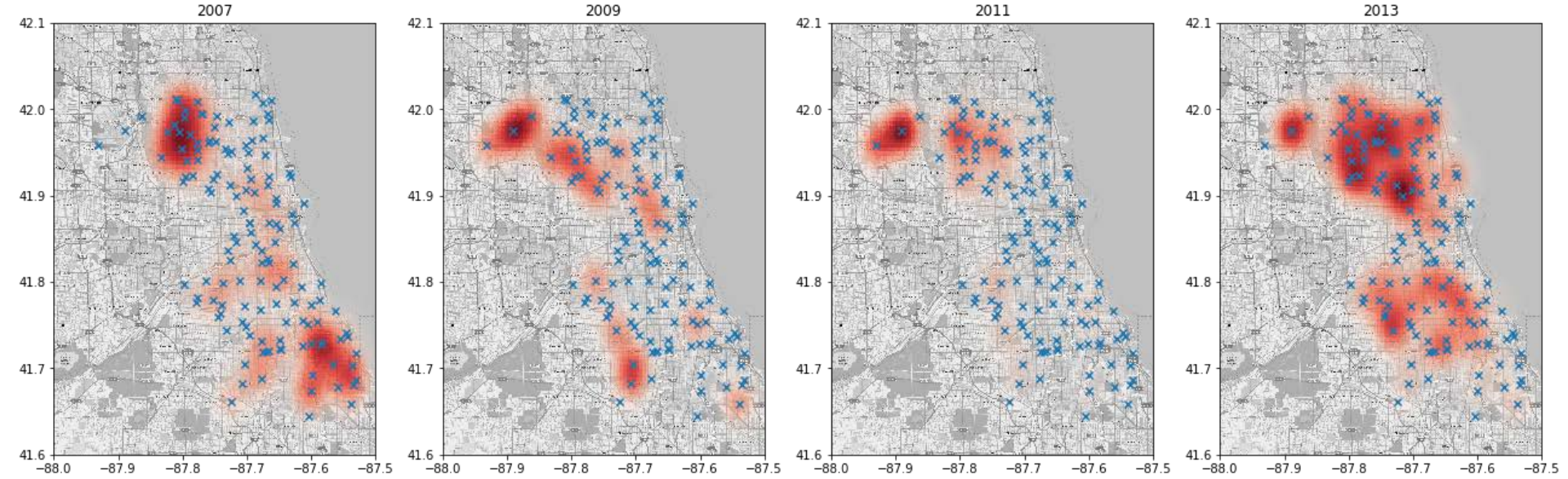




Block

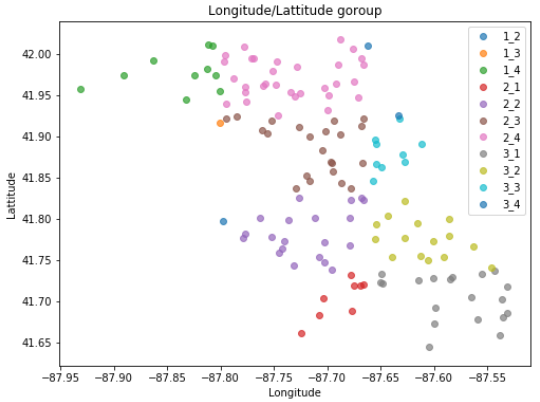
Feature Engineering – Testing Data

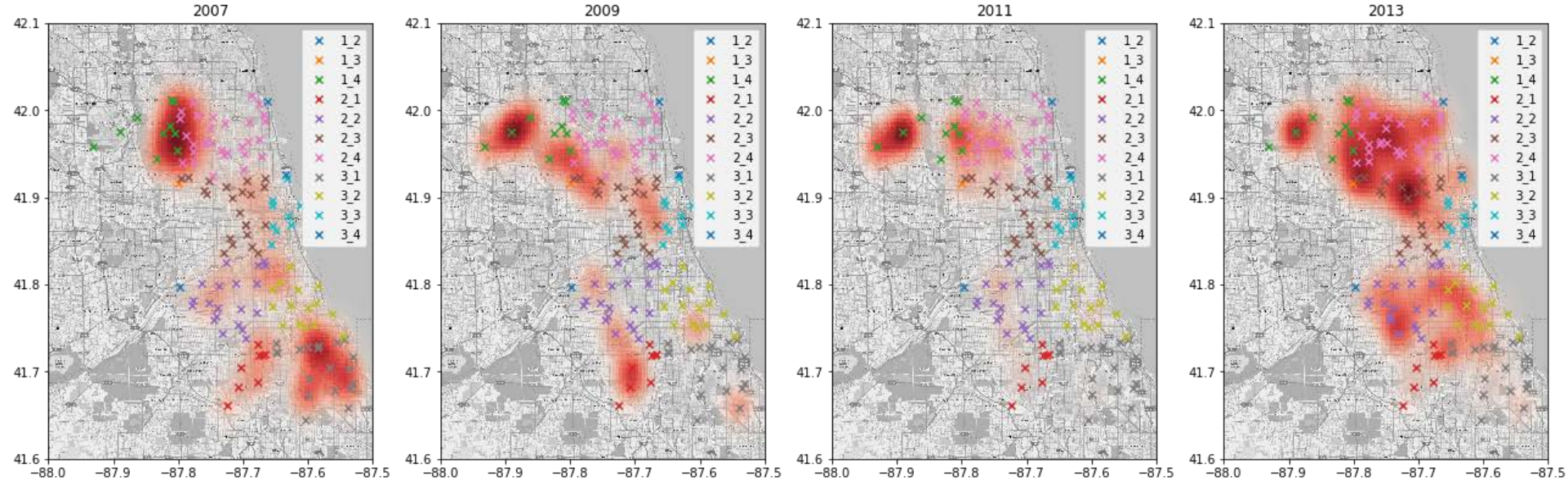
- How to represent loacation:
  - Address/Street name: 138/128 unique
  - Lattitude and longitude: Effect not linear
- Look at the heatmap of WnvPresent



Feature Engineering – Testing Data

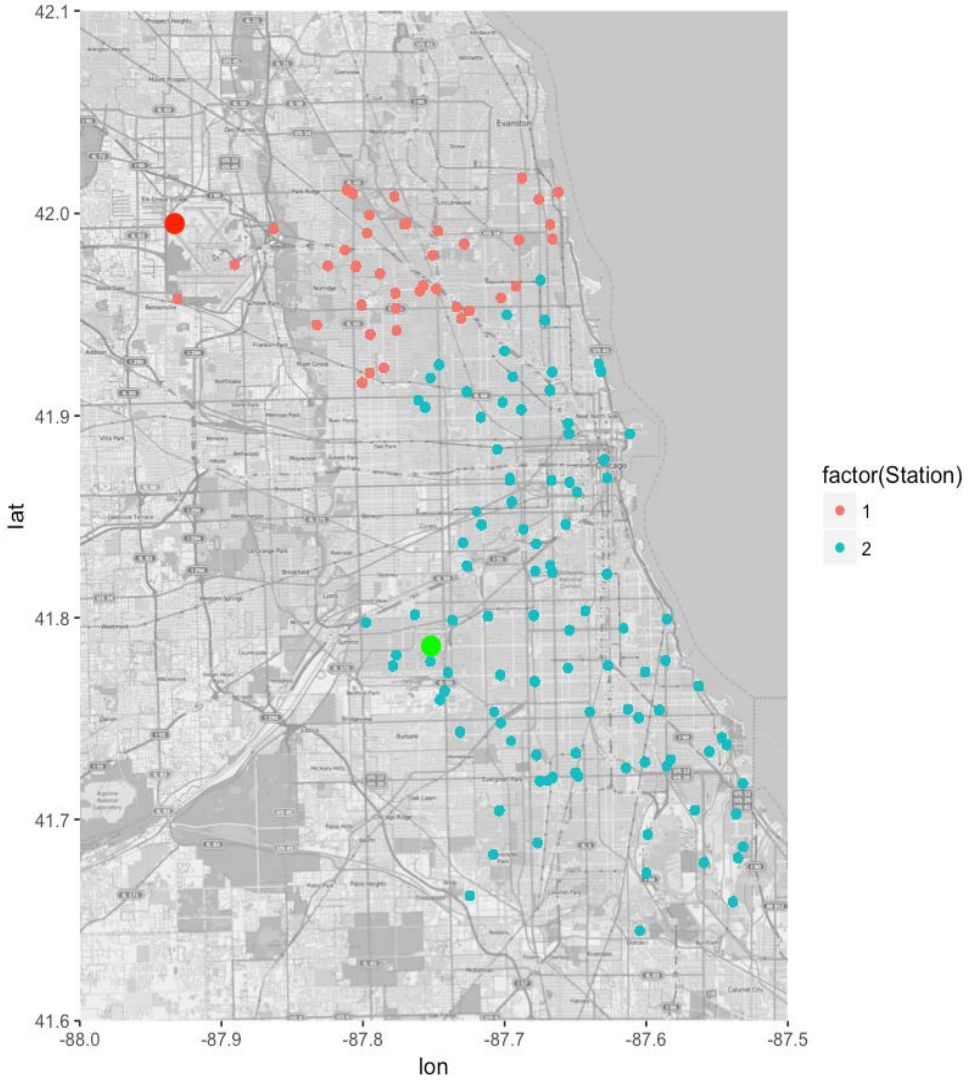
- Groups by lattitude and longitude:
  - Lattitude: 4 groups
  - Longitude: 3 groups
- Overlay on the heatmap of WnvPresent



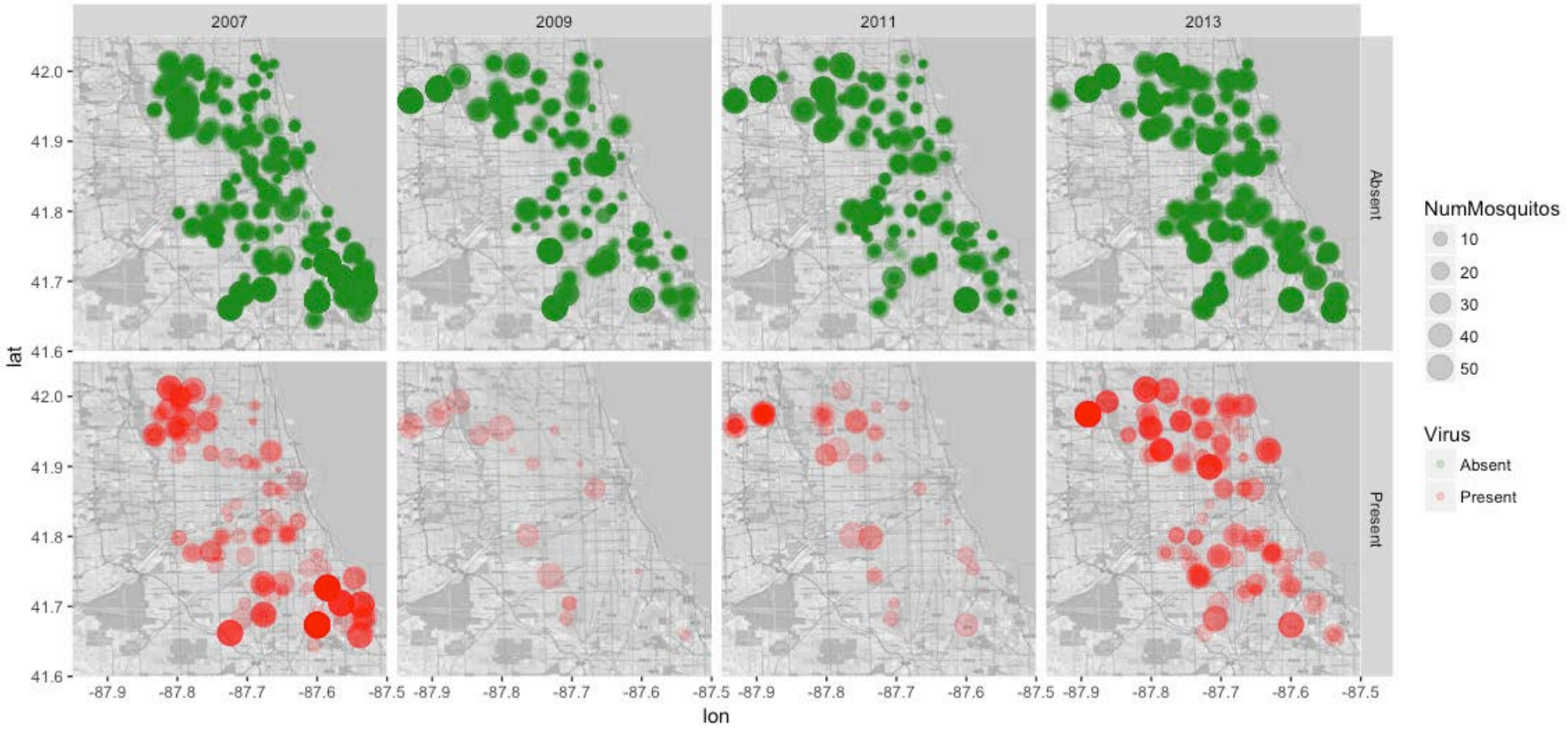


Data Cleaning – Merging with Weather Data

- Impute missing data:
  - ‘M’-> nan, ‘T’-> 0.005
- Remove features:
  - Snow fall, sea level, sunrise, sunset, wind speed, etc.
- Average data from two weather stations
- Moving window stats:
  - Take moving average 5D
  - Moving average/sum 30D high temp/precipitation



Build prediction model



Classification Model

- Shuffle data and split to training (80%) and validation (20%) sets
- Construct a custom transformer that will do one-hot encoding for categorical features
- Build a ML pipeline: transform data, optimize model (GridSearchCV) and predict new data
- Compare accuracy, F1 score, precision, recall and ROC

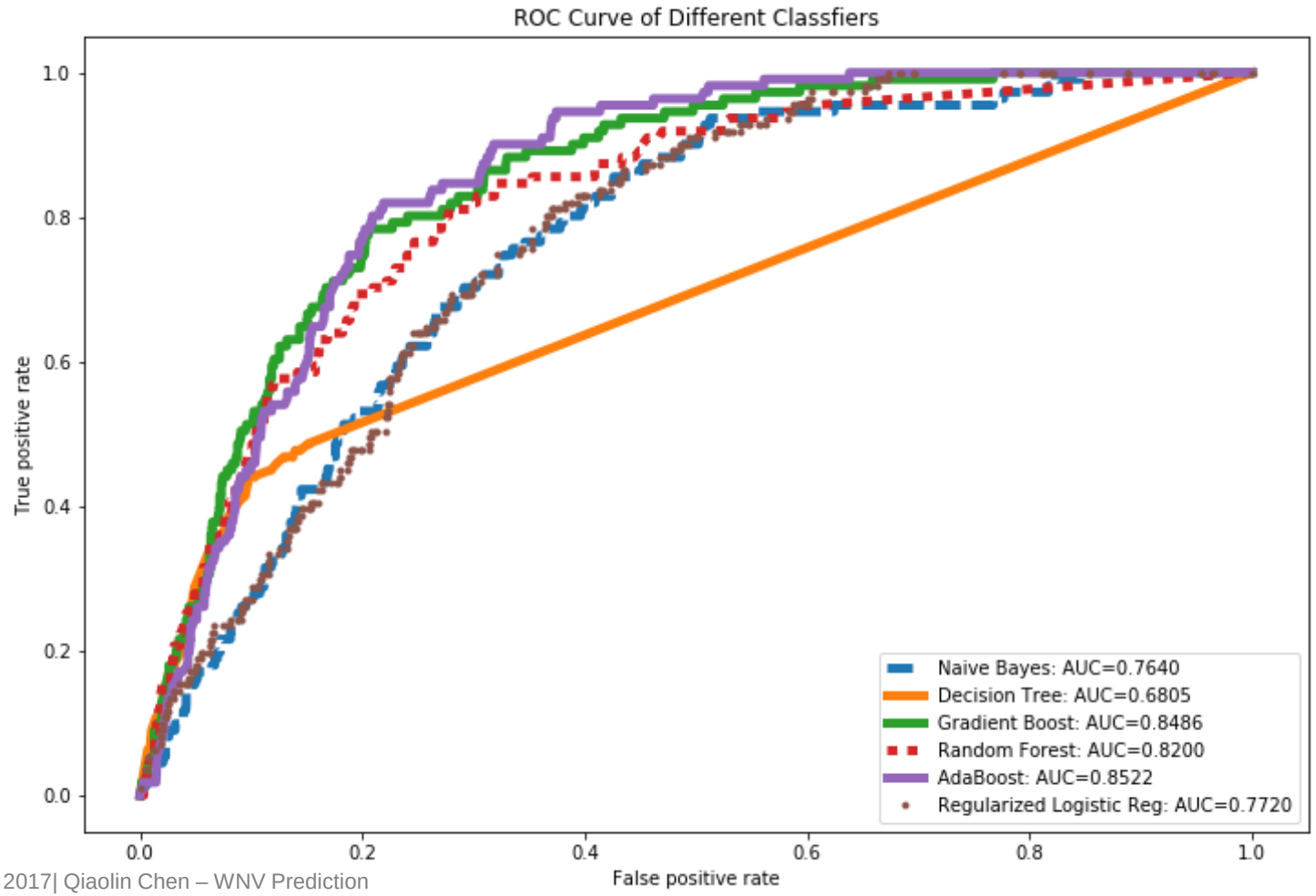
Classification Model Comparison

- Evaluation Metrics:
  - Accuracy: not appropriate, WNV yes =5.2%
  - AUC of ROC , Precision, Recall and F-1 score

Model	AUC	Test F1 Score	Test Precision	Test Recall	Test Accuracy
Naive Bayes	0.764	0.208	0.123	0.676	0.728
Decision Tree	0.681	0.154	0.267	0.108	0.937
Gradient Boosting	0.849	0.035	0.400	0.018	0.947
Random Forest	0.820	0.017	0.125	0.009	0.944
AdaBoost	0.852	--	--	--	0.9472
Regularized logistic regression	0.772	--	--	--	0.9472

Model Selection

- 'Total_Mosq', 'Pip_pct', 'PR_pct', 'Res_pct', 'Tmax', 'Tmin', 'Tavg', 'DewPoint', 'WetBulb', 'PrecipTotal', 'StnPressure', 'Tmax_sum30d', 'PrecipTotal_30d', 'Species', 'Month', 'lat_long'



Classification Model Comparison

- Accuracy is not a good metric: imbalance, p is small
  - All models (except Naive Bayes) have similar accuracy scores close to predicting no WNV at all (0.9477).
  - Naive Bayes have lower accuracy score, particularly when the number of features is large.
- Logistic regression:
  - Not good. predicting No for all and have low AUC.
- High AUC Models (>0.82):
  - Gradient Boosting(0.85), AdaBoost(0.85) and Random Forest(0.82).
- Final model: gradient boosting classifier

Feature Importance from Gradient Boosting Model

Feature	Importance
Total_Mosq	0.102874
PrecipTotal_30d	0.086703
Tmax_sum30d	0.082222
PR_pct	0.065226
Tmax	0.062006
Pip_pct	0.056996
Res_pct	0.056568
StnPressure	0.055900
Month_8	0.049460
PrecipTotal	0.049146
lat_long_1_4	0.045494
DewPoint	0.041716

Feature	Importance
WetBulb	0.031591
Species2_CULEX RESTUANS	0.029403
Species2_OTHER	0.026995
Species2_CULEX PIPIENS	0.024117
lat_long_2_1	0.020604
Species2_CULEX PIPIENS/RESTUANS	0.019385
lat_long_2_4	0.019154
Tavg	0.018433

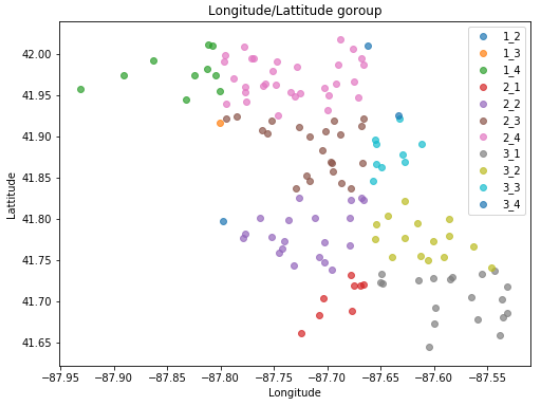
Suggestion to Chicago Dept of Public Health

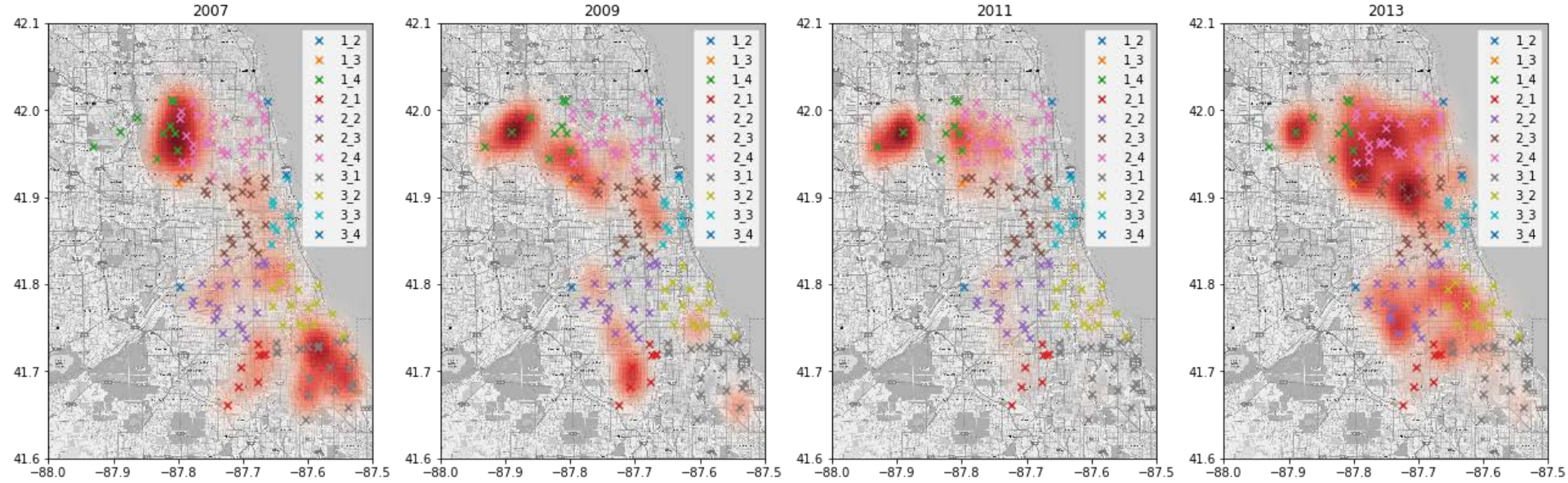
Monitor High Risk Region Closely

	NumMosquitos			WnvPresent (%)		
lat_long	count	sum	mean	count	sum	mean
1_4	1557	27729	17.8	1557	141	9.1
3_1	2025	48348	23.9	2025	108	5.3
2_1	708	9544	13.5	708	37	5.2
2_4	1717	13076	7.6	1717	87	5.1
2_3	1446	12316	8.5	1446	66	4.6
2_2	1382	12879	9.3	1382	62	4.5
3_2	974	6657	6.8	974	37	3.8
3_3	697	4490	6.441894	697	13	1.9

Suggestion to Chicago Dept of Public Health

- Focus on areas: Lattitude: 4 groups
  - Longitude: 3 groups
- Overlay on th





Suggestion to Chicago Dept of Public Health

- When to spray:
- August is the moth with most mosquitoes and most positive tests, follow by September

Month	NumMosquitos			WnvPresent (%)		
	count	sum	mean	count	sum	mean
5	84	230	2.74	84	0	<0.01
6	1571	16578	10.55	1571	1	0.06
7	2606	37248	14.29	2606	46	1.8
8	3751	58036	15.47	3751	377	10.0
9	2218	21029	9.48	2218	125	5.6
10	276	1918	6.95	276	2	0.7

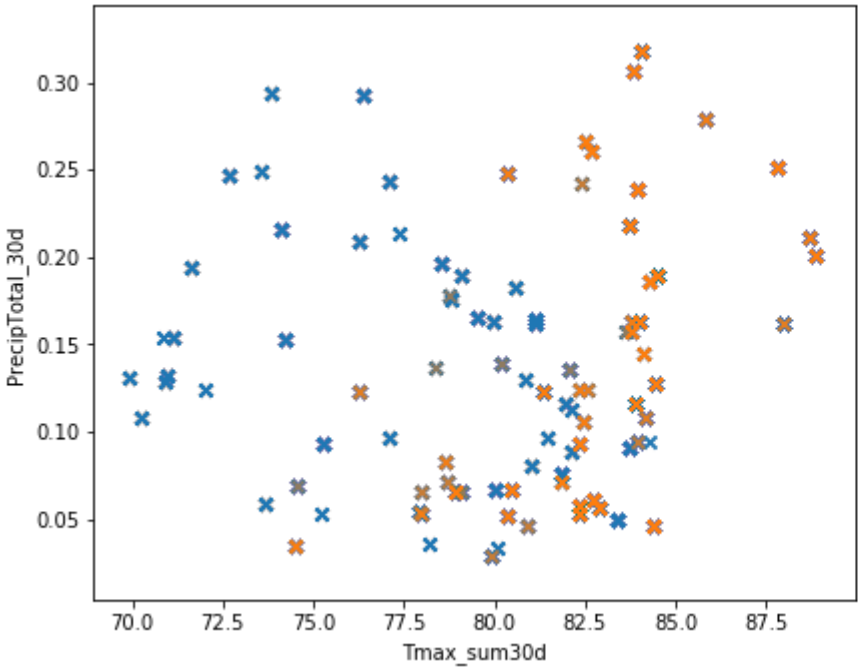
Suggestion to Chicago Dept of Public Health

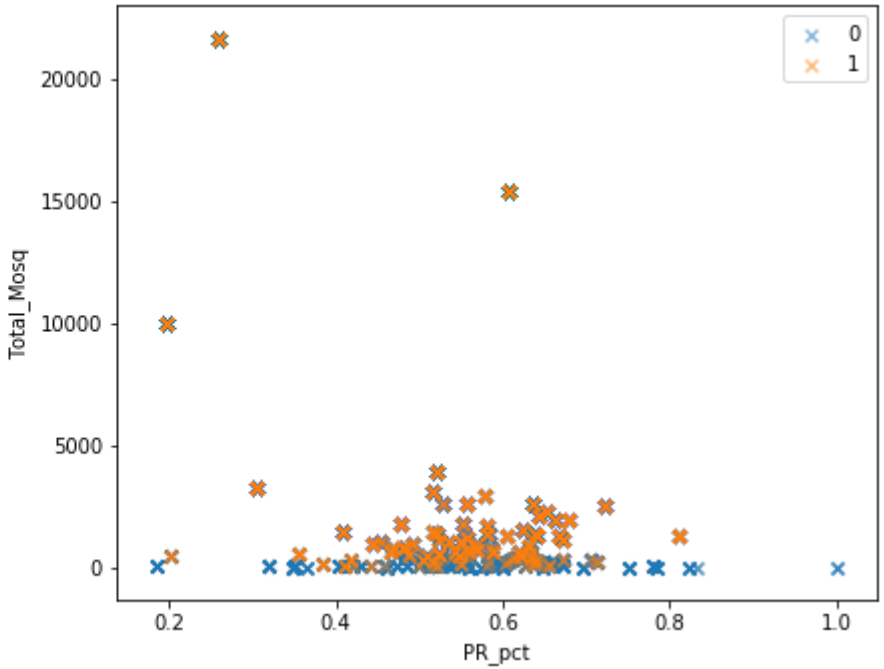
- Which Species:
- CULEX PIPIENS: highest probability of WNV test positive

Species	NumMosquitos			WnvPresent (%)		
	count	sum	mean	count	sum	mean
CULEX PIPIENS	2699	44671	16.55	2699	240	8.9
CULEX PIPIENS/RESTUANS	4752	66268	13.95	4752	262	5.5
CULEX RESTUANS	2740	23431	8.55	2740	49	1.8
OTHER	315	669	2.12	315	0	<0.01

Suggestion to Chicago Dept of Public Health

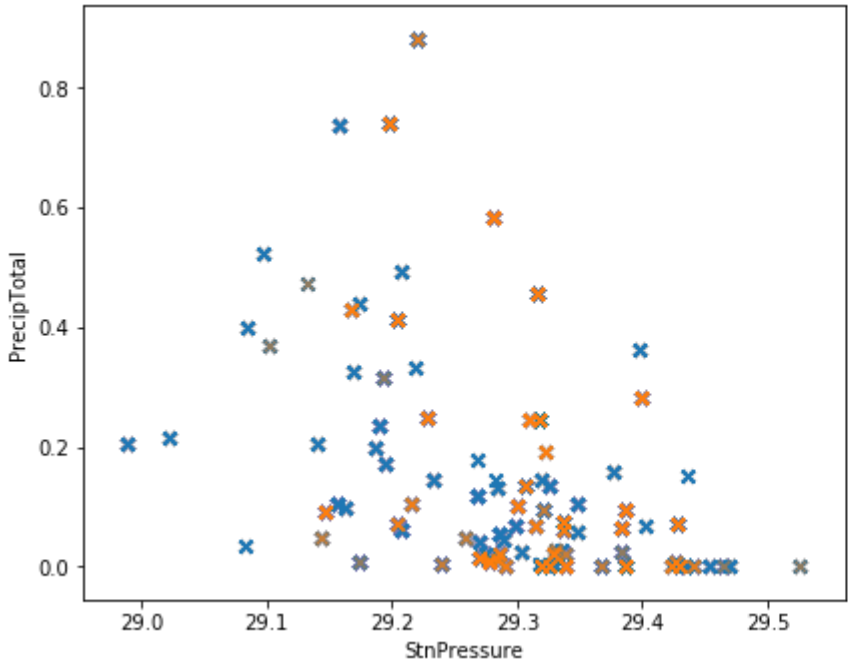
- Hot and dry conditions are more favorable for WNV than cold and wet
- Traps with high total mosquito counts have high probabilities of WNV test positive

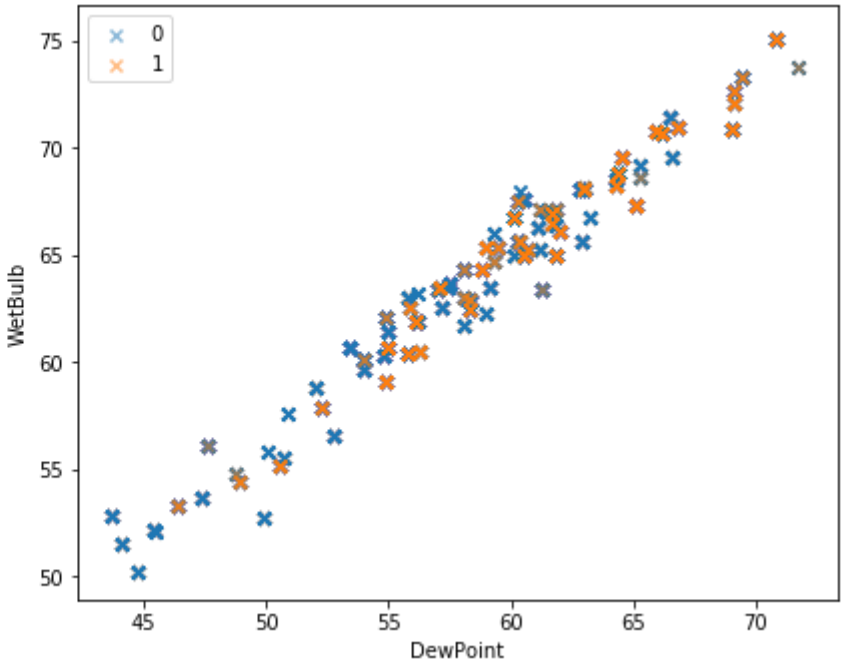




Suggestion to Chicago Dept of Public Health

- Low precipitation and high pressure are favorable
- Dew point and wet bulb are high correlation





Summary of West Nile Virus Prediction Project

- Goal: Given weather, location and testing, predict when and where different species of mosquitoes will test positive for WNV

1. Data wrangling
2. Feature generation and selection
3. Exploratory analysis
4. Model building and optimization
5. Insights and suggestions for the City of Chicago