Project 4:
Predict West Nile Virus

Geh Si Rong, Lee Meng Chin,
Teo Zhan Rui, Wee Zi Jian

TABLE OF CONTENTS

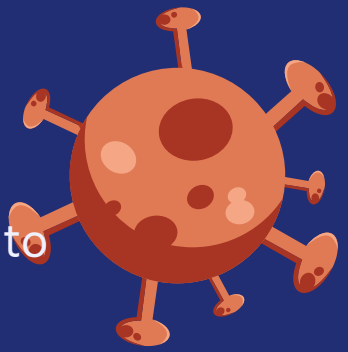
Problem Statement

The West Nile virus (WNV) is a mosquito-borne illness that can cause severe neurological disease and death in humans.

Since 2004, the Chicago Department of Public Health has increased surveillance and control efforts in a bid to prevent transmission of this virus.

Given weather, location, testing, and spraying data, our goal is to **predict whether the WNV is present** in a given location.

Based on our predictions, we will devise a cost effective strategy to deploy pesticides in WNV-hotspots.



Pre-Processing: Train / Test

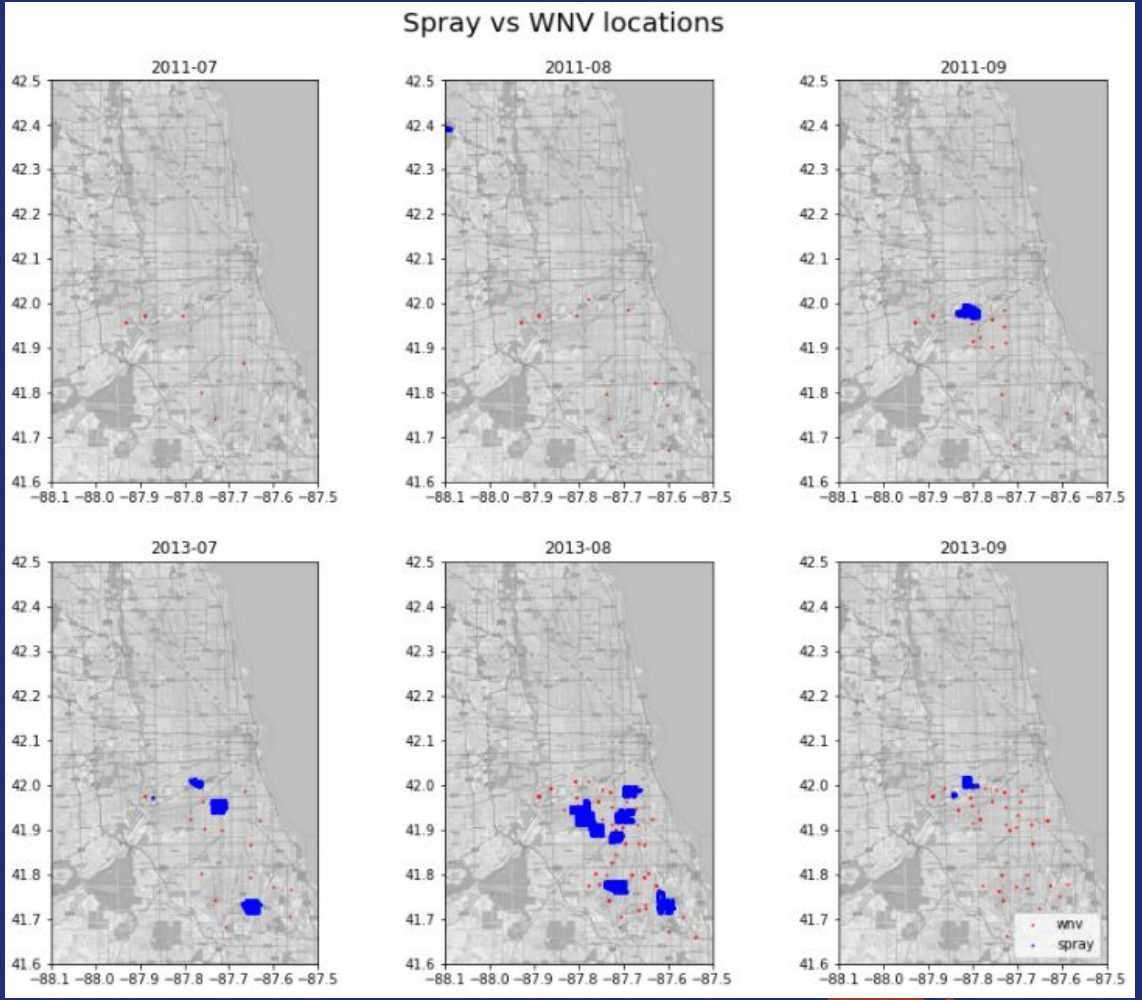
- Train: 10,506 rows, 12 columns (2007, 2009, 2011, 2013)
- Test: 116,293 rows, 11 columns (2008, 2010, 2012, 2014)
- Relevant columns:
  - Date
  - Species
  - Longitude
  - Latitude
  - WNV present
- Set date as index
- Assign nearest weather station to each trap
- Group by mosquito species
- Convert species to categorical features

Pre-Processing: Weather

- 2,944 rows, 22 columns
- Daily data from May-October 2007-2014
- Impute missing values (‘M’) and trace values (‘T’) with 0 or mean
- Convert weather conditions (CodeSum) to categorical variables
- Compute 14 day rolling average/sum of various weather data
- Compute lagged (3, 5, 7, 10 days) versions of rolling weather data
- Assign weather data to train/test data based on nearest weather station

Pre-Processing: Spray

- 14,835 rows, 4 columns
- 2011 and 2013 spray locations and dates
- Based on plots of sprayed locations vs WNV presence, spraying does not appear to reduce WNV presence in subsequent months



Pre-Processing: Spray

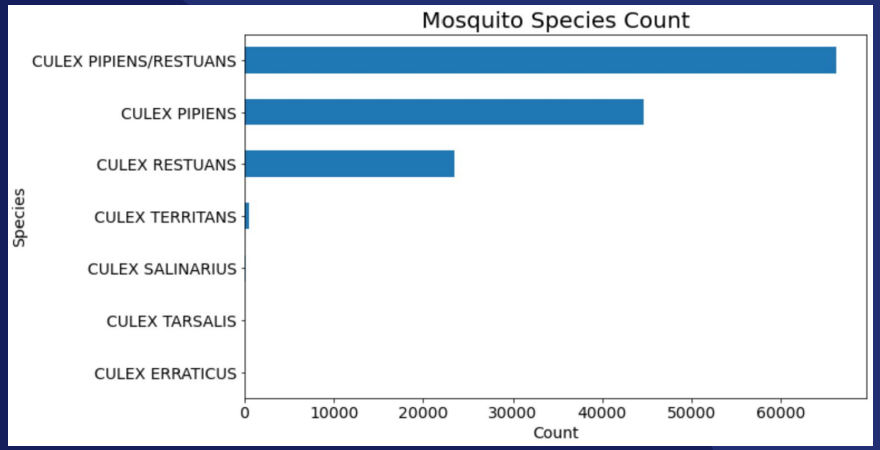
- 2011 and 2013 train data locations were checked if they had been sprayed within a certain radius within the past 10 days
- Spraying within 10m, 30m and 50m of a location within the last 10 days has **marginal effect** on the number of mosquitoes caught or the presence of WNV
- Since spray data for 2008, 2010, 2012 and 2014 is unavailable as well, **spray data will not be used** in modelling

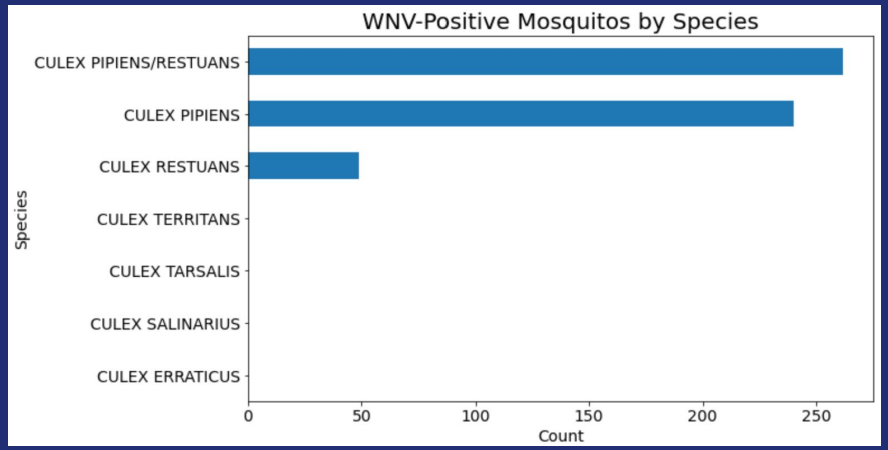
Sprayed radius within last 10 days vs WNV

	wnv	wnv_binary	num_mos
sprayed_10m_binary			
0	0.077421	0.064742	14.662800
1	0.115385	0.115385	11.384615
sprayed_30m_binary			
0	0.077114	0.064170	14.647205
1	0.103896	0.103896	13.370130
sprayed_50m_binary			
0	0.076364	0.063497	14.640280
1	0.109524	0.104762	13.828571

EDA: Mosquito Counts

- 3/7 species found with WNV
  - Most frequently caught species
- Species expected to be an important feature in predicting WNV



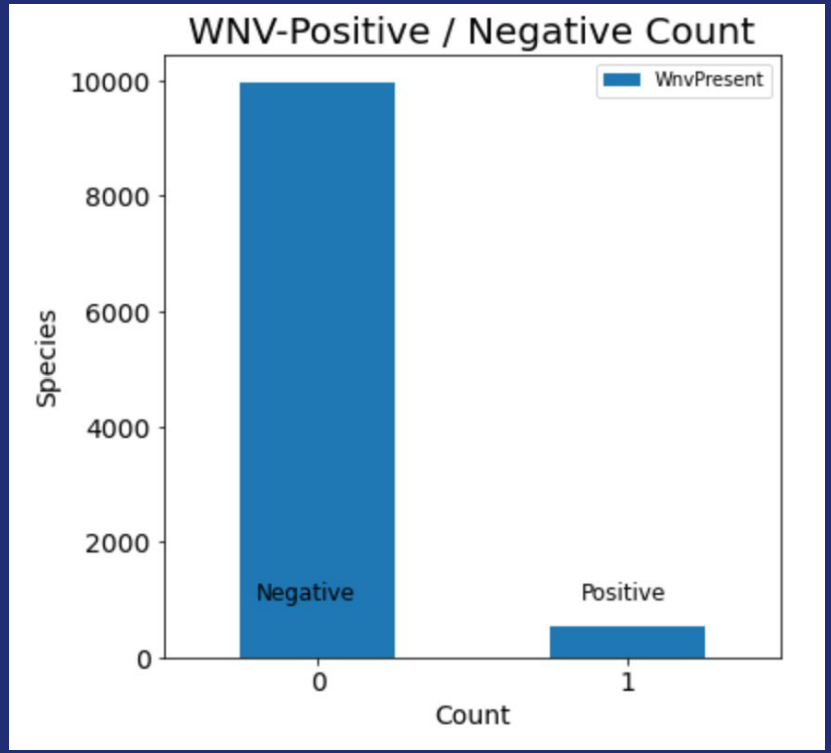


EDA: WNV Positive/ Negative Counts

- WNV-negative vs WNV-positive

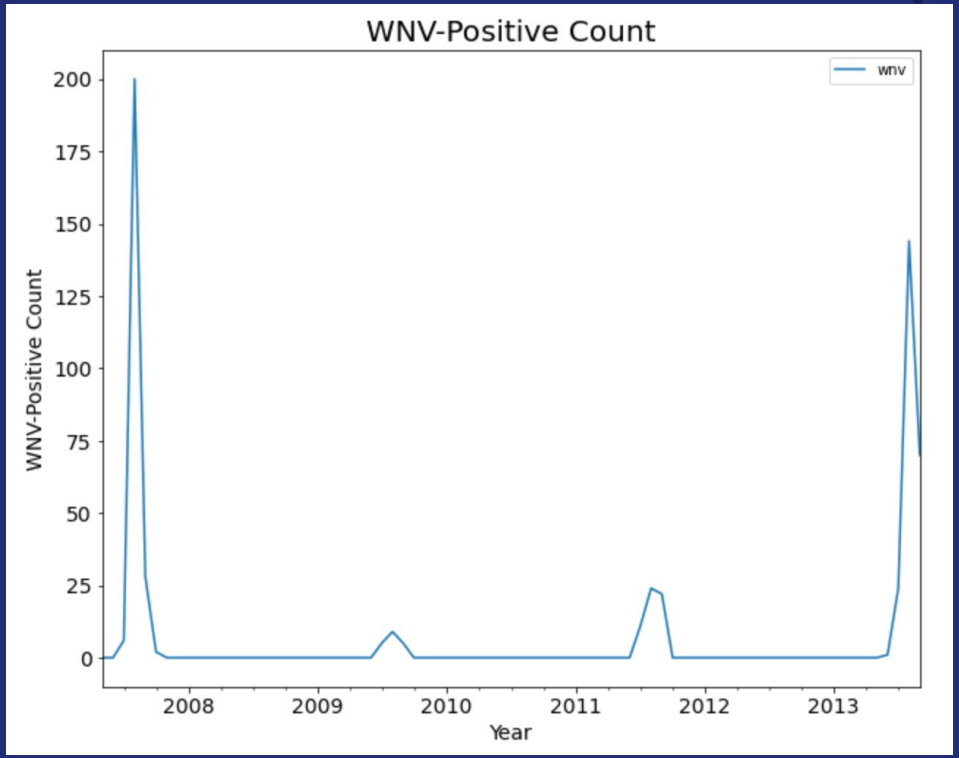
  9,955 **551**

- Imbalanced classes
  - Minority class was resampled using SMOTE

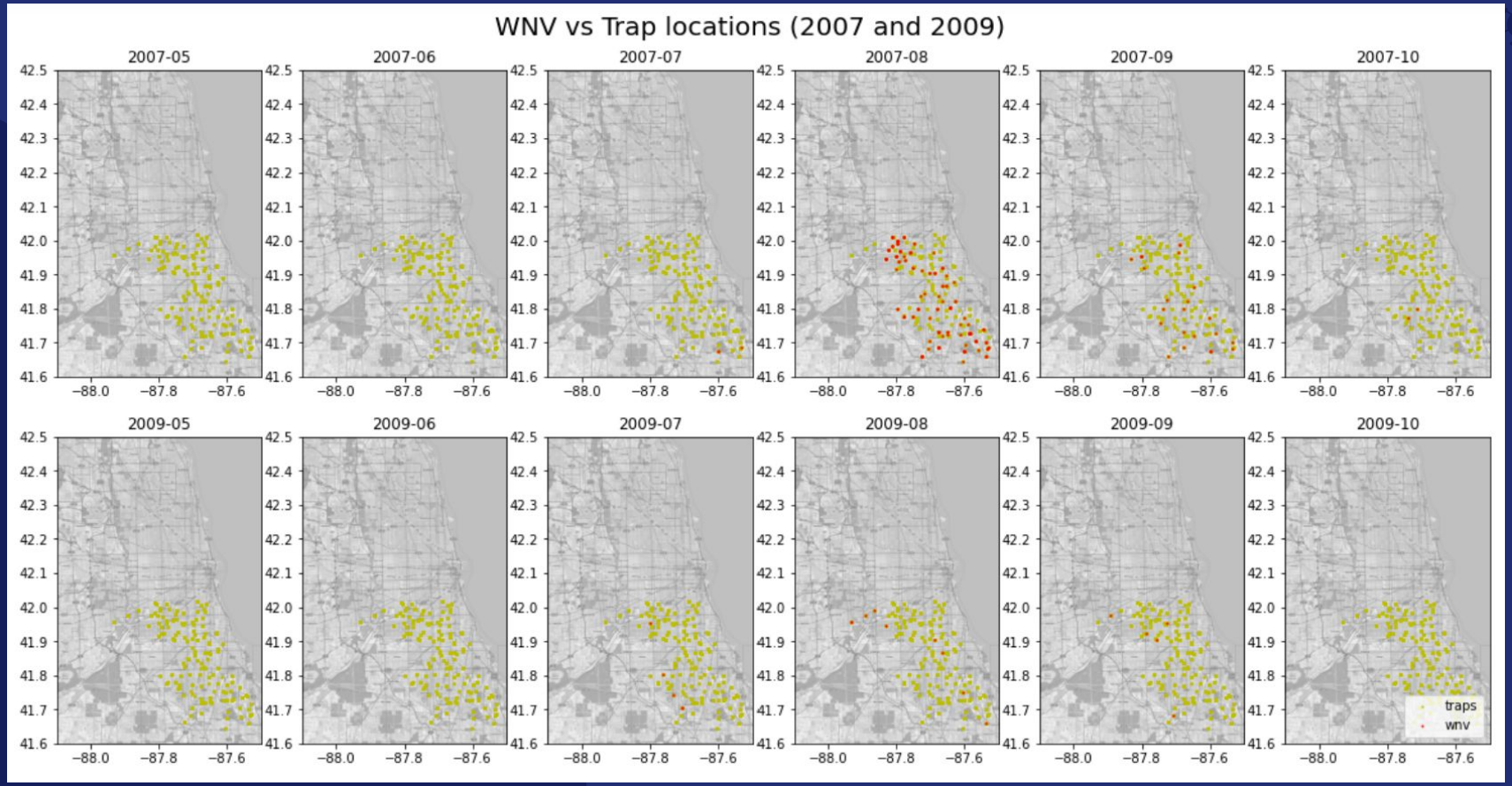


EDA: WNV-Positive by Date

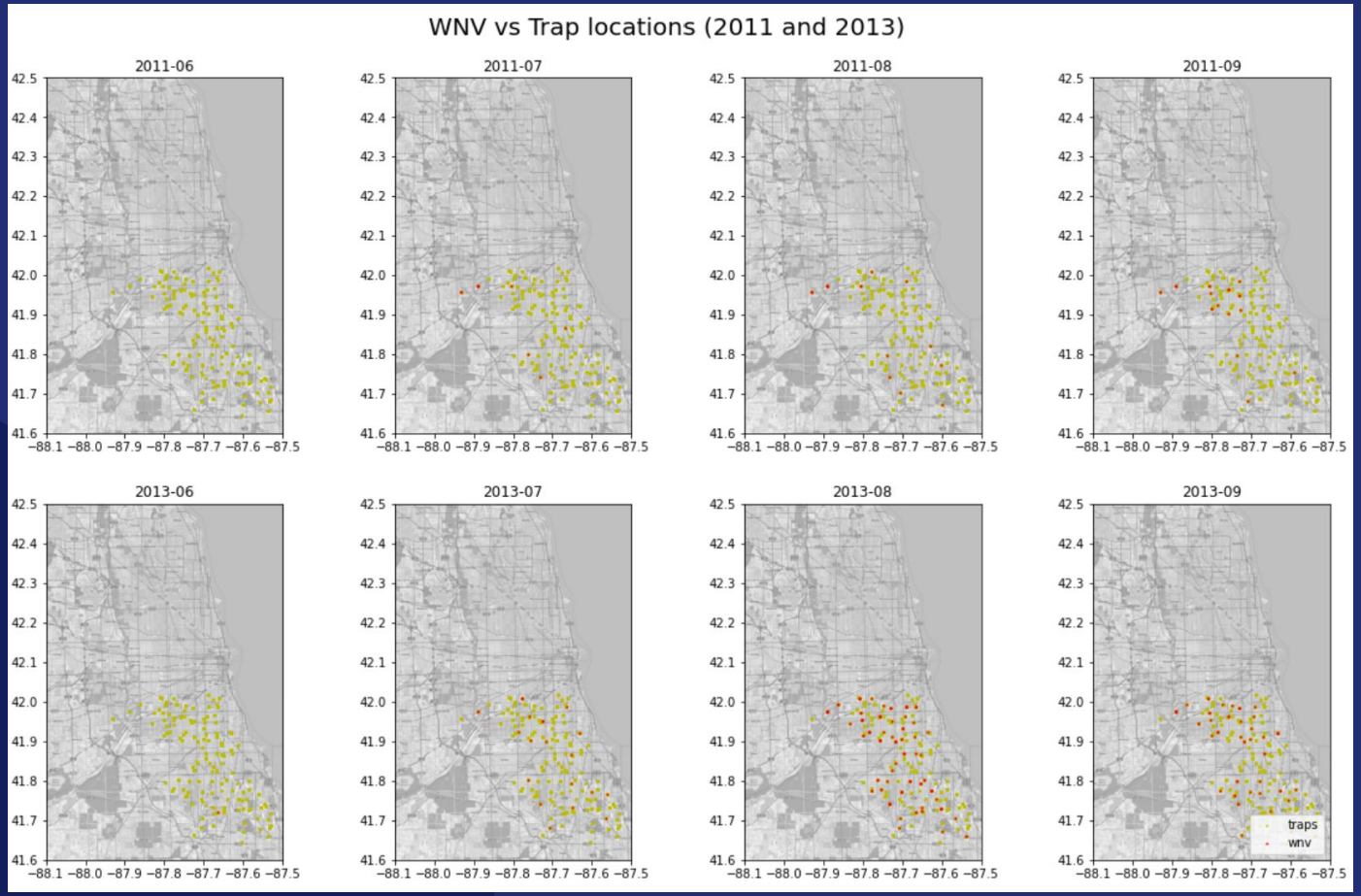
- WNV-Season: Jul to Sep
- Huge spikes in 2007 and 2013



EDA: WNV-Positive by Location

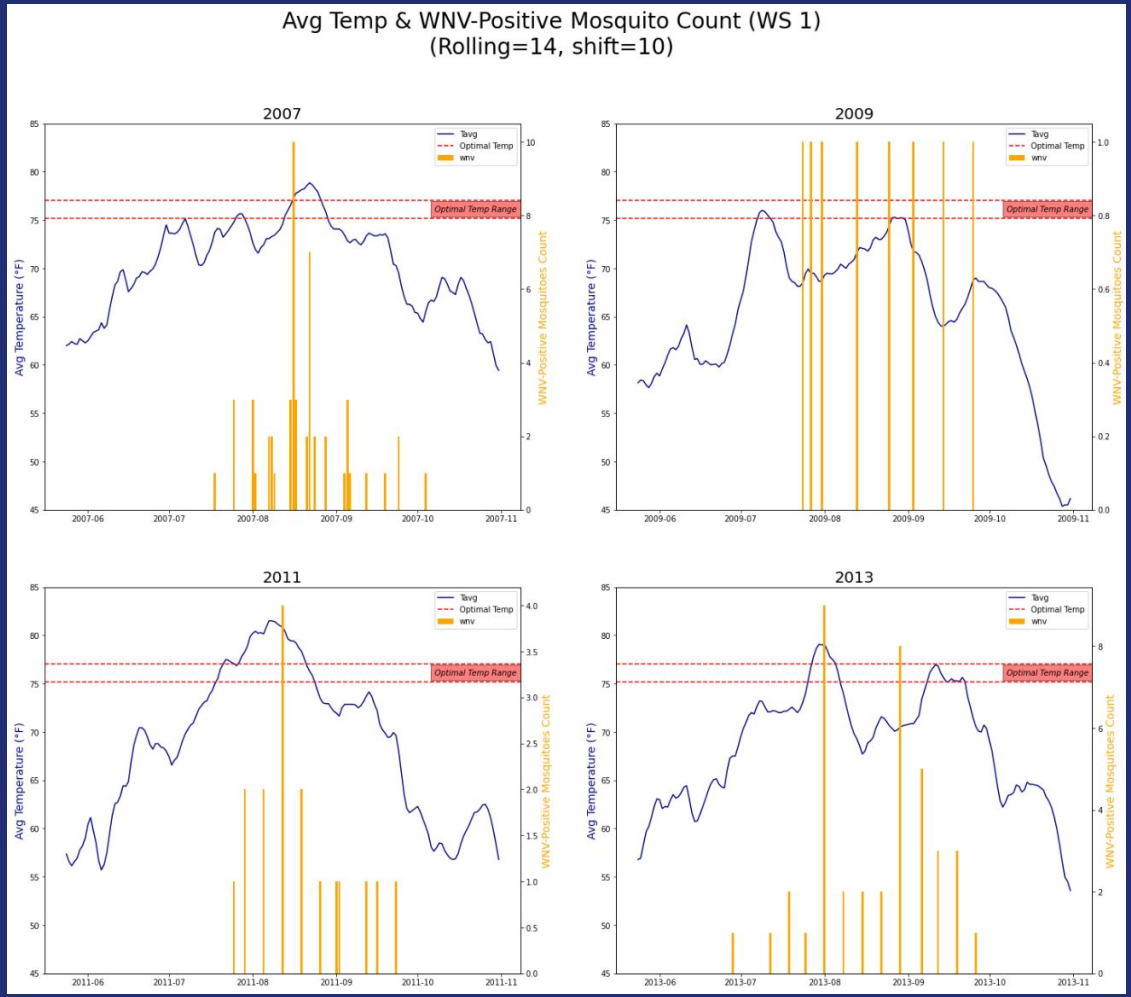


EDA: WNV-Positive by Location



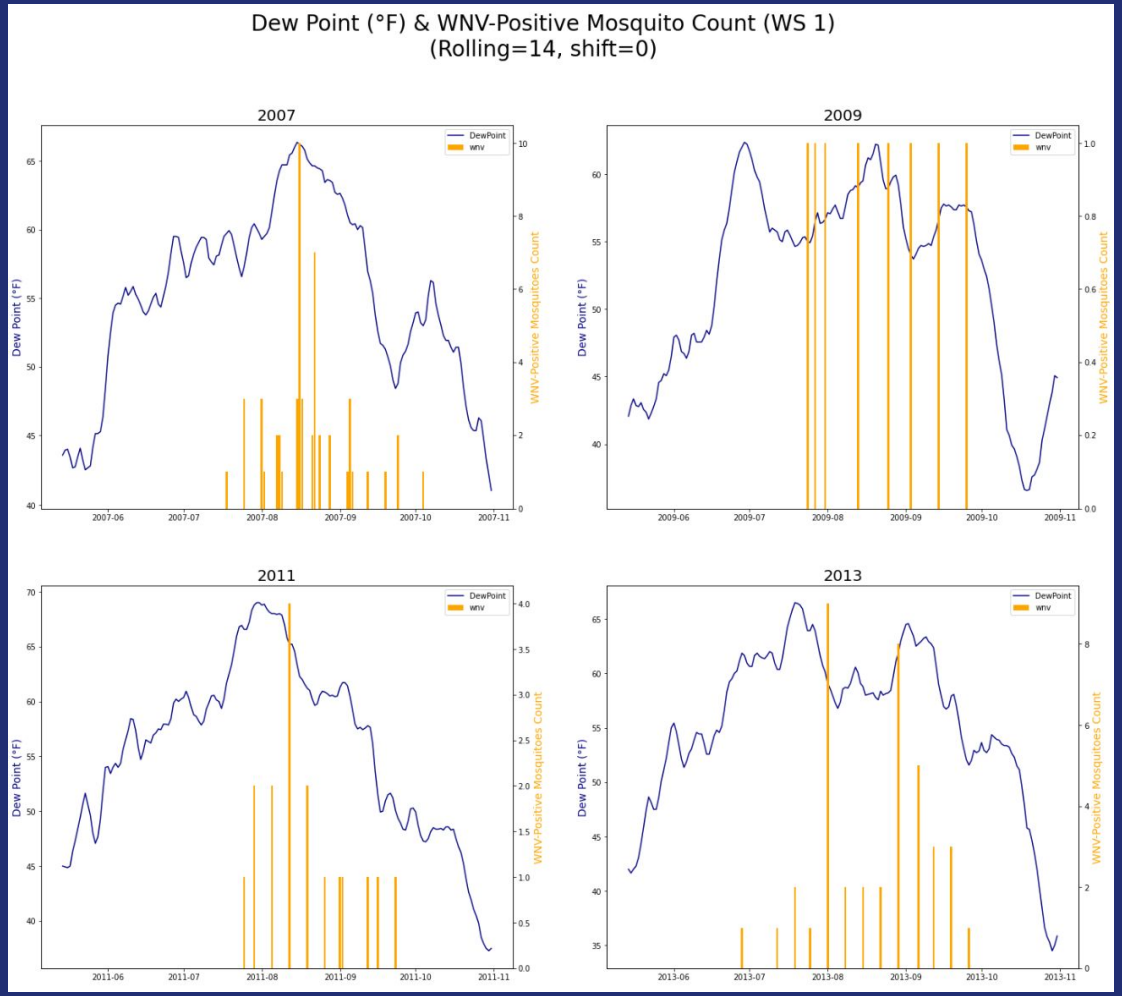
EDA: Temperature

- Immediate & delayed impacts observed
- Impact of Temperature
  - ↑ Reproduction
  - ↑ Biting
  - ↑ Virus replication
- Optimal temperature: 75.2 - 77°F



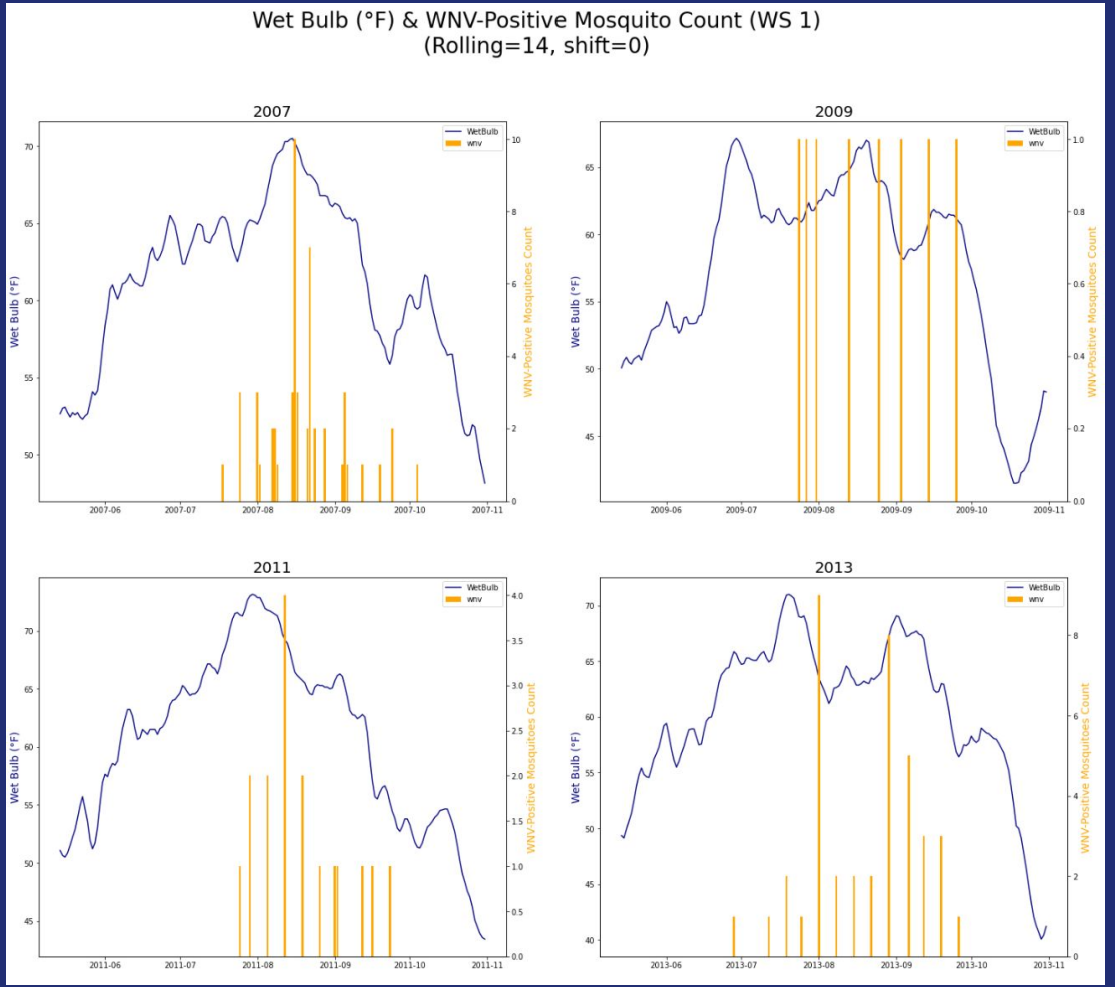
EDA: Dew Point

- Peak between July and September
- High dew points linked to higher WNV counts



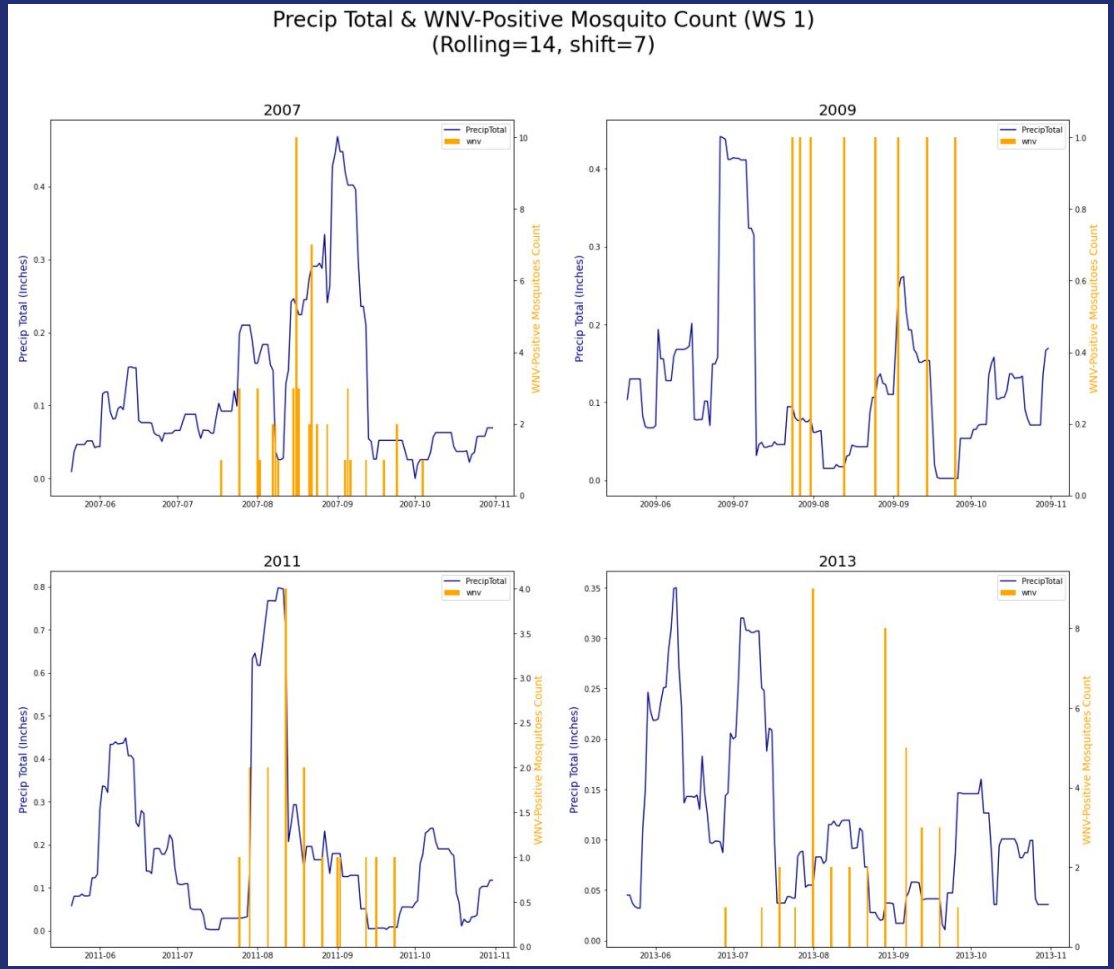
EDA: Wet Bulb

- Peak between July and September
- Wet bulb - humidity relationship
- High humidity offsets higher temperatures



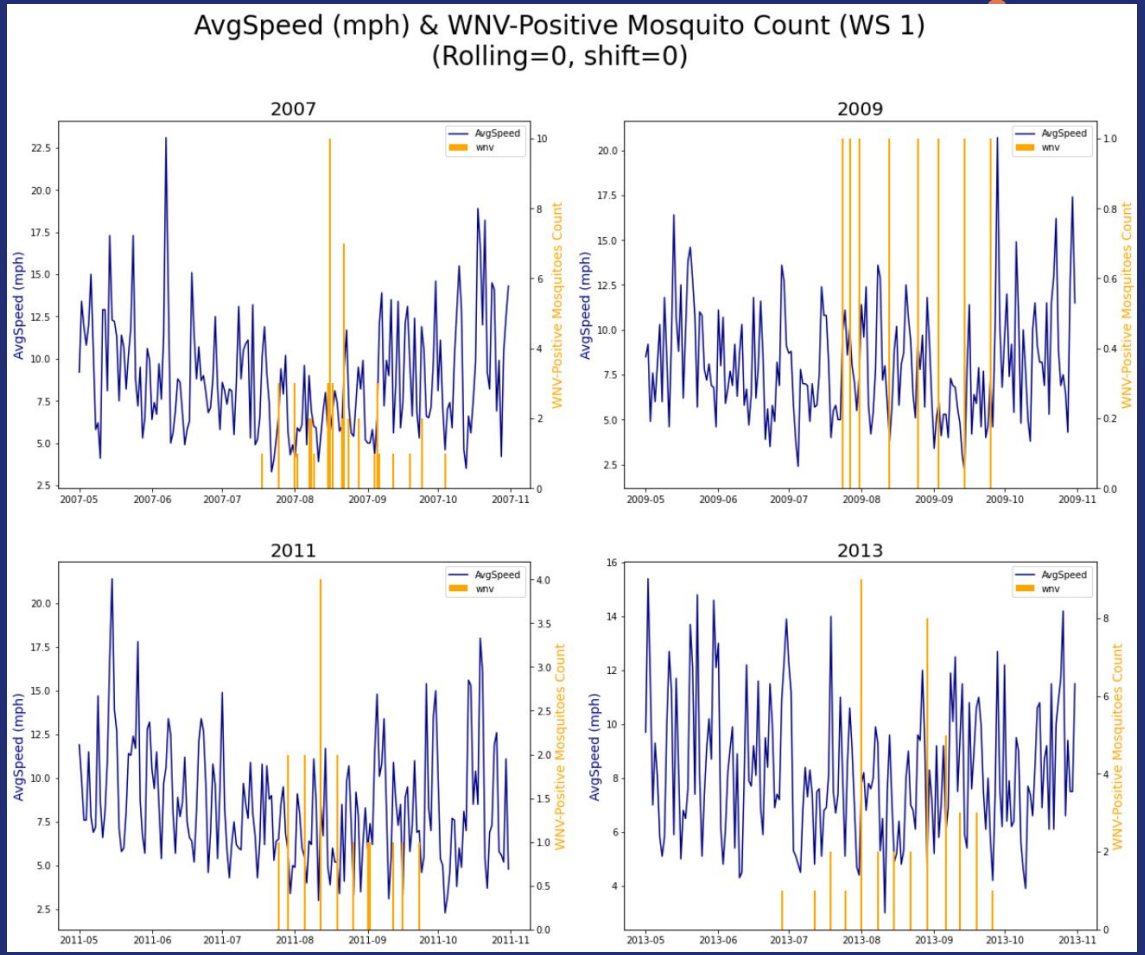
EDA: Precipitation

- Precipitation encourages breeding
- Spike in WNV-positive counts after heavy precipitation (2009, 2011, 2013)



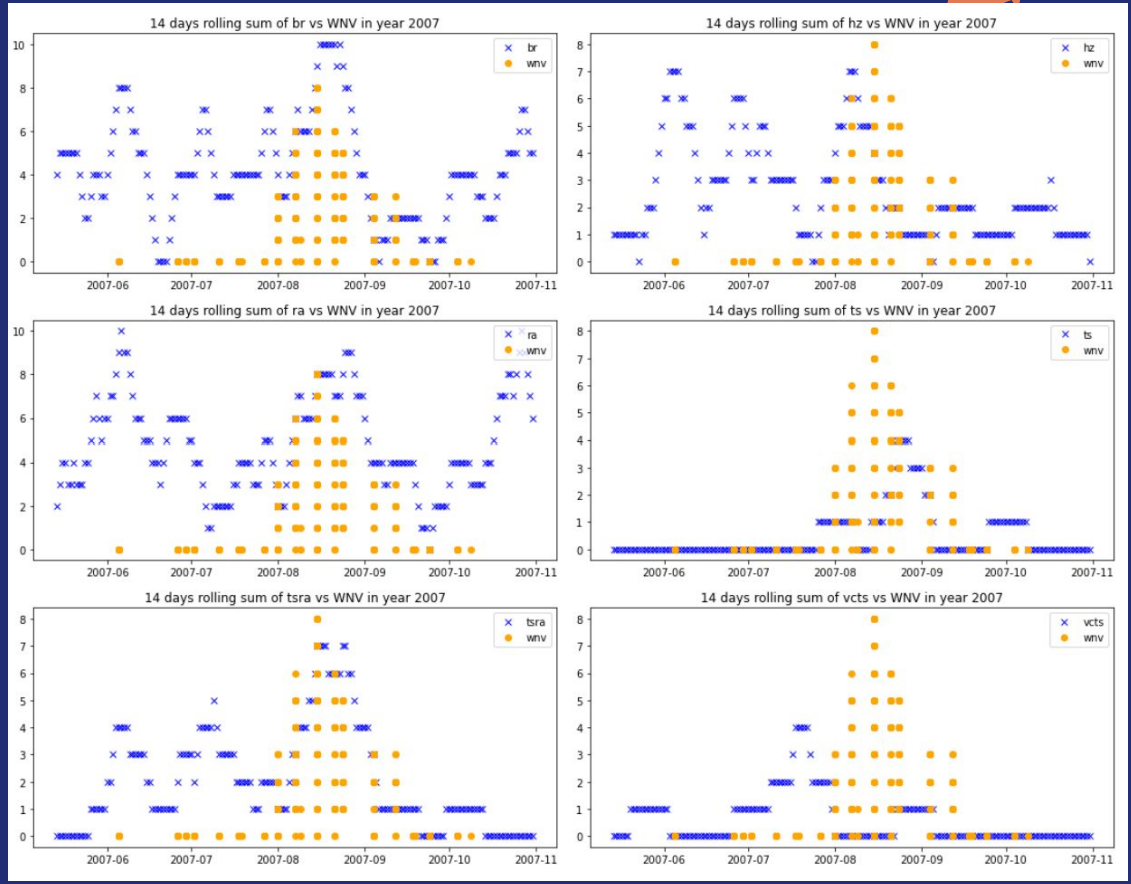
EDA: Wind

- Lower wind speeds show higher WNV numbers
- Mosquitoes caught by traps during lower wind speeds



EDA: Weather Conditions

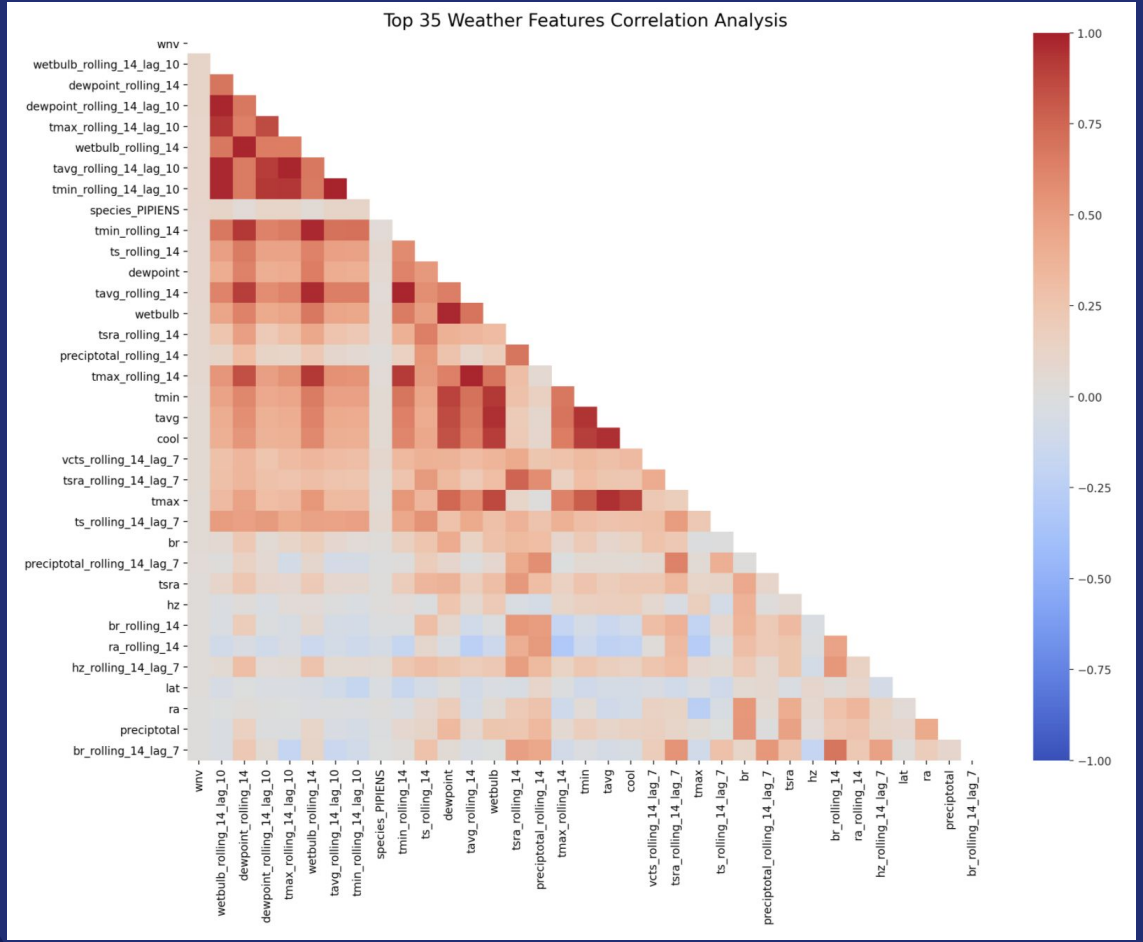
- Weather conditions with highest correlation to WNV:
  - Thunderstorm
  - Thunderstorm / rain
  - Vicinity thunderstorm
  - Rain
  - Mist
  - Haze



EDA: Top Features

- Temperature-related features most correlated with one another
- Rolling average and time lags increase correlation between weather features and WNV

wnv	1.000000
wetbulb_rolling_14_lag_10	0.118518
dewpoint_rolling_14	0.118297
dewpoint_rolling_14_lag_10	0.117584
tmax_rolling_14_lag_10	0.107532
wetbulb_rolling_14	0.106450
tavg_rolling_14_lag_10	0.106093
tmin_rolling_14_lag_10	0.101629
species_PIPIENS	0.094056
tmin_rolling_14	0.082480
ts_rolling_14	0.080939



Modelling

Models Tested:

- Logistic Regression
- K-Nearest Neighbors
- Random Forest
- Extra Trees
- Support Vector Machine
- XGBoost

Grid Search with Pipeline:

1) standard scaler

2) resampling with SMOTE

3) classifier

Modelling Results

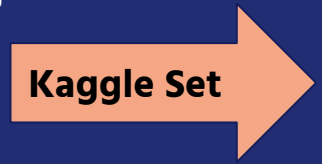
	LR	KNN	RF	ET	SVC	XGB
train_acc	0.818826	0.974271	0.898951	0.910162	0.880890	0.915546
val_acc	0.795181	0.740469	0.823425	0.821004	0.812502	0.823421
test_acc	0.820431	0.775824	0.857993	0.849129	0.855147	0.855532
train_auc	0.741826	0.909938	0.500000	0.504373	0.809036	0.793339
test_auc	0.746211	0.695973	0.504386	0.508772	0.782165	0.736855
train_recall	0.781341	0.944606	0.000000	0.008746	0.854227	0.685131
test_recall	0.780702	0.535088	0.008772	0.017544	0.798246	0.578947

Optimize for ROC AUC and Recall

Kaggle Submission Score

Using SVC with SMOTE, our best parameters are:

'sampling__sampling_strategy': 'auto',
 'svc__C': 0.2,
 'svc__degree': 3,
 'svc__kernel': 'poly'

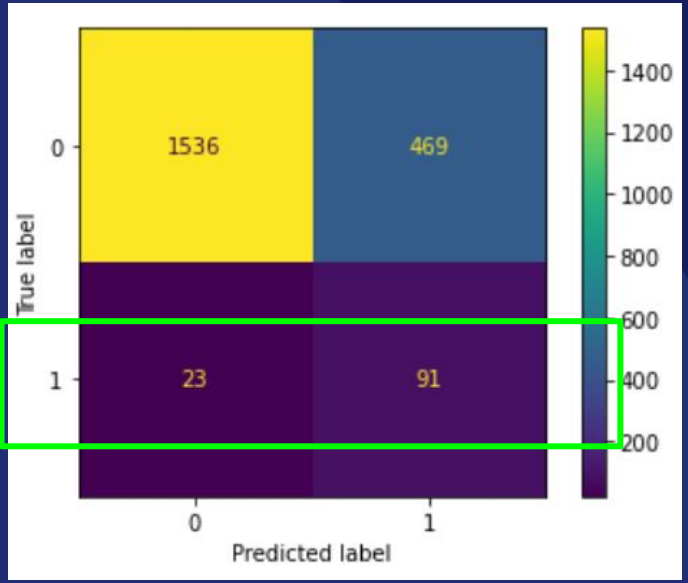


Kaggle Score: 0.683

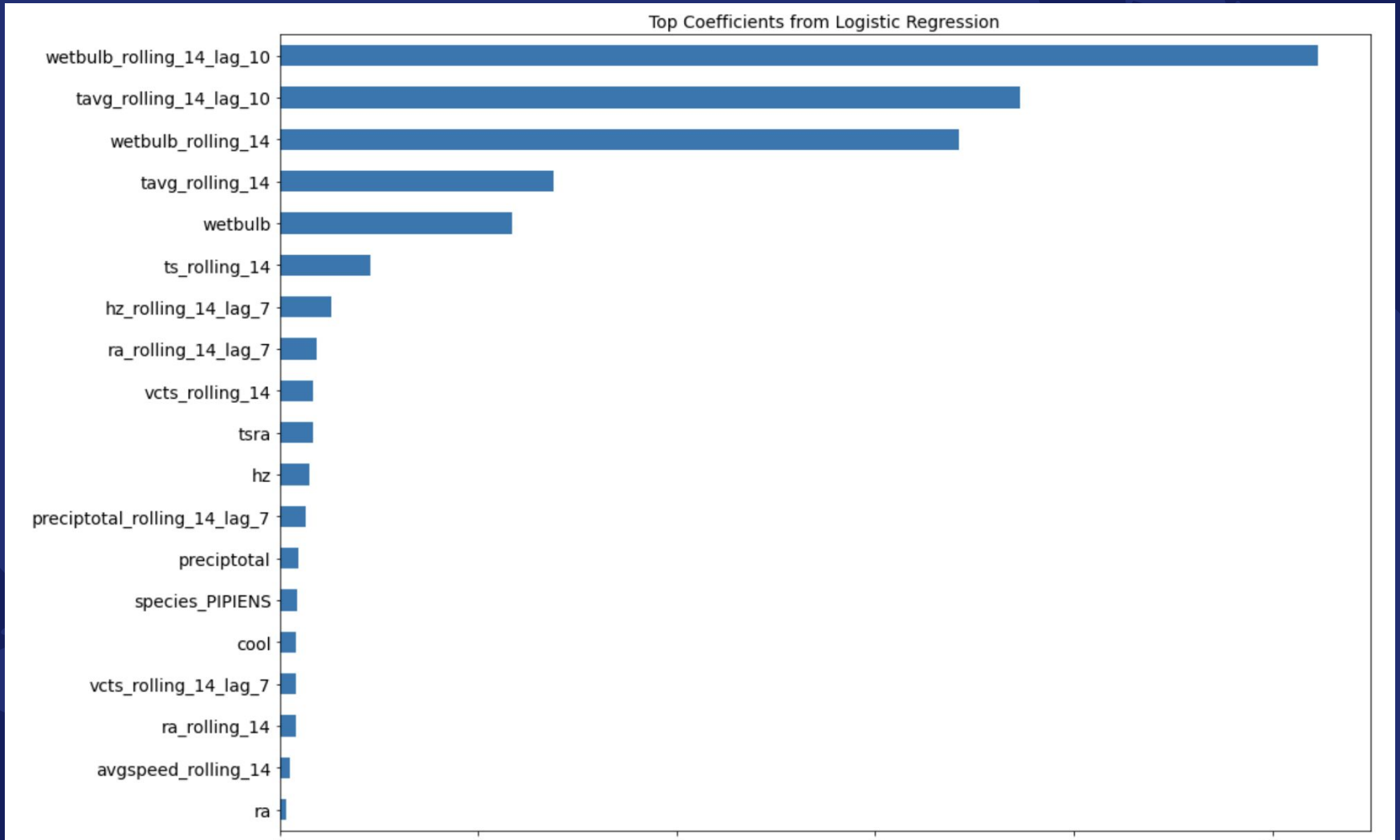
Final Model:
Support Vector Machine

Drawback:

- polynomial kernel does not produce coefficients
- Rely on second best model for examine features



confusion matrix for SVC


Top Coefficients from Logistic Regression
wetbulb_rolling_14_lag_10
tavg_rolling_14_lag_10
wetbulb_rolling_14
tavg_rolling_14
wetbulb
ts_rolling_14
hz_rolling_14_lag_7
ra_rolling_14_lag_7
vcts_rolling_14
tsra
hz
preciptotal_rolling_14_lag_7
preciptotal
species_PIPIENS
cool
vcts_rolling_14_lag_7
ra_rolling_14
avgspeed_rolling_14
ra

Cost Benefit Analysis

Estimated Epidemic Cost:

- Nationwide: $778 million over 15 years
- Louisiana (2005): $20 million
- Sacramento, California (2005): $2.98 million
- Average cost per infected person: $18,000 - $61,000 (depending on severity)

Cost of spraying:

- Vector Control Cost: $701, 790
- 15 prevented WNV cases would justify the cost

But is this enough?

Cost Benefit Analysis

Optimise spraying for weather conditions and months:

- Lower wind speeds (reduces spray drift)
- Temperatures below 86°F (30°C)
- Humidity above 45%
- July to September

Cost Benefit Analysis

Alternative Measures

- Eliminate mosquito breeding grounds
- Insect repellent
- Long-sleeve shirts and long pants

Conclusion / Recommendations

- Pesticide spraying is not enough
- Combination of spraying and alternative measures
- Spray in the right weather conditions
- Get data on number of WNV cases reported for better prediction

THANK YOU