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

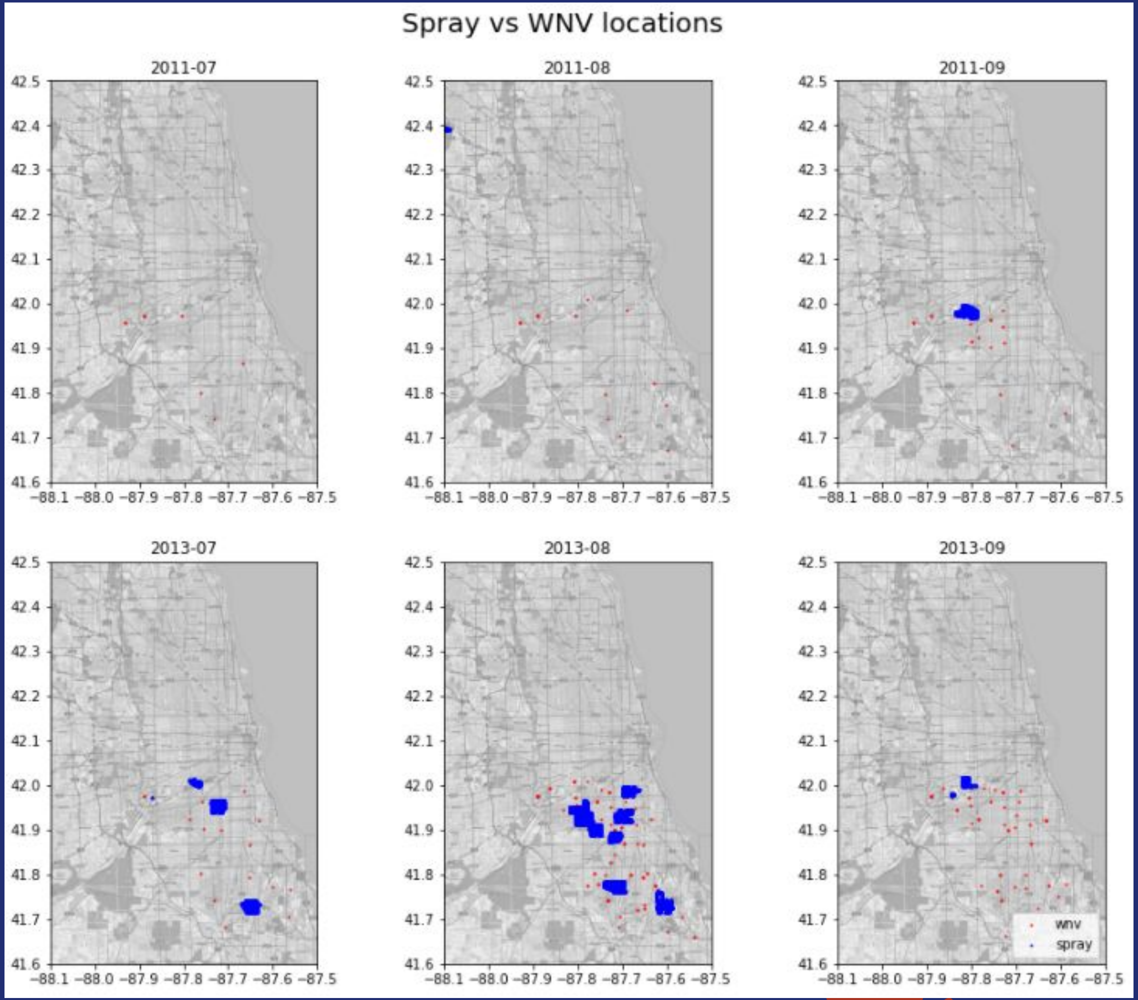
- Train: 10,506 rows, 12 columns (2007, 2009, 2011, 2013)
- Test: 116,293 rows, 11 columns (2008, 2010, 2012, 2014)
- Relevant columns:
  - Date
  - Species
  - Longitude
  - Latitude
  - WNV present
- Set date as index
- Assign nearest weather station to each trap
- Group by mosquito species
- Convert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

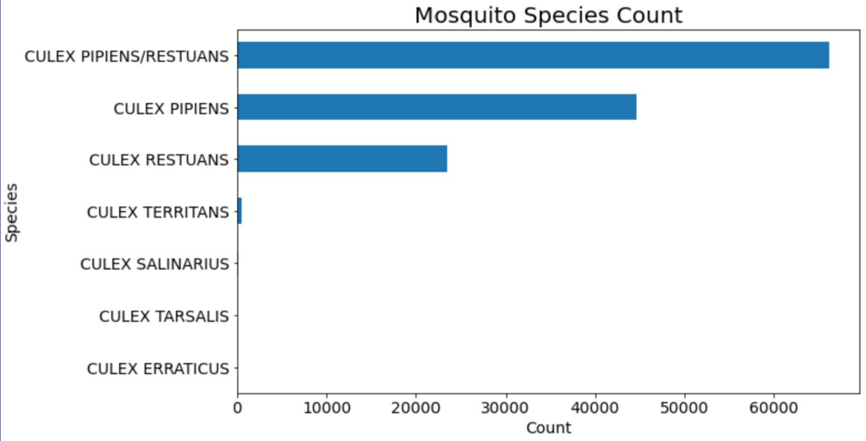
- 2011 and 2013 train data locations were checked if they had been sprayed within a certain radius within the past 10 days
- Spraying within 10m, 30m and 50m of a location within the last 10 days has **marginal effect** on the number of mosquitoes caught or the presence of WNV
- Since spray data for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

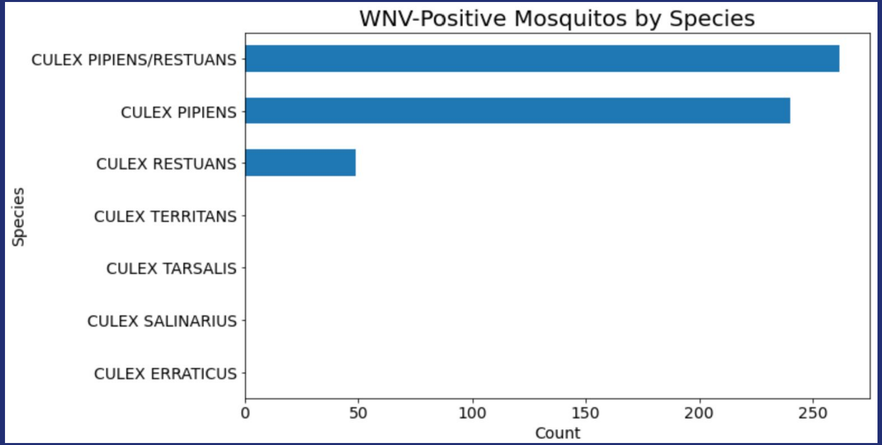
Sprayed radius within last 10 days vs WNV

	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

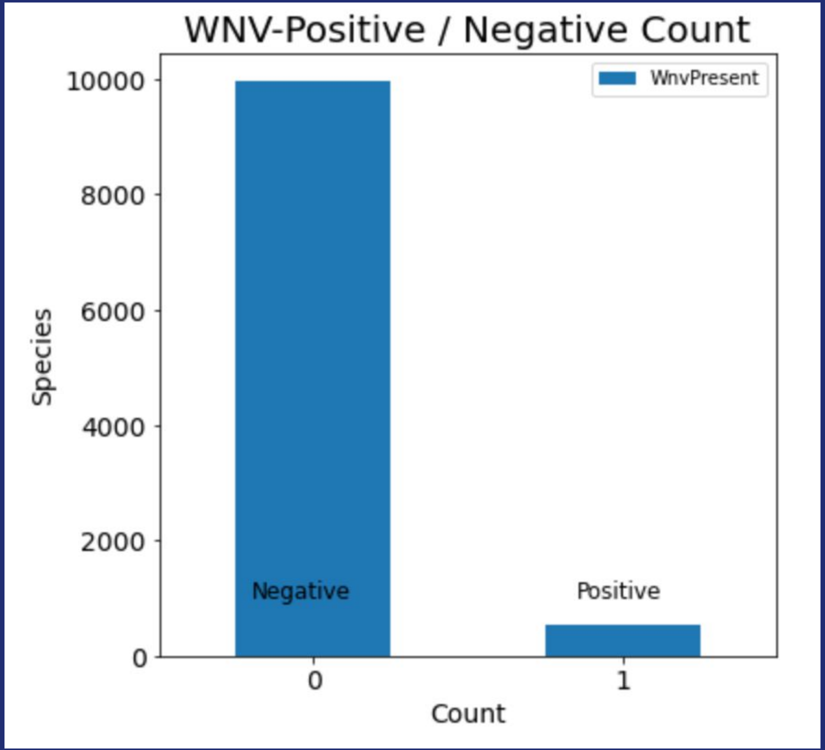
- Of the 7 species of mosquitoes caught, only 3 were found with the WNV. These were also the most frequently caught species
- As the distributions of total mosquito counts and WNV-positive mosquito counts differ, we should expect species to be an important feature in predicting WNV





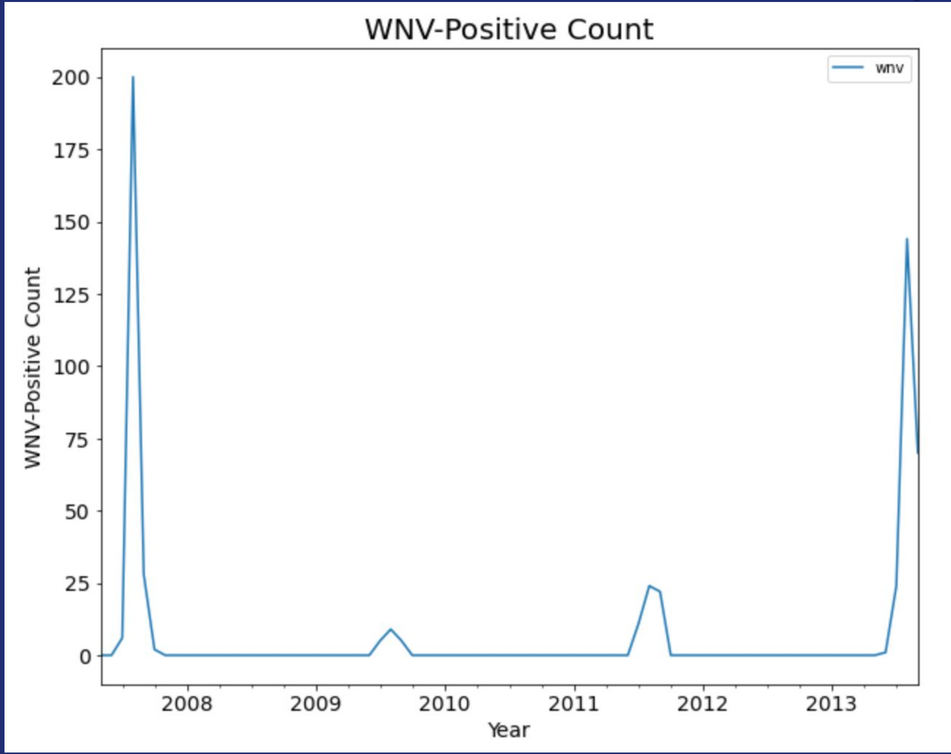
EDA: WNV Positive/ Negative Counts

- Grouped by species, there are **9,955** WNV-negative vs **551** WNV-positive findings in the train data
- As the classes are imbalanced, we resampled the WNV-positive data using SMOTE

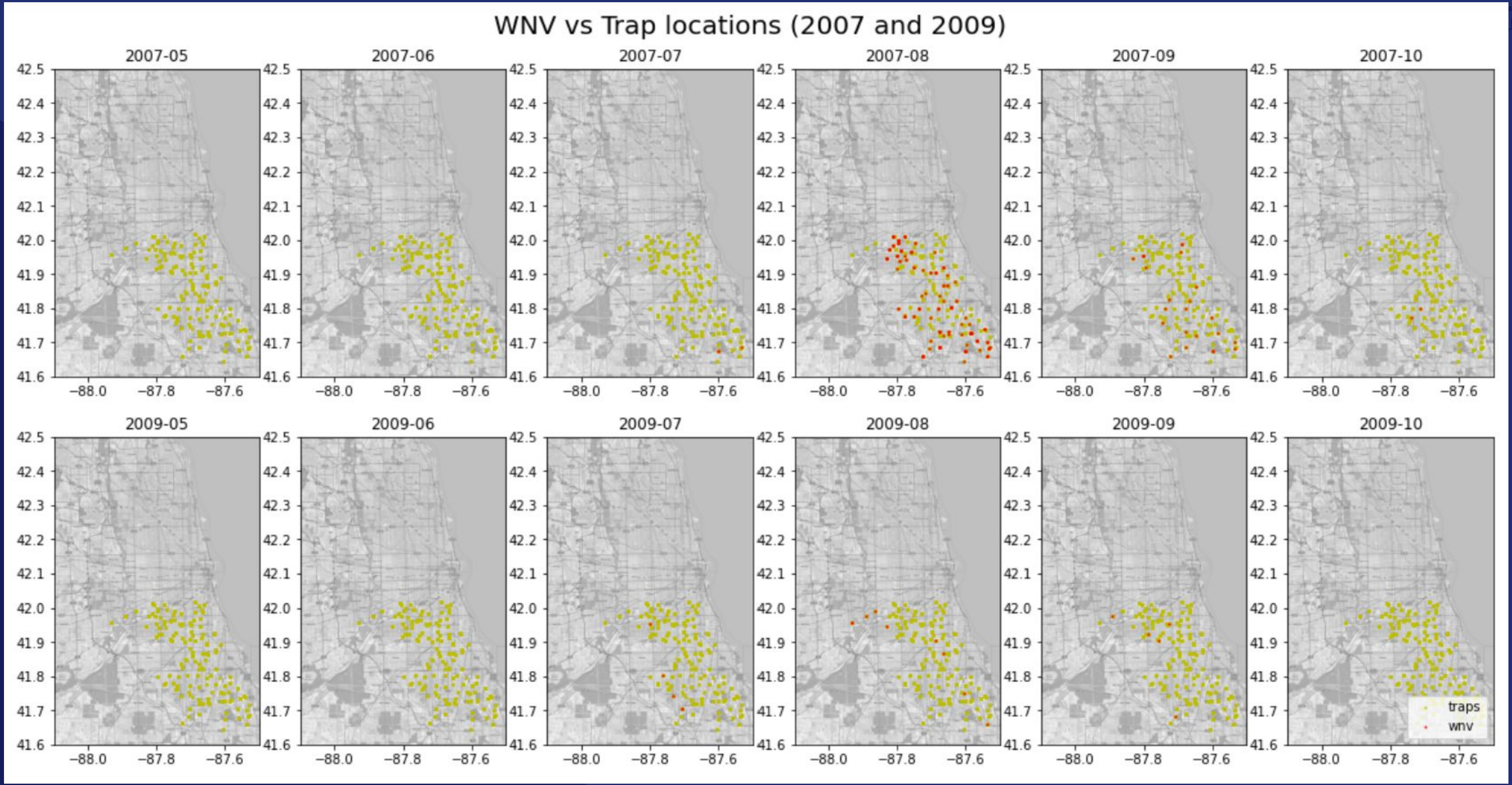


EDA: WNV-Positive by Date

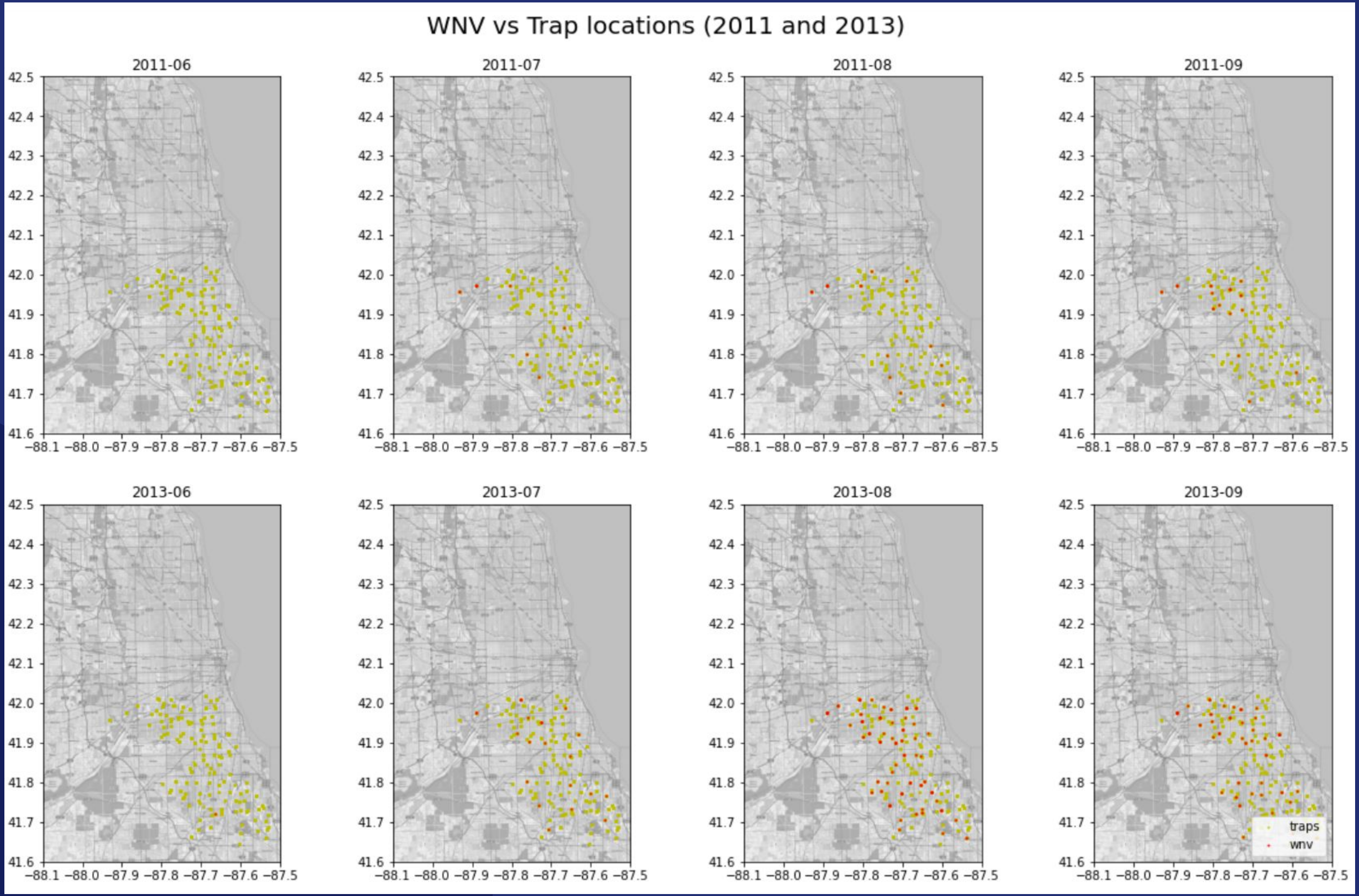
- WNV-positive mosquitoes were most frequently caught in August and September each year
- 2007 and 2013 had the most number of WNV-positive mosquitoes caught



EDA: WNV-Positive by Location

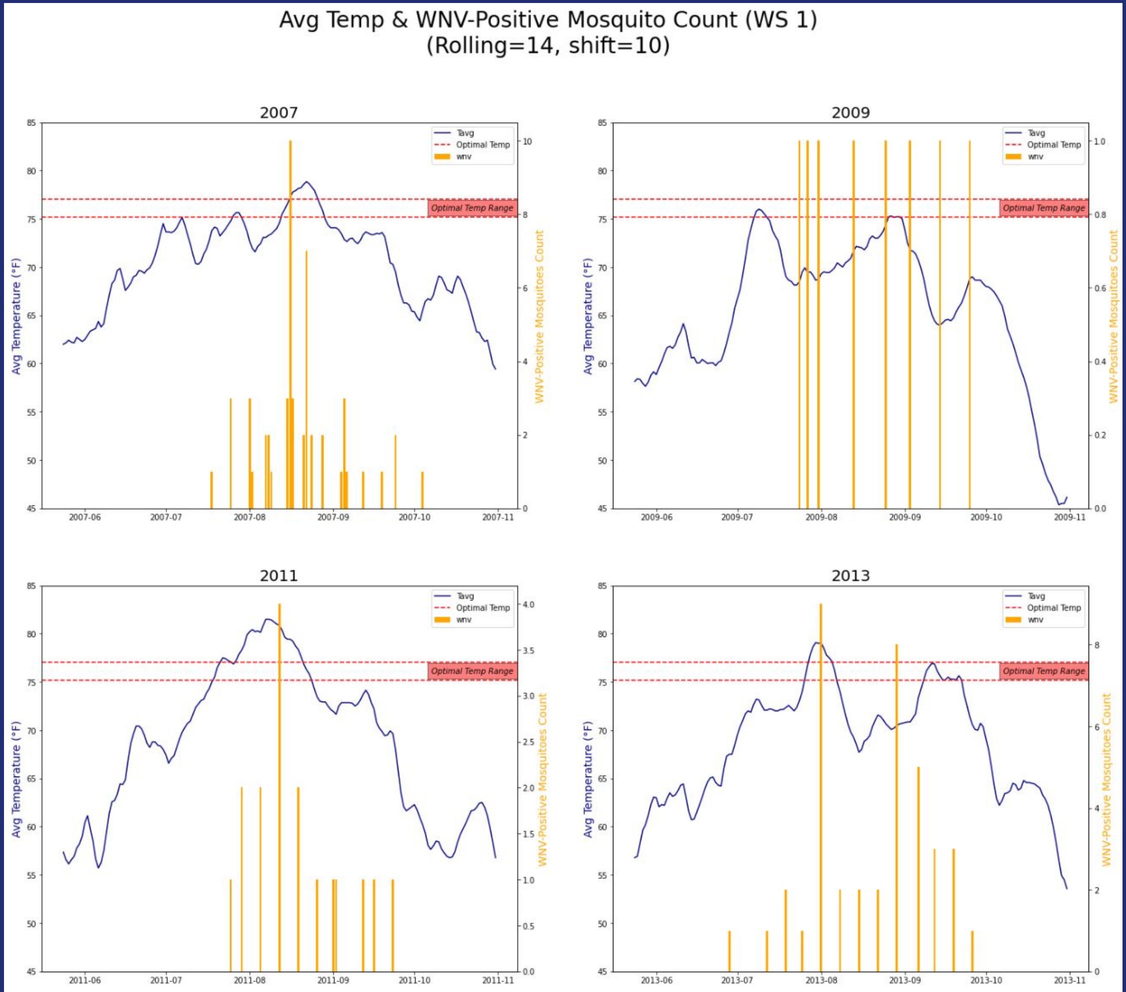


EDA: WNV-Positive by Location



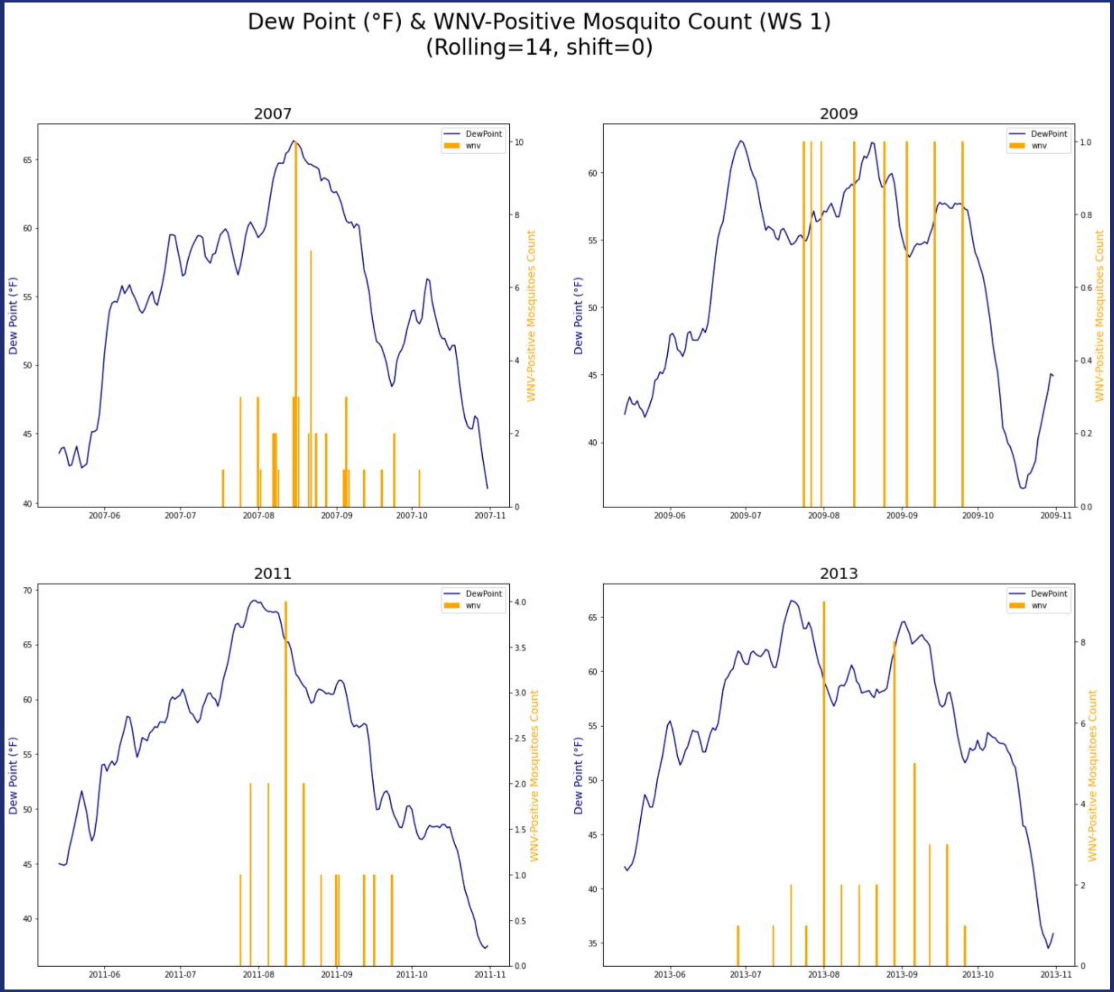
EDA: Temperature

- Temperature has an impact on these 3 factors:
  - Mosquito reproduction rate
  - Mosquito biting rate
  - Virus replication rate
- Optimal temperature for WNV to spread is between 75.2 to 77 degrees Fahrenheit
- Higher temperatures have both an immediate and delayed impact on the WNV-positive mosquito count



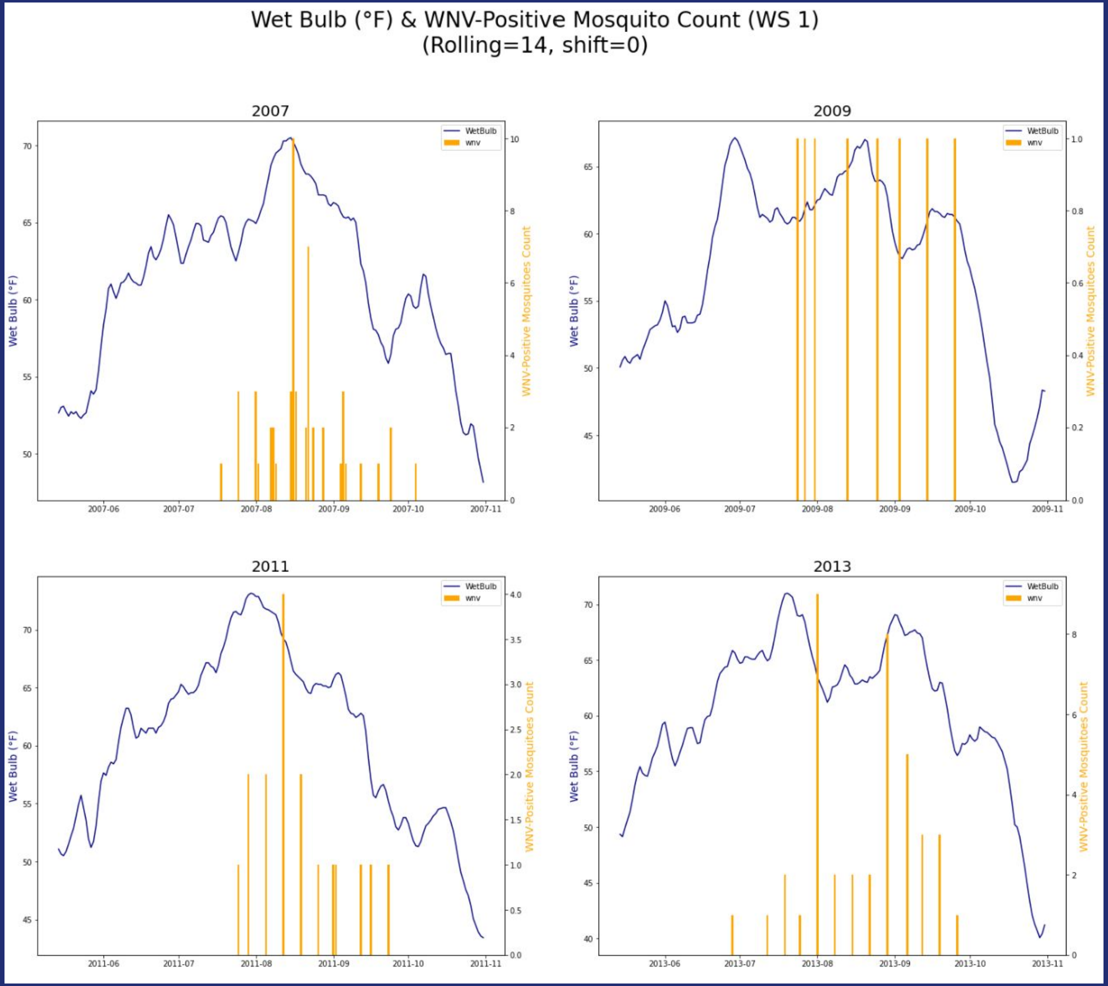
EDA: Dew Point

- Dew point tends to peak between July and September
- Generally, higher dew points do result in higher WNV counts



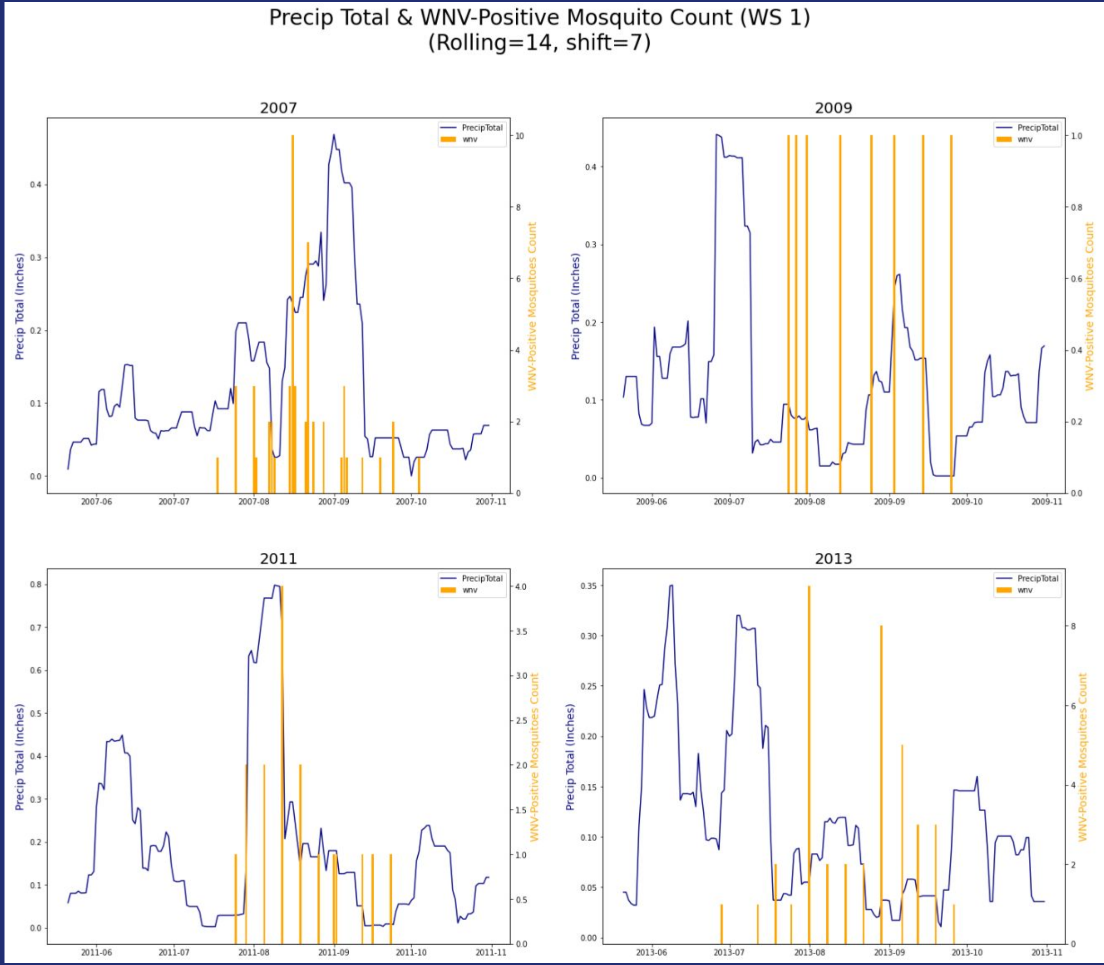
EDA: Wet Bulb

- Wet bulb tends to peak between July and September
- Generally, higher wet bulb temperatures are associated with higher levels of humidity, which offsets the higher temperatures, and thus results in higher WNV counts



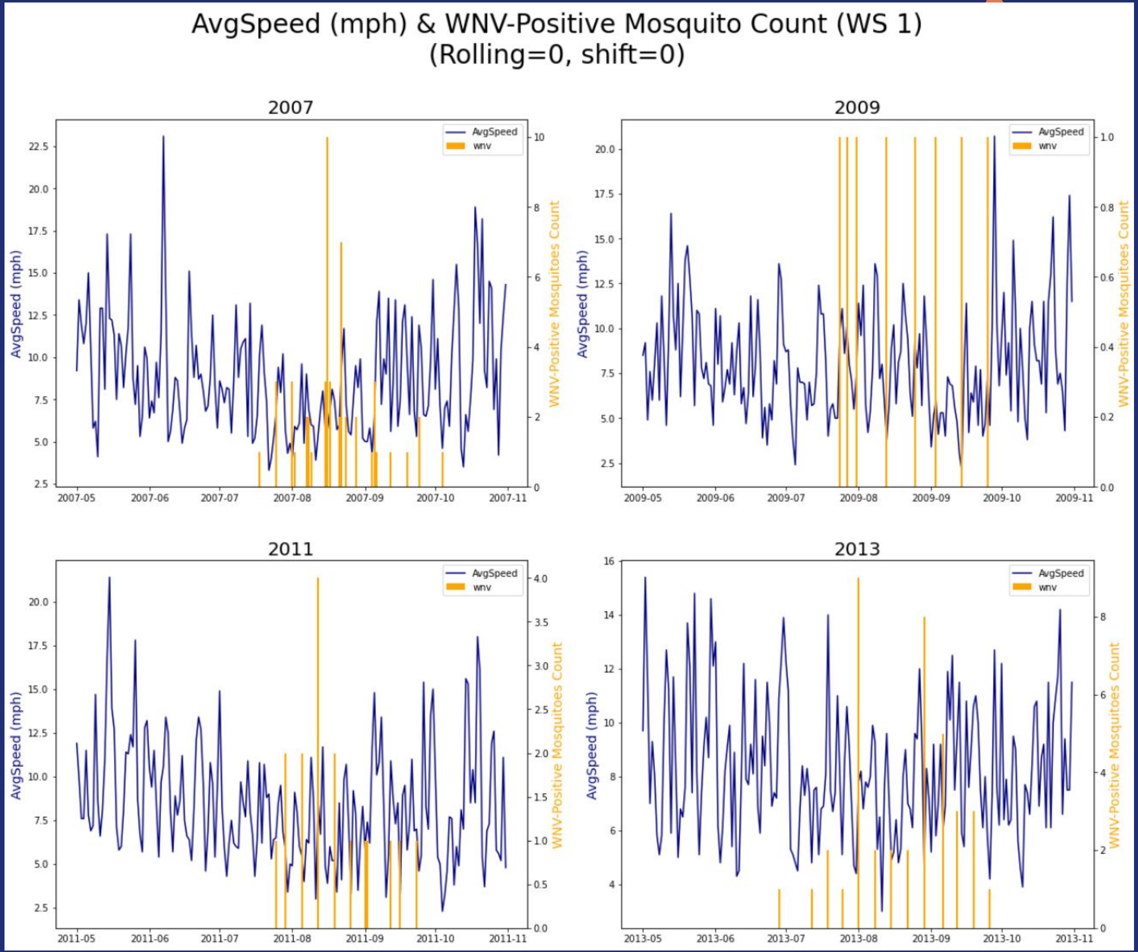
EDA: Precipitation

- Higher precipitation increases the amount of water surfaces for mosquitoes to breed
- This explains the spikes in WNV-positive counts after periods of heavy precipitation as shown in the years 2009, 2011 and 2013



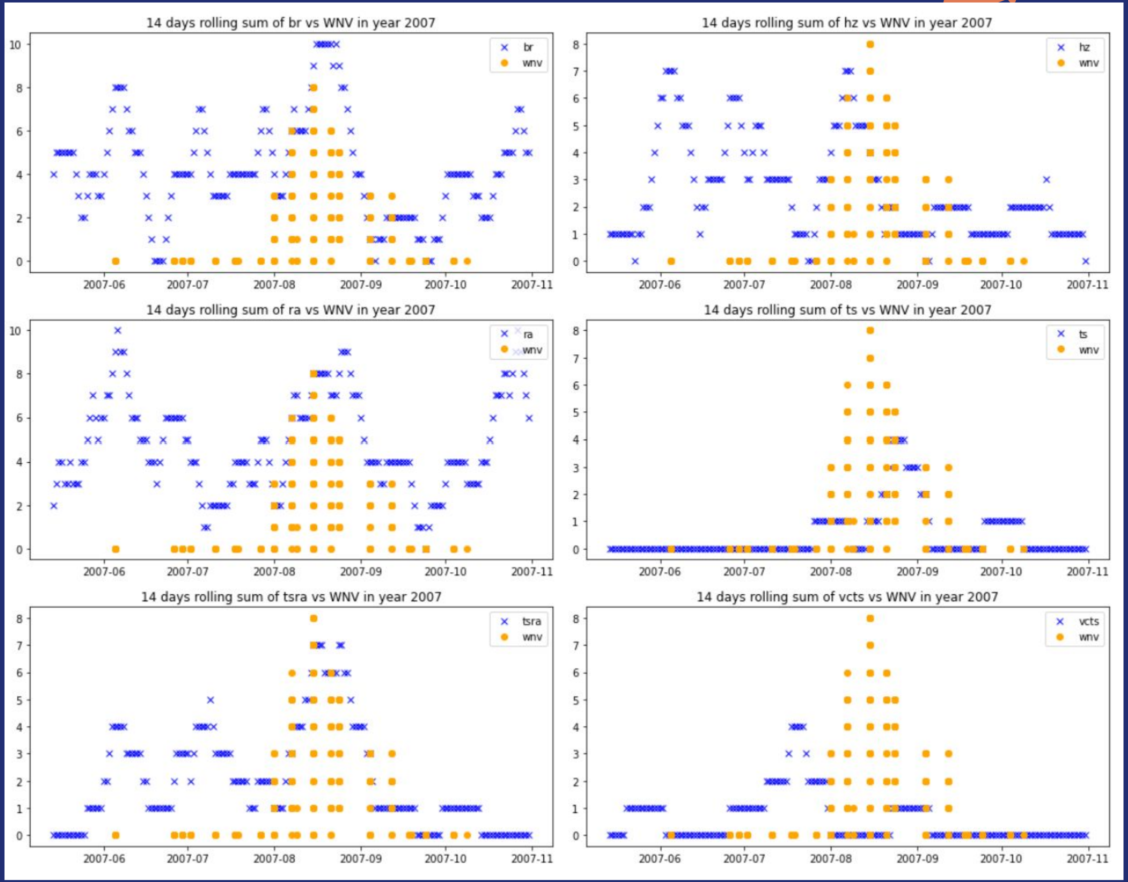
EDA: Wind

- Lower average wind speeds show a larger number of WNV-positive mosquitoes being detected
- It is likely that fewer mosquitoes are detected when wind speeds are higher, and they are only detected by traps during lower wind speeds



EDA: Weather Conditions

- Weather conditions associated with rain (thunderstorm, thunderstorm / rain, vicinity thunderstorm, rain, mist, haze) were found to be most highly correlated to the presence of WNV

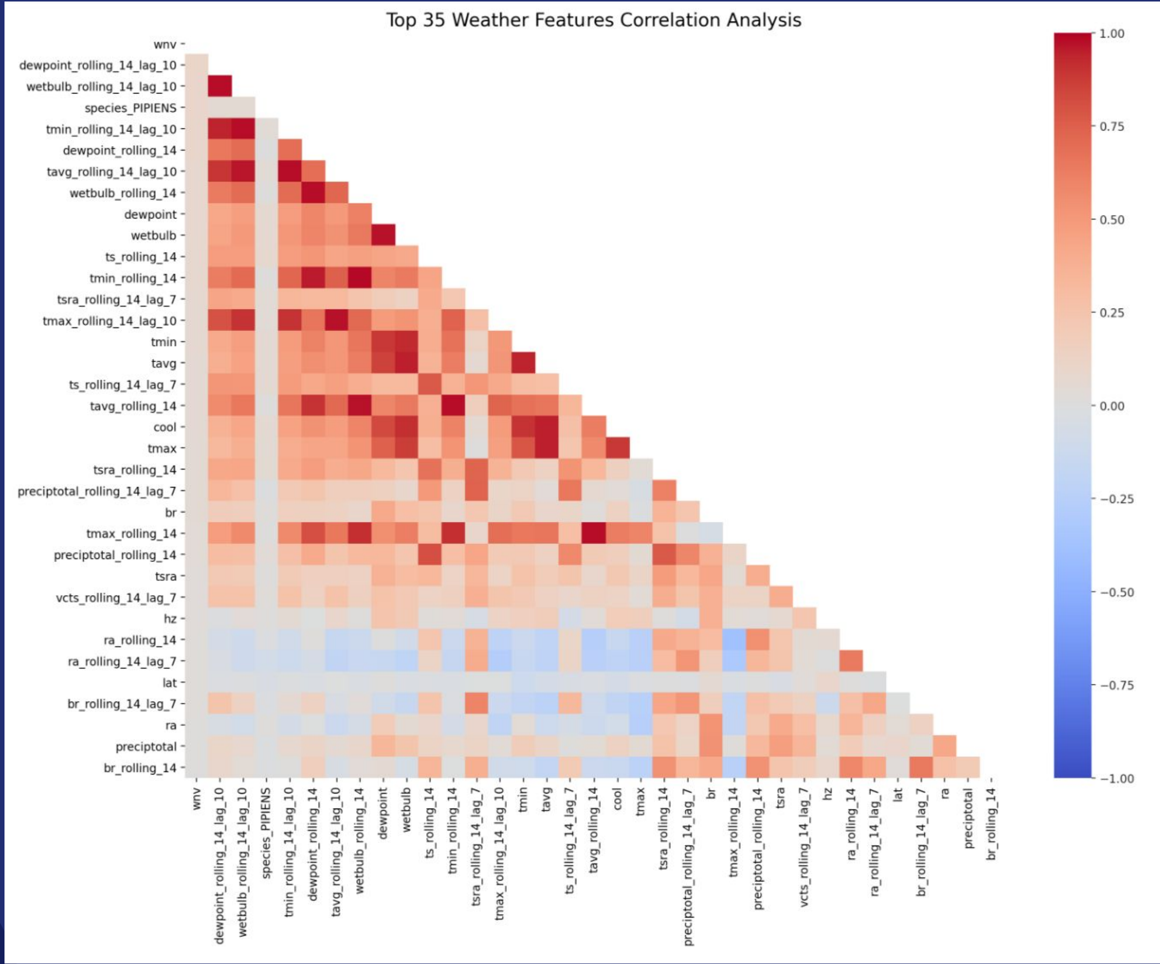


EDA: Top Features

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

Top 10 correlation with WNV

wnv	1.000000
dewpoint_rolling_14_lag_10	0.103517
wetbulb_rolling_14_lag_10	0.100037
species_PIPIENS	0.094056
tmin_rolling_14_lag_10	0.088623
dewpoint_rolling_14	0.088508
tavg_rolling_14_lag_10	0.081336
wetbulb_rolling_14	0.080345
dewpoint	0.079021
wetbulb	0.076981
ts_rolling_14	0.075088



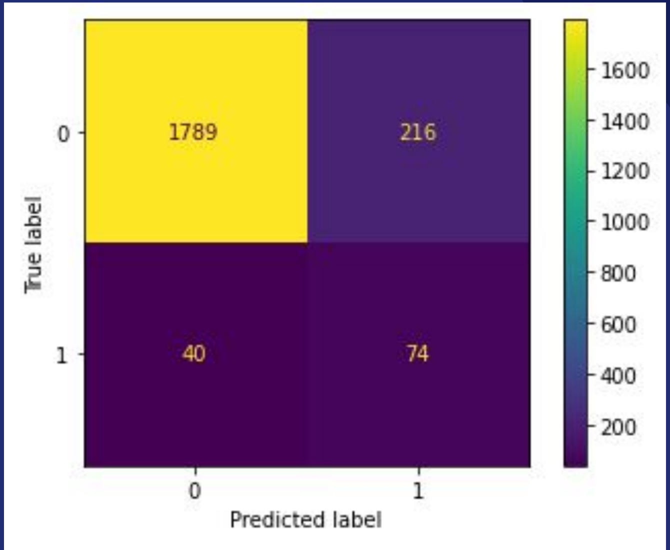
Modelling

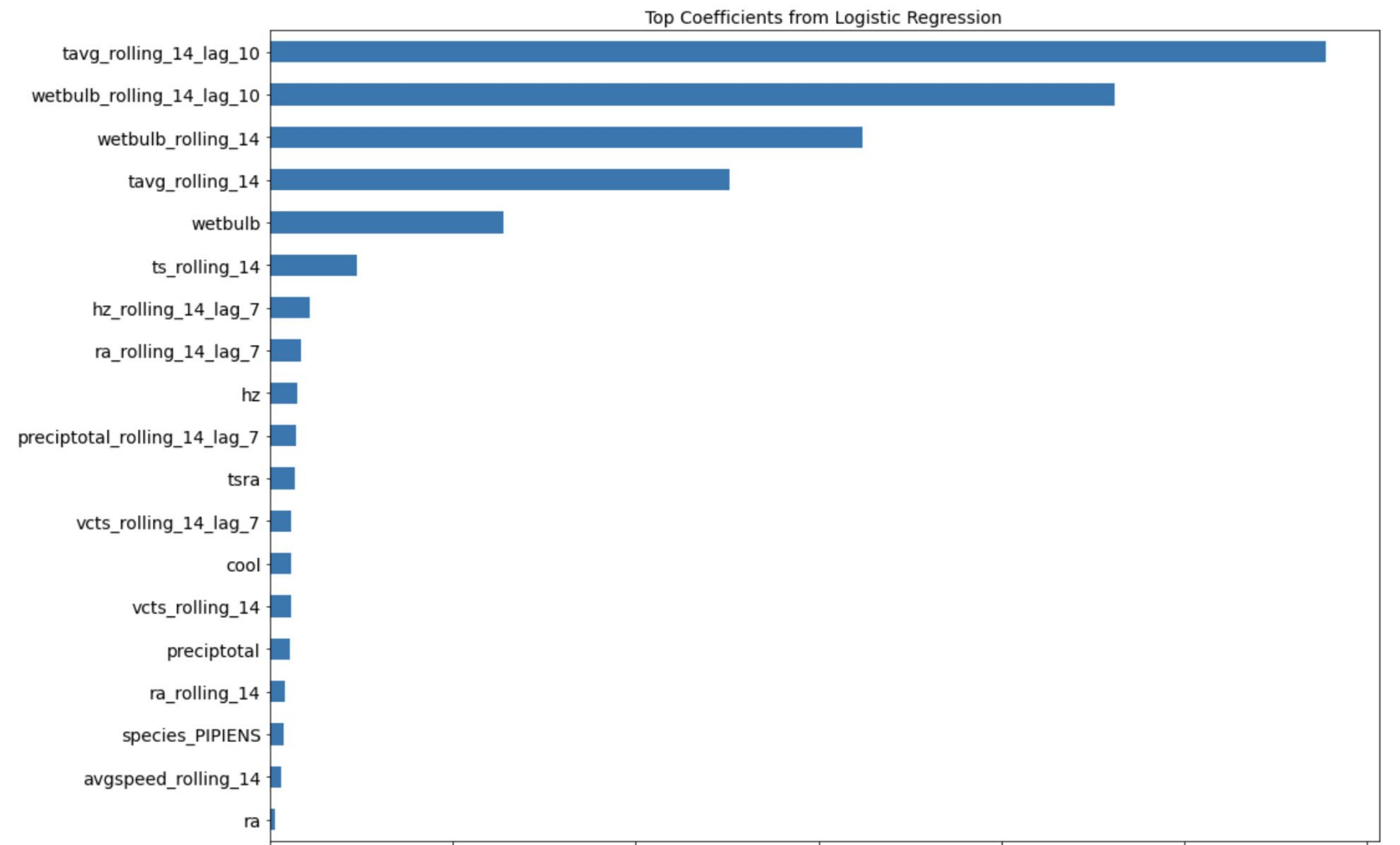
Models Tested:

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

Modelling Results

	LR	KNN	RF	ET	SVC	XGB
train_acc	0.815192	0.975321	0.901253	0.909615	0.878503	0.918950
val_acc	0.786722	0.742261	0.823720	0.820316	0.819098	0.820091
test_acc	0.825388	0.775244	0.854261	0.848064	0.853835	0.863735
train_auc	0.746376	0.916562	0.500000	0.508746	0.808571	0.796382
test_auc	0.756978	0.697616	0.504386	0.508772	0.787651	0.770696
train_recall	0.801749	0.959184	0.000000	0.017493	0.845481	0.693878
test_recall	0.798246	0.543860	0.008772	0.017544	0.798246	0.649123




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

Kaggle Submission Score

Using XGBoost with smote, our best parameters are:

```
{'sampling__k_neighbors': 5,
 'xgb__alpha': 0.1,
 'xgb__colsample_bytree': 0.9,
 'xgb__gamma': 0.4,
 'xgb__gpu_hist': 'gpu_hist',
 'xgb__lambda': 0.2,
 'xgb__learning_rate': 0.1,
 'xgb__max_depth': 5,
 'xgb__min_child_weight': 2,
 'xgb__n_estimators': 100,
 'xgb__n_jobs': -1,
 'xgb__subsample': 0.8,
 'xgb__verbosity': 2}
```

Kaggle Set → Kaggle Score: 0.636

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

THANK YOU