West Nile Virus Prediction

Introduction:

- West Nile Virus:- disease transmitted through mosquito bite
- Develop severe symptoms leading death
- First reported in 2002 in Chicago

Problem:

- City of Chicago and Chicago Department of public health starts surveillance program to control the mosquitoes
- Identify the potential outbreak region and spray disinfectants to control the mosquitoes

Objective:

- ◉ Develop machine learning algorithm to predict when and where the mosquito will test positive for the West Nile virus

Methodology

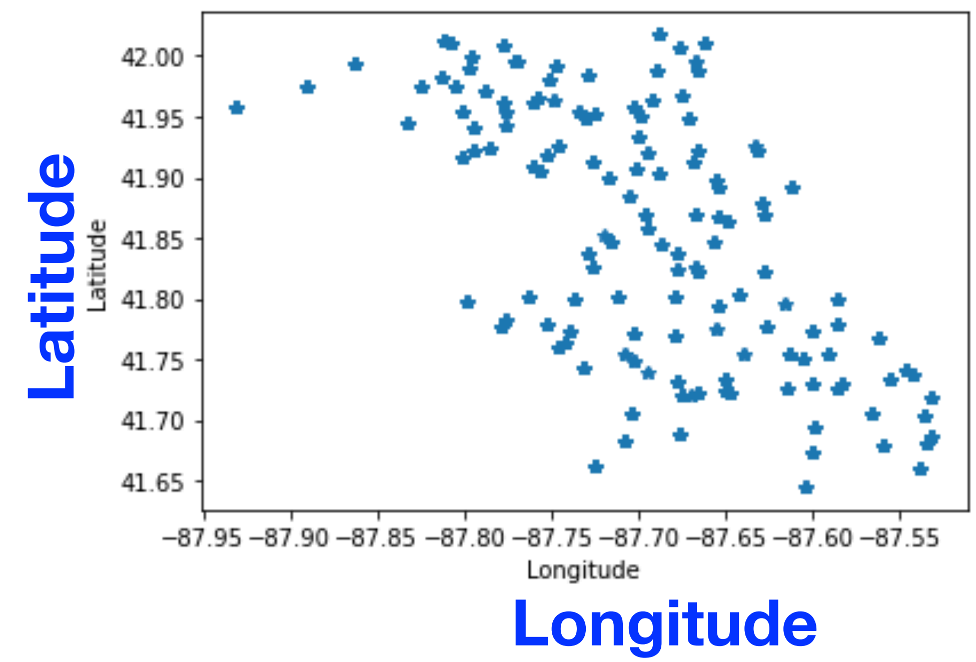
Data analysis:

- Two sets of data: GIS data and weather data
- GIS data:
    - provides location information, date, mosquito species and a label indicating presence or absence of virus features describing ticket prices
    - 10506 rows and 11 features
    - No missing data
- Weather data:
    - Weather parameters from two weather stations describing temperatures, pressure, precipitation, sunrise, sunset, etc
    - 2944 rows with 21 features
    - Missing parameters in rows and columns

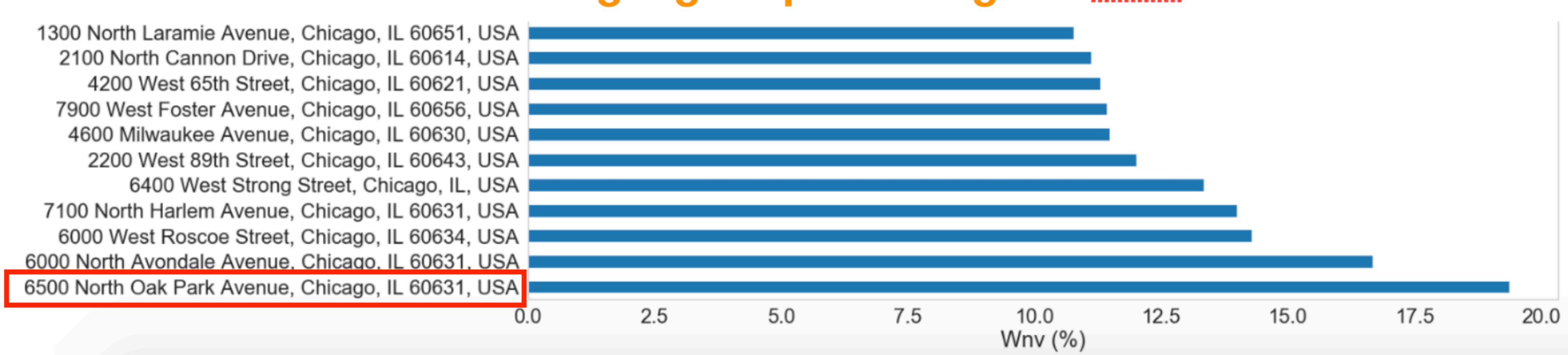
GIS data

- Samples collected from 136 locations every two years apart from 2007 to 2013
- Latitude and longitude of locations are provided along with street name, address and block number
- Mosquitoes are trapped from May to October
- Label 0 and 1 indicates presence and absence of virus
- The highest sample collected from O'hare international airport with 8.8% containing the virus
- The highest virus percentage as 19.35% was observed at 6500 North Park Avenue, Chicago IL, 60631, USA

City of Chicago



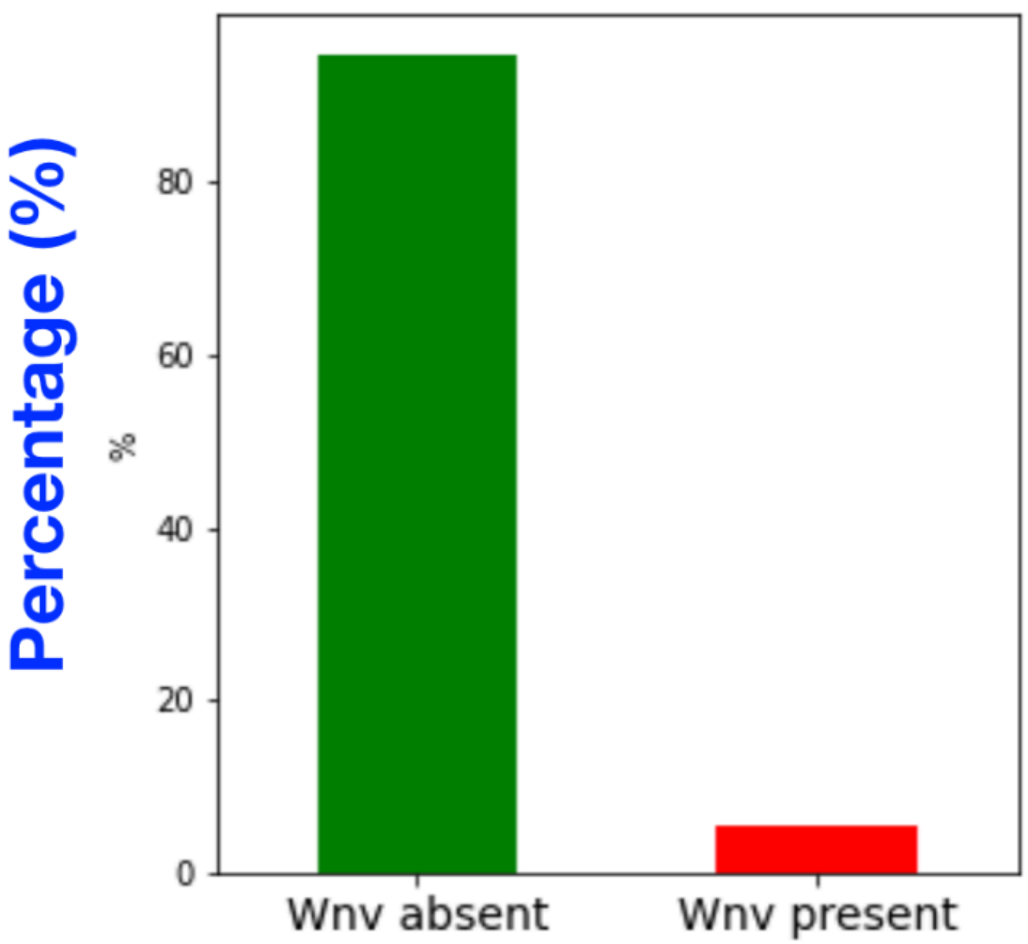
Address observing higher percentage of Wnv virus



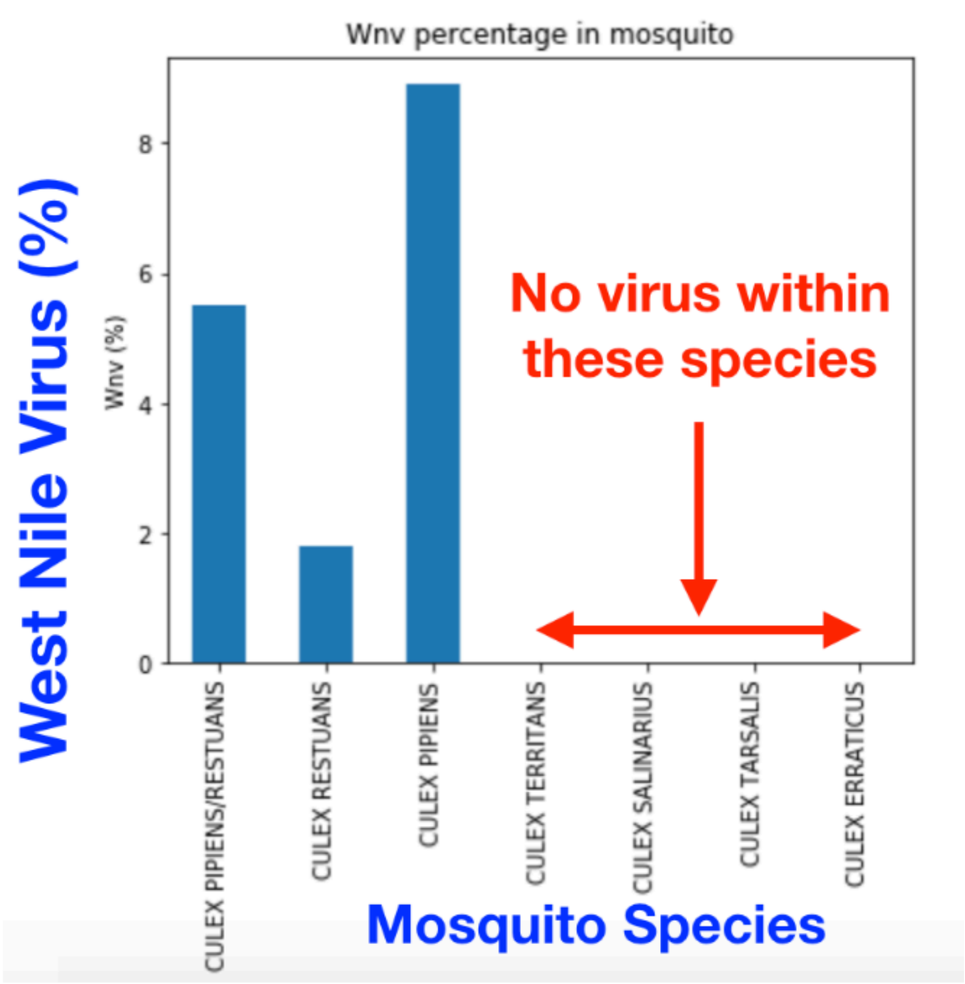
Exploratory data analysis

- Six species of mosquitoes with two frequently appeared together in most locations
- Virus present only in two species of mosquitoes:- Culex Restuans and Culex Pipiens
- Highly imbalanced data with 5.2% probability of observing virus

Probability



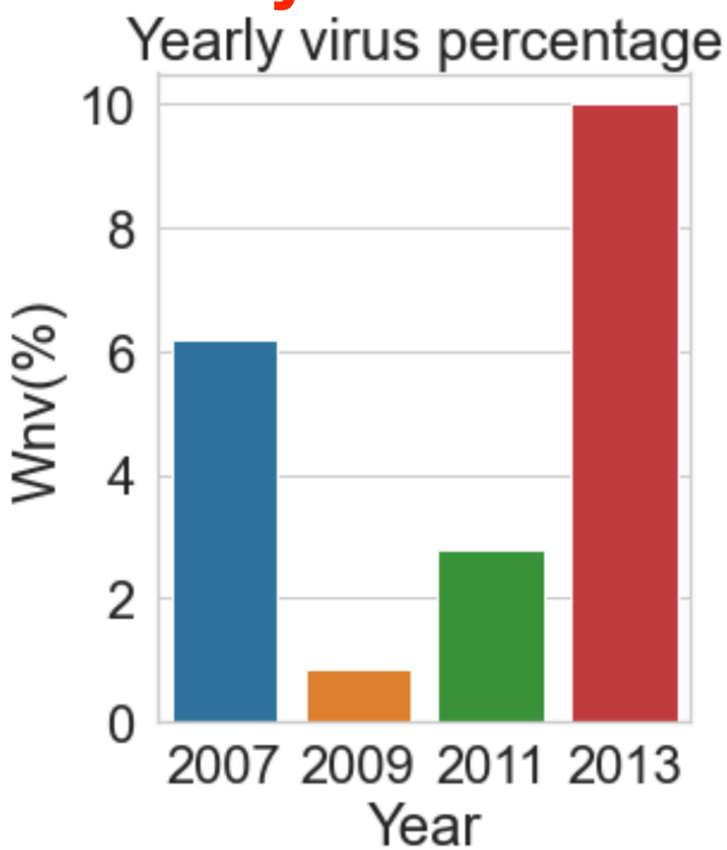
Mosquito species



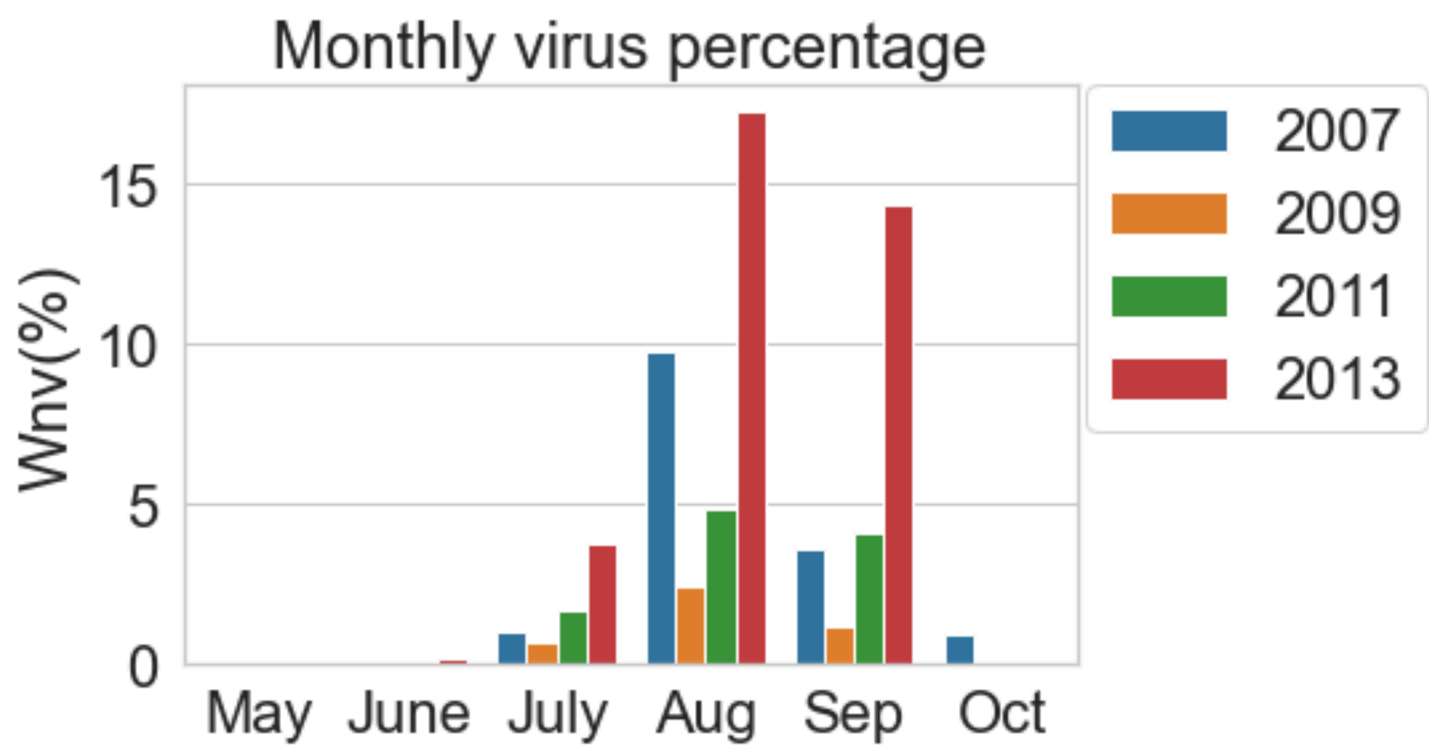
Exploratory data analysis

- Virus observation probability is higher in 2013 and fall season
- Virus appears in June, becomes critical in August and then disappears in October

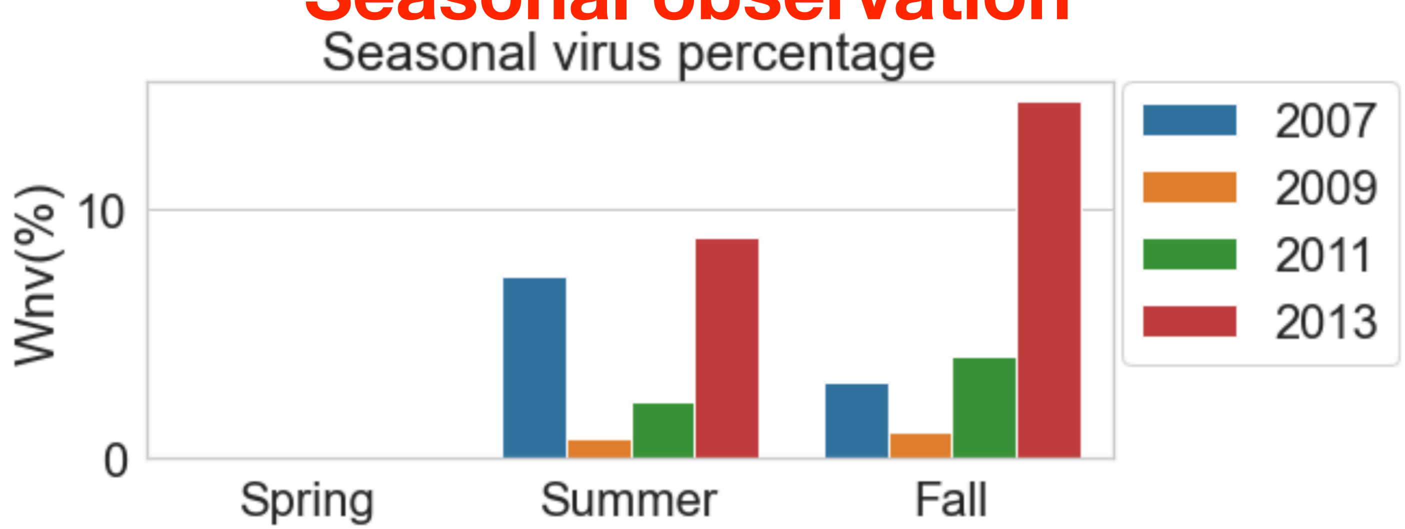
Yearly observation



Monthly observation



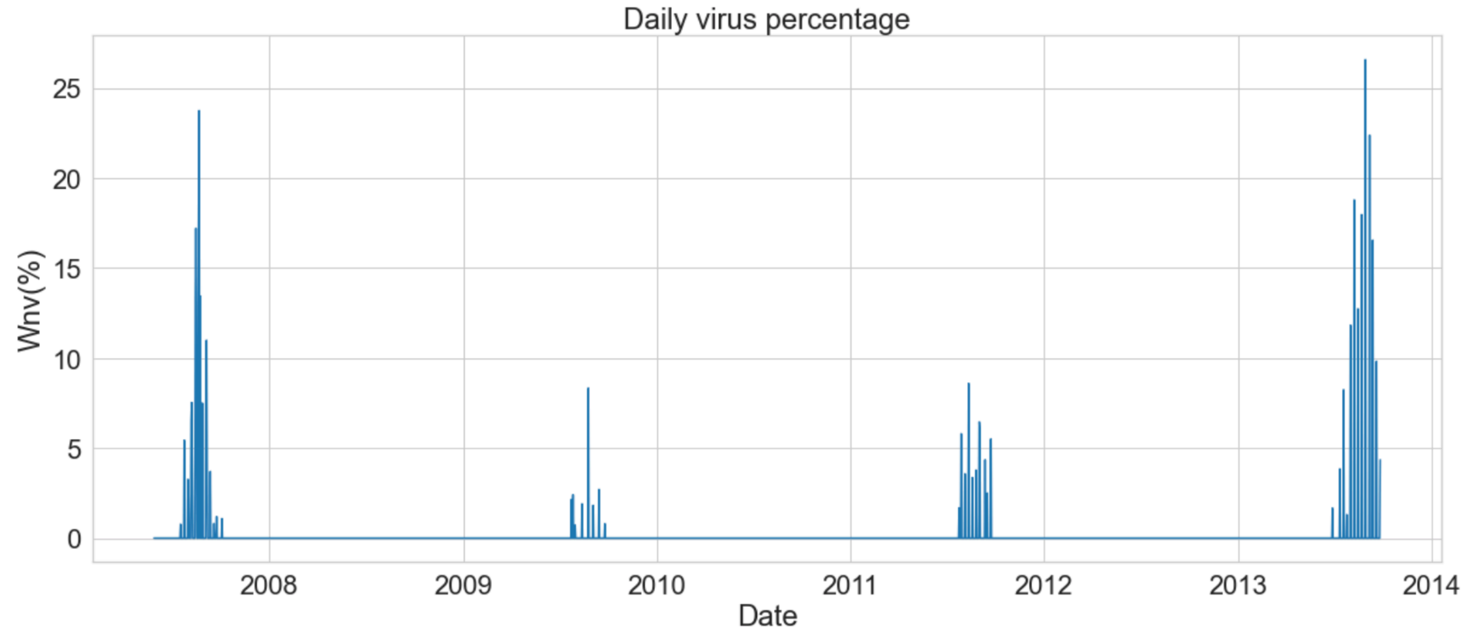
Seasonal observation

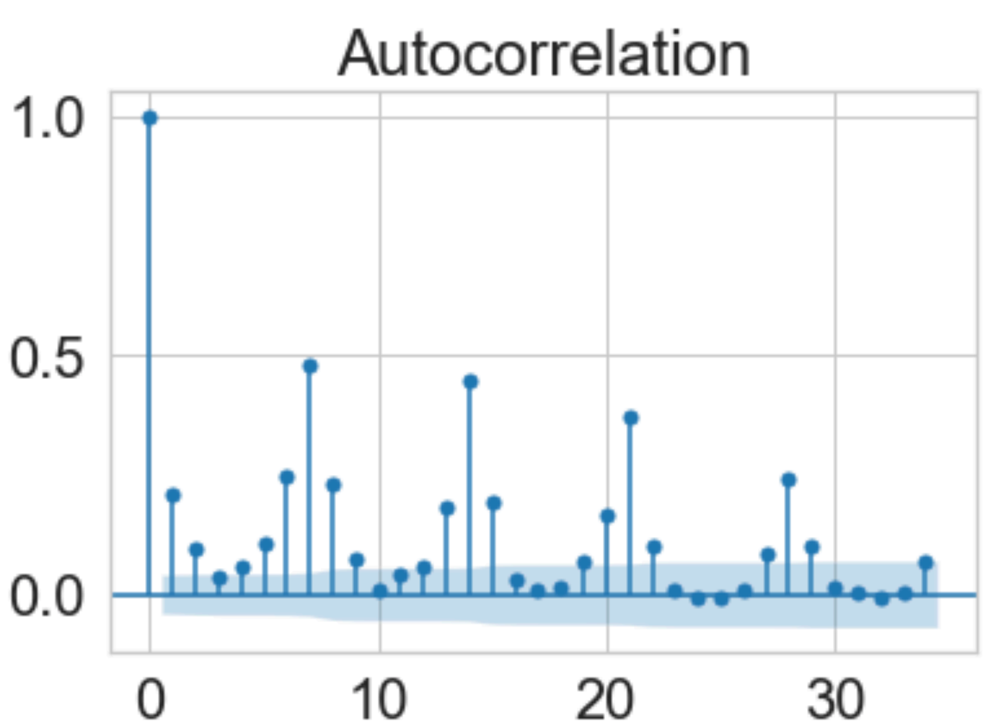


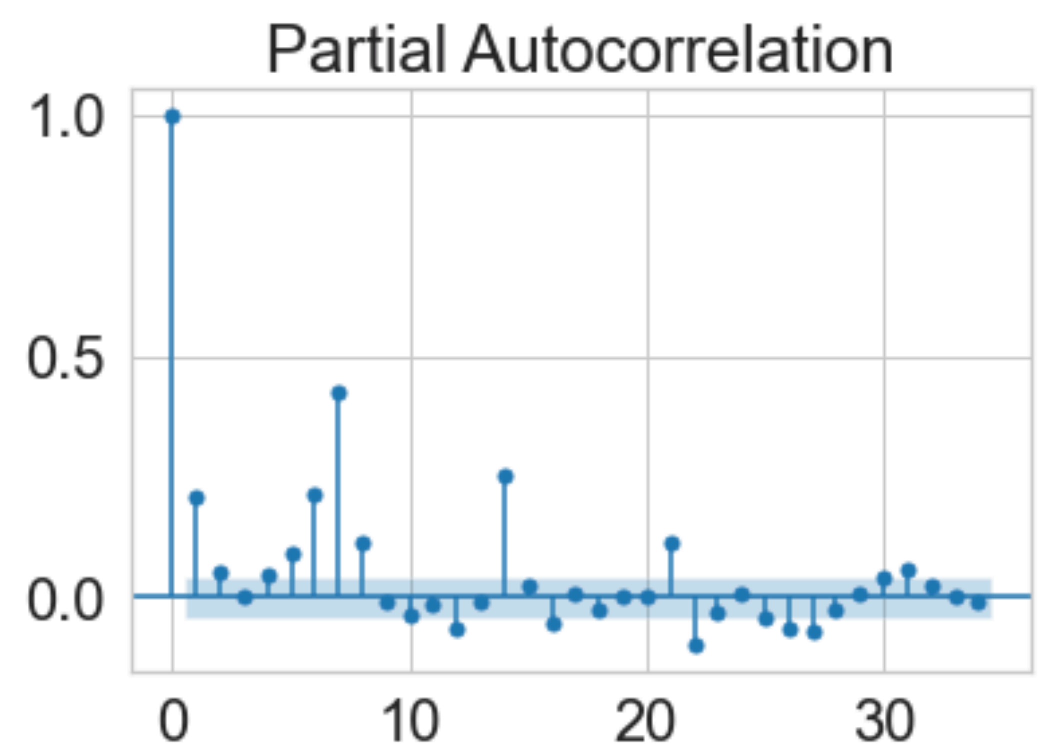
Exploratory data analysis

- Repeating pattern of virus
- Correlation with 7, 14 and 21 days are significantly higher than other days
- The pattern repeats in partial autocorrelation indicating periodicity in virus observation

Daily virus observation probability

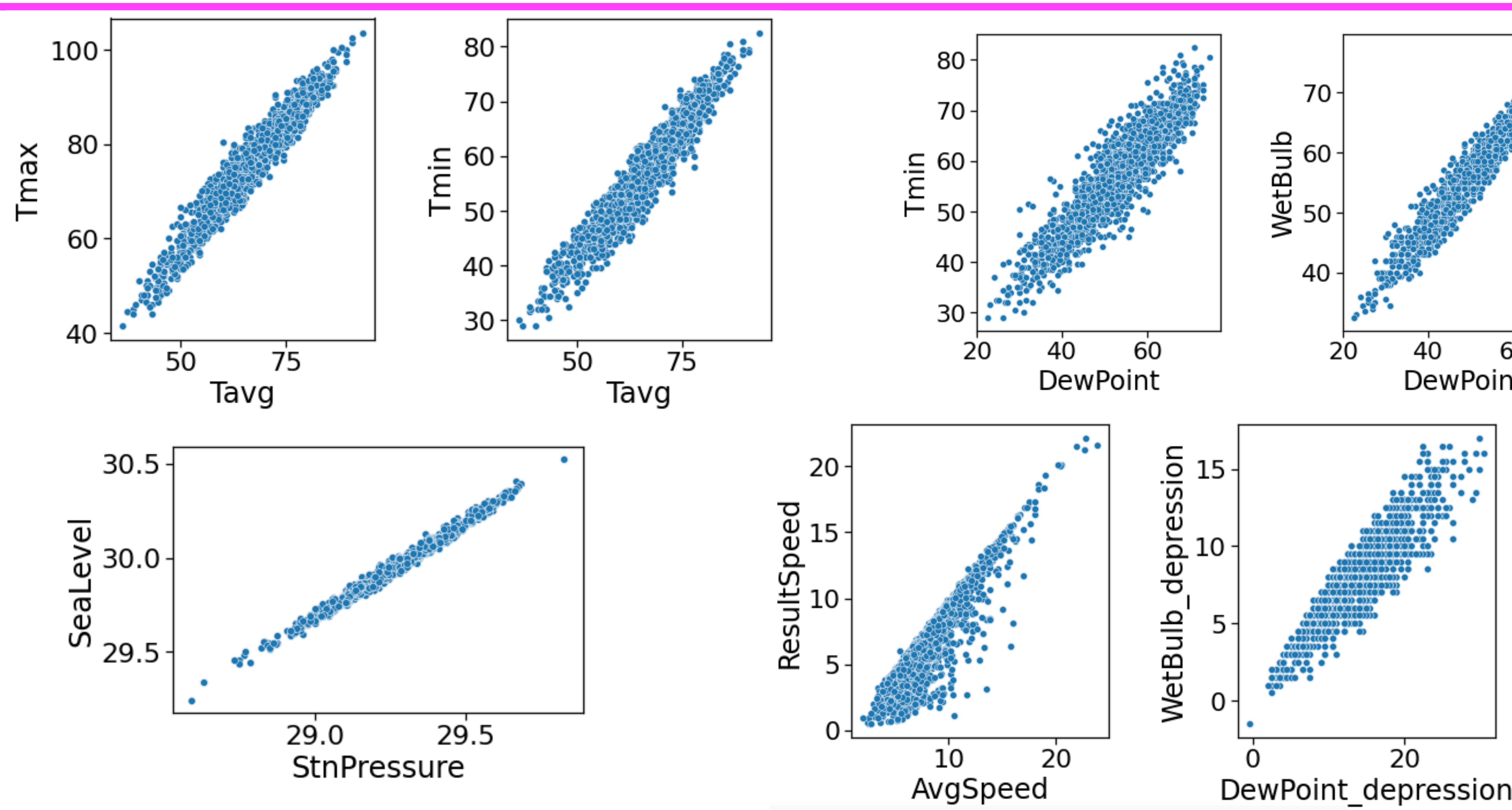






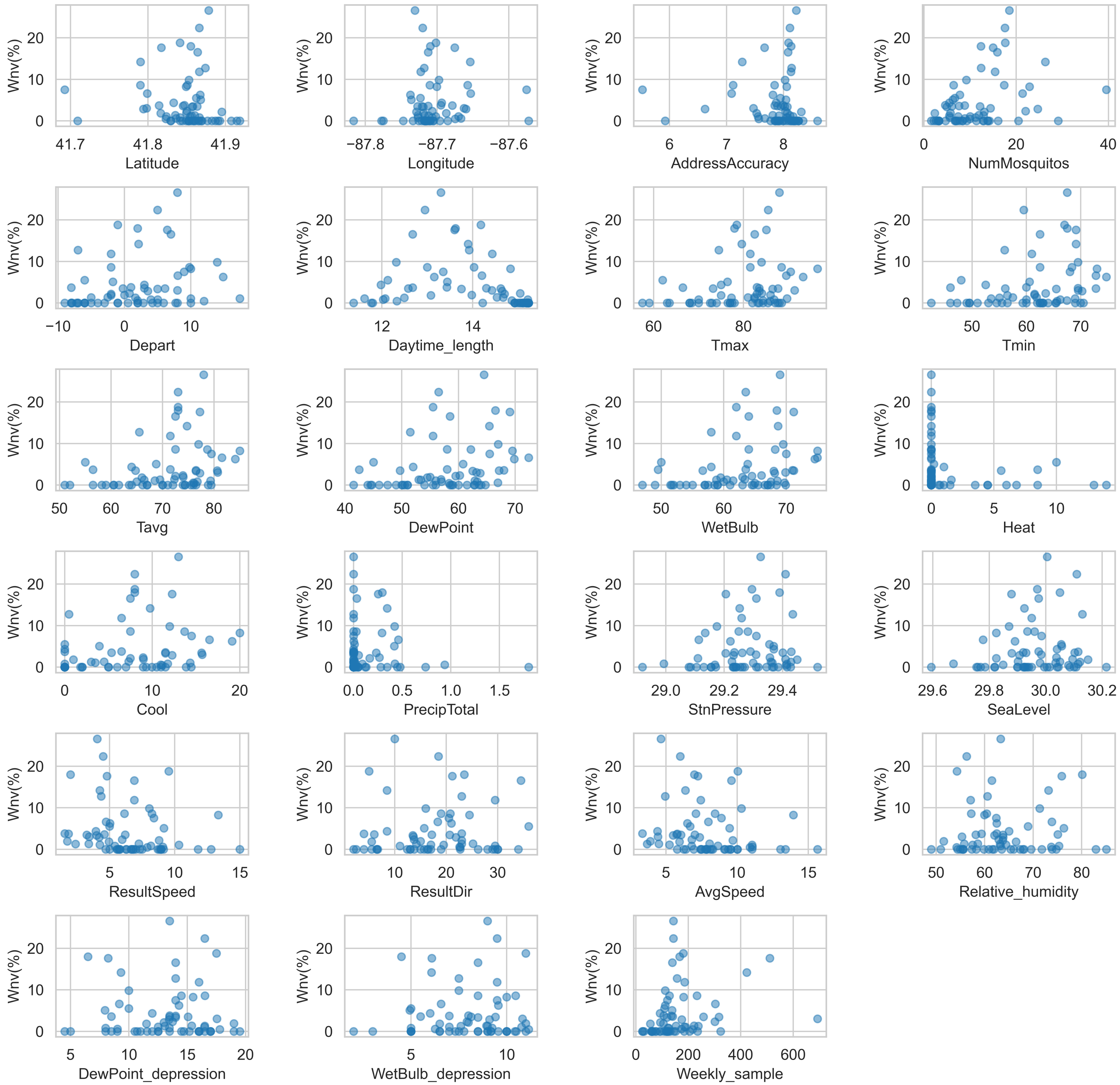
Weather data

- Addition of new features such as wetbulb depression, dewpoint depression and relative humidity
- Many weather parameters are correlated in heat map and scatter plots
- Correlated weather parameters
  - Average, maximum and maximum temperature
  - Sea level and pressure
  - Result speed and Average speed
  - Dewpoint and bulb



Exploratory data analysis

Relation between weekly virus observation probability and weather parameters



Feature engineering

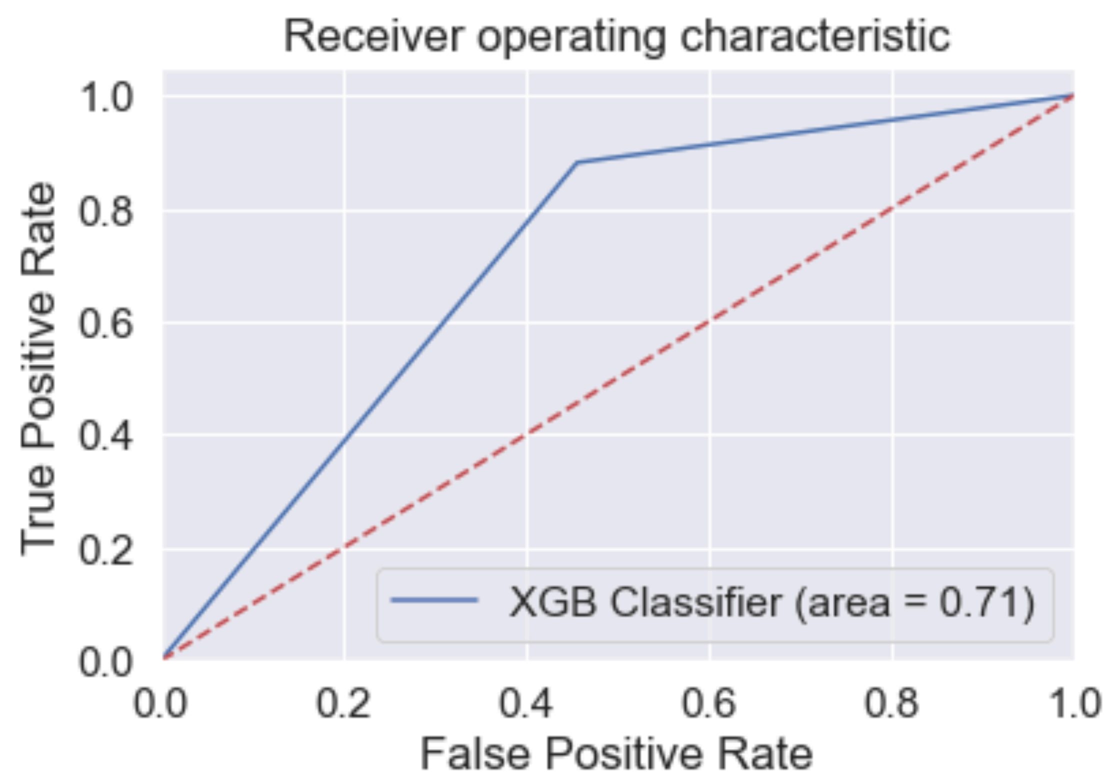
- Available weather parameters from both stations are averaged
- Fill missing values with the average of backward and forward filling method
- Remove correlated weather parameters retaining only one
- Merge weather data with GIS data on date
- Add additional features such as day of week, day of month, week of year, month, season and year for gaining seasonality on virus observation
- Addition of daily virus observation probability based on 7, 14 and 21 days lag
- Hot encoding of month, season and mosquito species
- Removal of hot encoded month and mosquito species without virus
- Drop object columns describing address retaining only latitude and longitude

Model Preprocessing

- Apply information value (IV) technique for the selection of important features
  - select only features displaying IV statistics within a range of 0.1-0.8
  - IV < 0.1: not useful for modeling
  - IV > 0.8: biased relationship with a dependent variable
- Reduce multicollinearity issue using variance inflation factor (VIF) technique
  - Features exhibiting VIF greater than 5 exhibit extreme multicollinearity and are removed
- 11 features related to time lag, weather, mosquito species and time periods are retained at the end

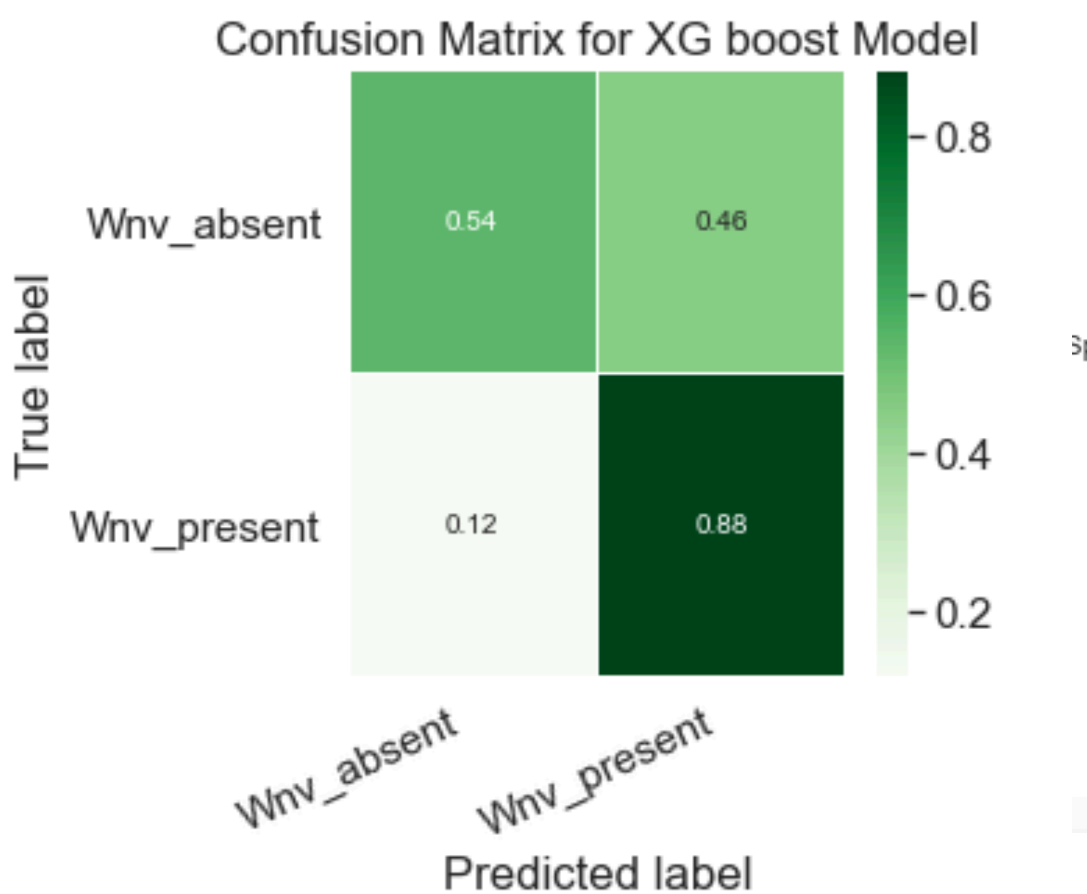
Data Modeling

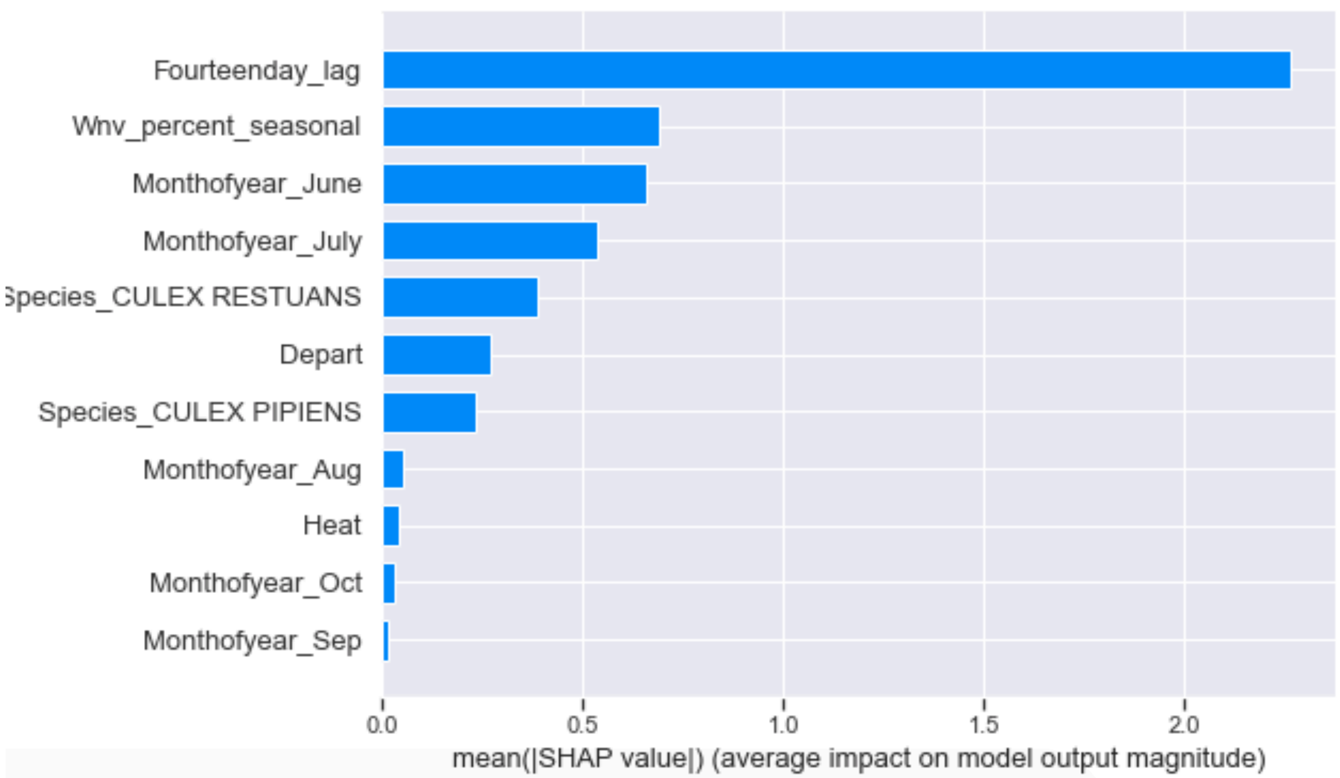
- Partitioning of data into training and test sets with a size ratio 7/3
- Data modeling using XGB classifier
  - Selection of best parameters using grid search method with five fold cross validation
  - Best score: 0.81
  - metric - area under the curve
- Model prediction:
  - AUC of 0.71 on test set



Results:

- Model predicts the presence of virus with a probability of 0.88
- The cost of incorrectly classifying the presence of virus (false negative) is riskier for virus outbreak than false positive
- Smaller magnitude of false negative. Thus, the model is critical in predicting the virus although the precision for virus absence is low.
- Fourteen day lag is the most important feature in classification decision

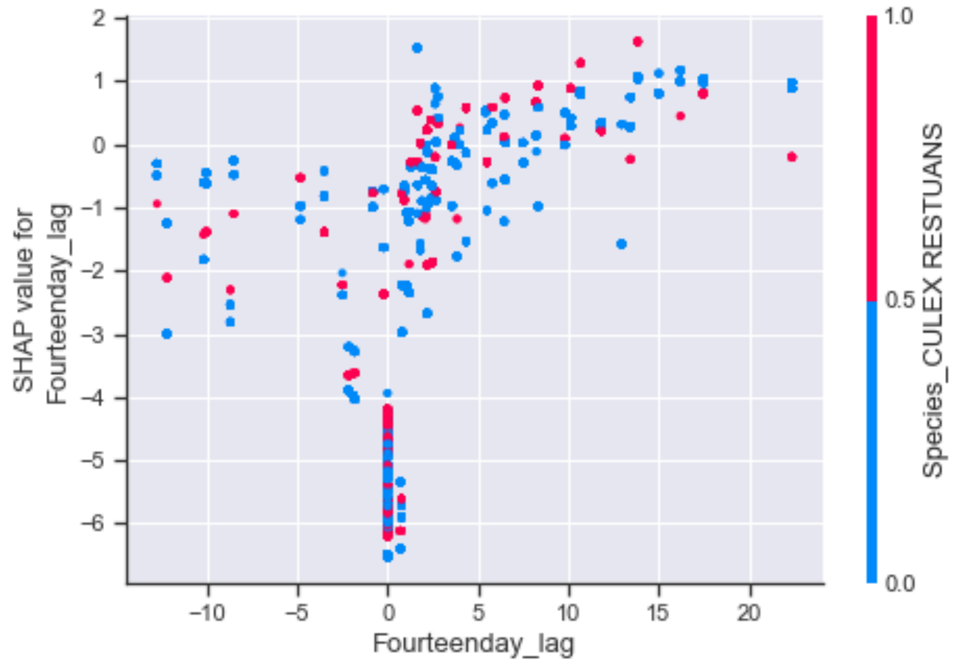


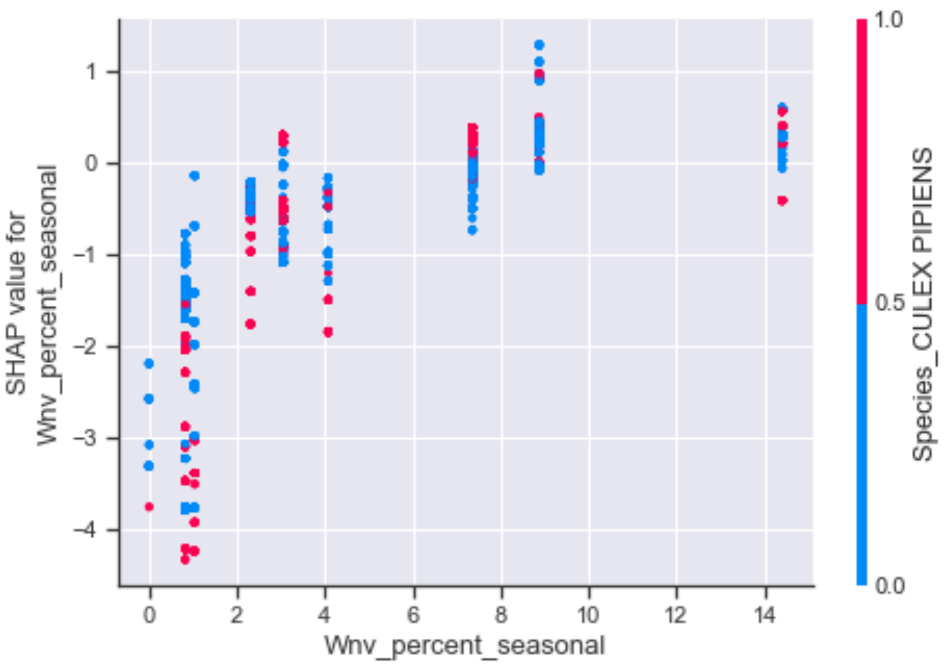


Impact of features on virus prediction

- Fourteen day lag and seasonal virus observation probability are positively correlated with the presence of virus
- Months alone are negatively correlated except August but their interaction with lag values and mosquito species increases virus predictions
- Impact of weather parameters depart and heat index are minimum compared to fourteen day lag, seasonal and monthly observations

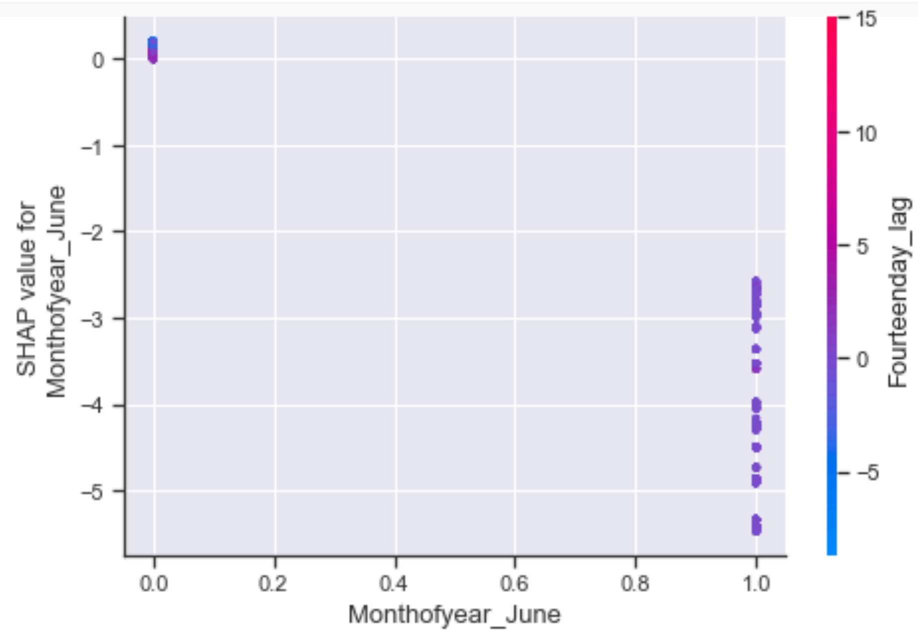
Partial dependence plots

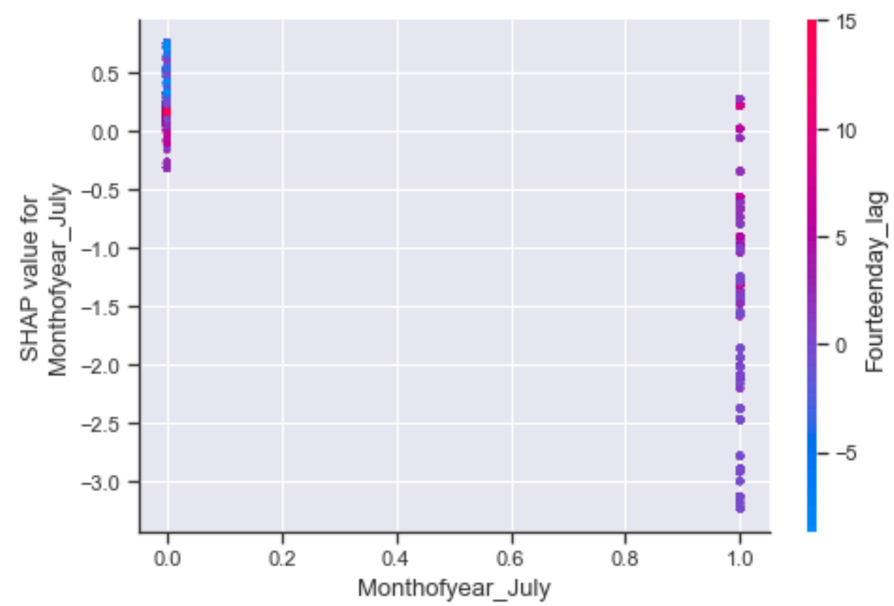


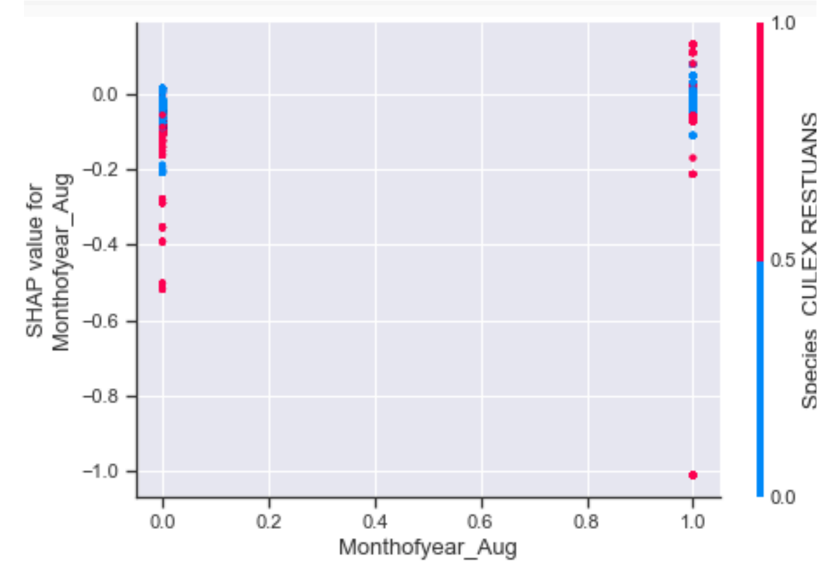


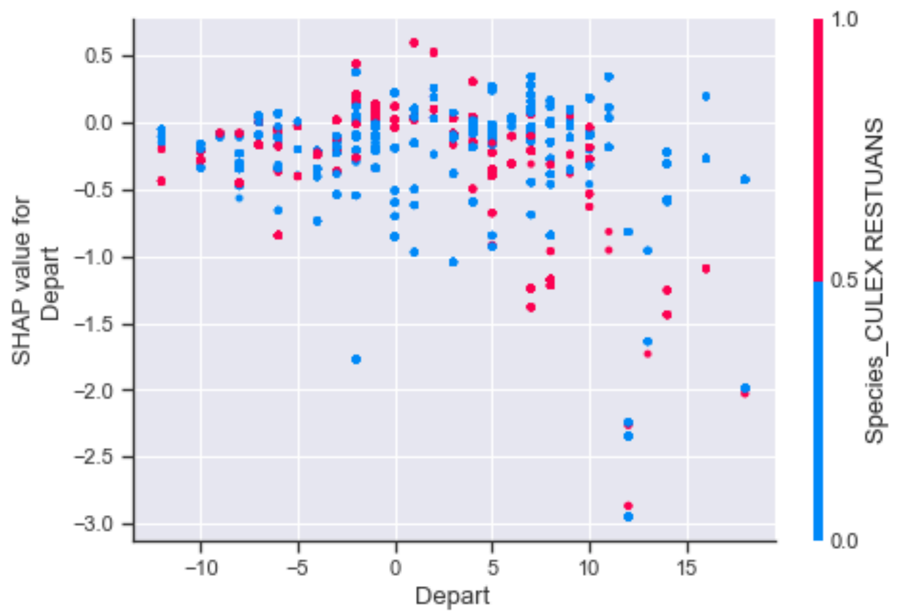
- Shap value increases as the magnitude of fourteen day lag value becomes positive. It exhibits some sort of linear relationship with the target variable and the spread suggests it's interaction with the mosquito species, Culex Restuans
- Virus prediction increases monotonically with the seasonal virus observation probability

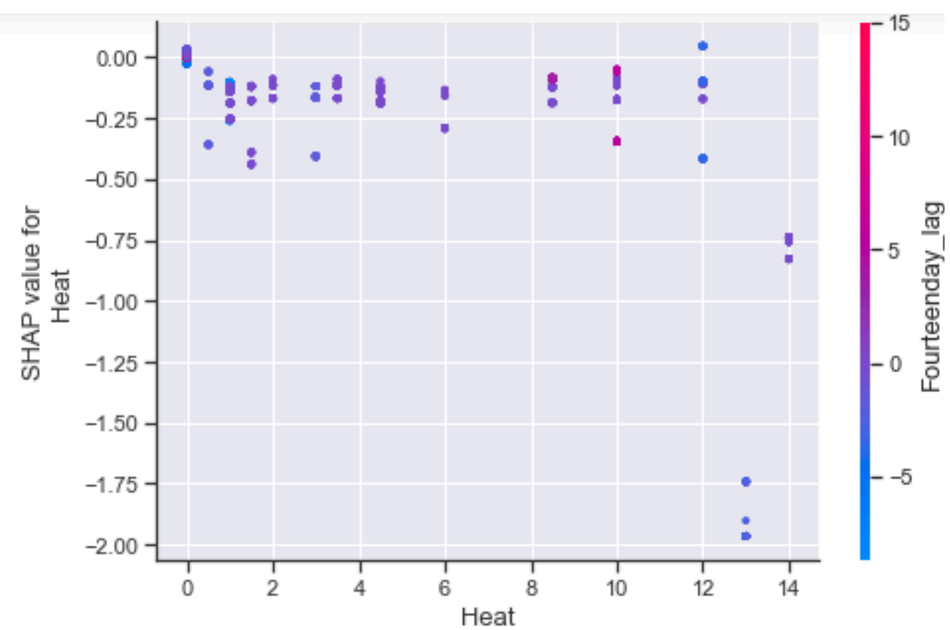
Partial dependence plots











- Higher shap value in August than June and July
- Shap value horizontal across a wide range of temperature departure form 30 years average temperature and heat index

Conclusions

- Virus prediction depends on time series data.
- Prediction increases if the virus was observed two weeks prior (fourteen day lag)
- In June and July, virus prediction depends on fourteen day lag value.
- In August, virus prediction increases for mosquito species Culex Restuans and Culex Pipens
- Weather parameters have minimum affect for virus predictions in the developed model
- Model performance may be improved by creating rolling features and percentile of virus observation over multiple rolling time windows.