West Nile Virus Prediction

Introduction

- Transmitted through mosquito bites
- Can lead to sever symptoms
- The Chicago Department of Public Health (CDPH) began monitoring the virus by2004

Objective

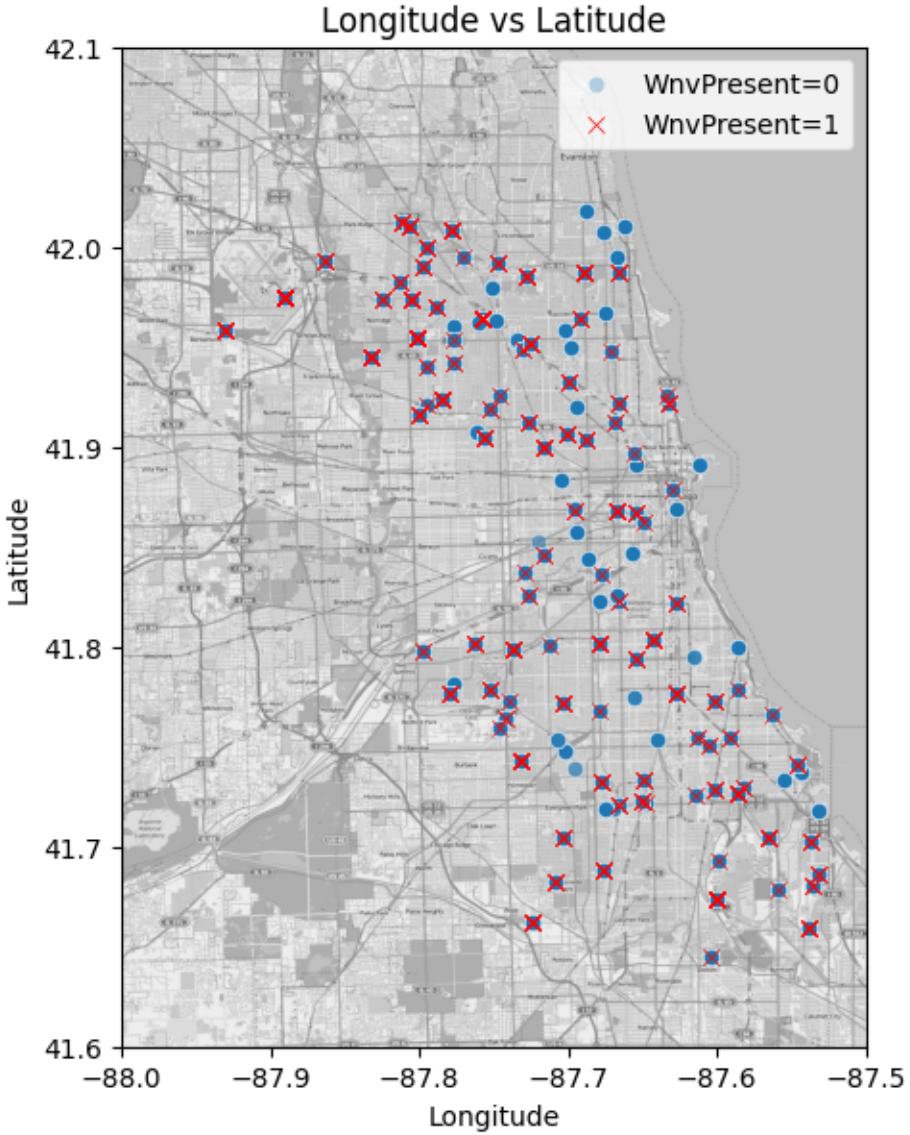
- Create a machine learning model to predict time and location of the presence of the virus to allocate disinfectant spray accordingly

Data

- GIS Data
  - Date, Location information, mosquito species, label indicating presence of the virus
- Weather data
  - Weather information including, temperature, pressure, precipitation, and dewpoint from two stations
  - Average from the two weather stations
- Merge on date

GIS Data

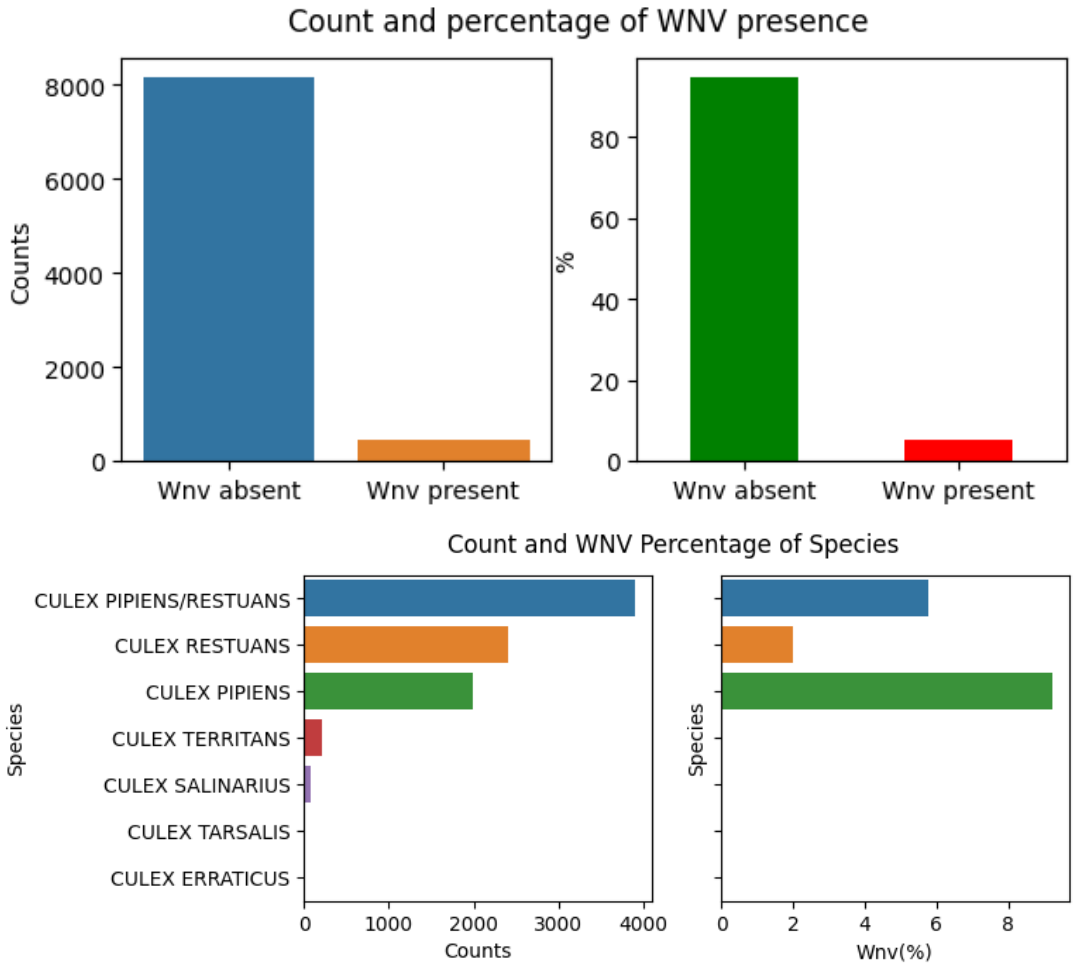
- Contains samples from 138 different locations, collected in 2007, 2009, 2011, and 2013
- Samples collected between may and october
- WnvPresent indicates the presence of the virus (1 = presence, 0 = absence)



Exploratory data analysis

balance and species

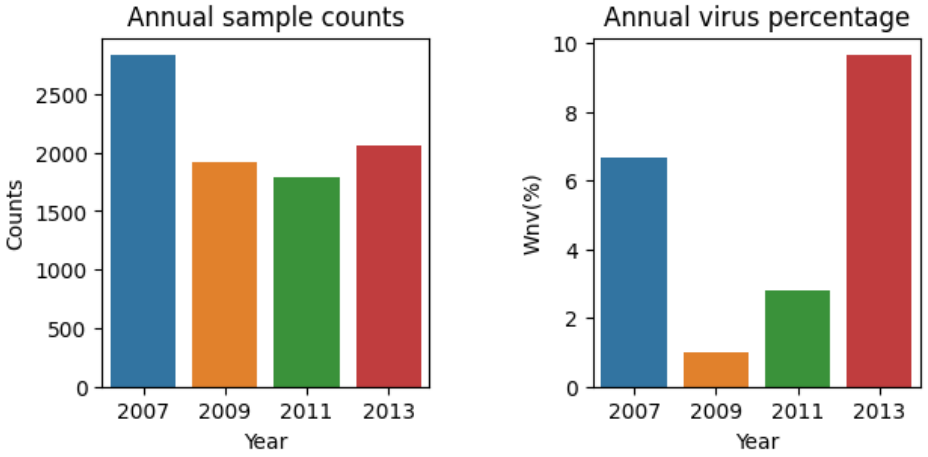
- Highly imbalanced dataset
- Only two species tested positive for the virus

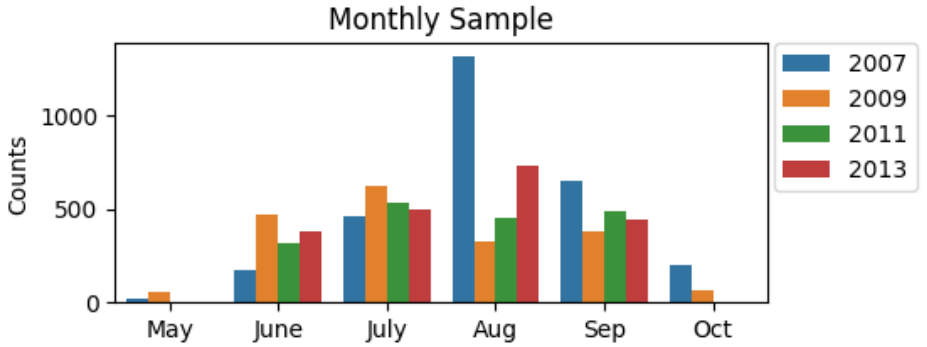


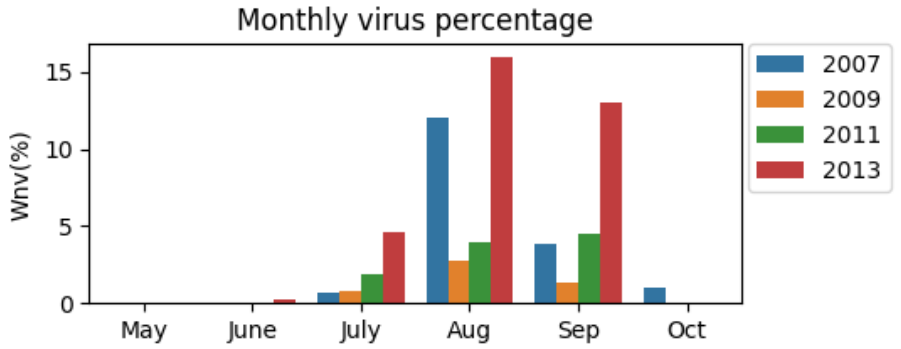
Exploratory data analysis

Seasonality

- Virus is significantly more present in 2013 and in the month of August
- No clear indicator of general trend
- Data is limited to only four years



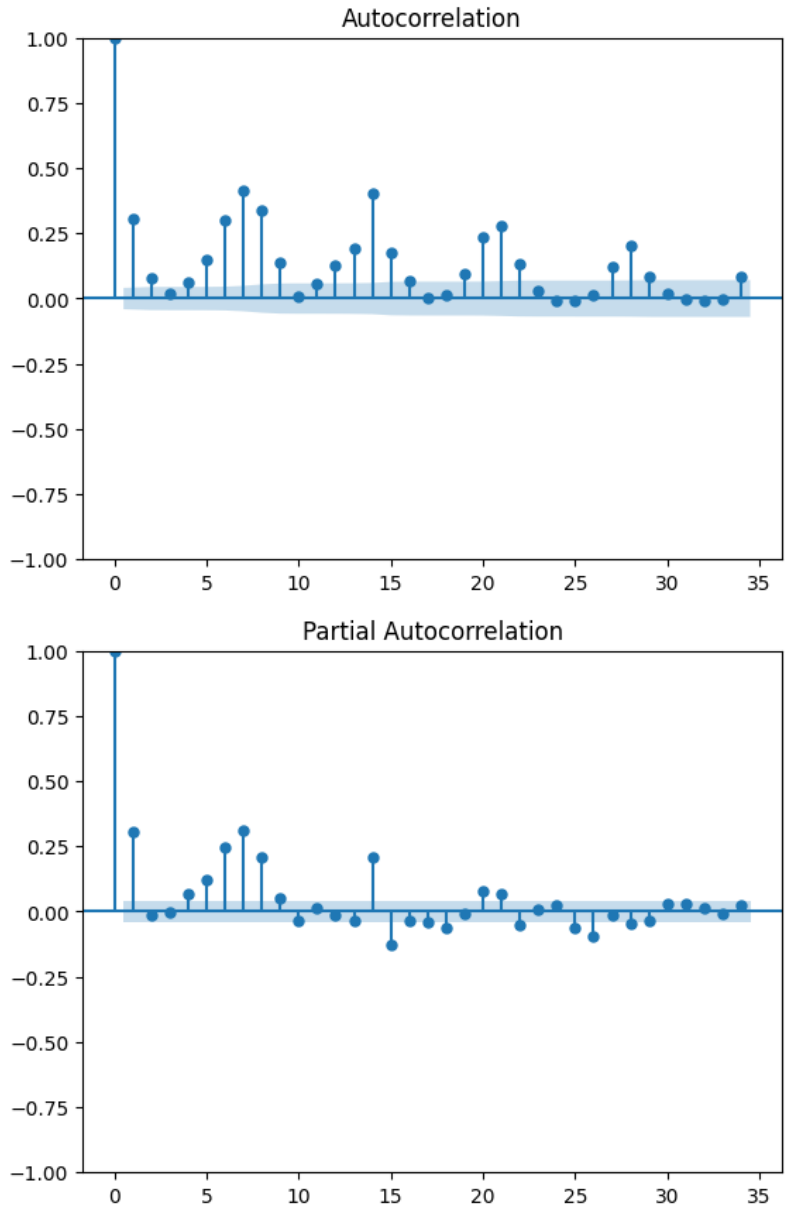




Exploratory data analysis

Daily virus presence

- 7 day repeating pattern
- High correlation with 7, 14, and 21 day lags



Feature engineering

- 7, 14, and 21 day lag features added to dataset
- Date column split into separate day, month and year columns
- Wetbulb depression, dewpoint depression and relative humidity features added to datset
- Annual, seasonal, monthly and weekly virus percet added to dataset
- Highly correlated and redundant features removed from dataset
- One-hot-encoded Species, Season, and Month_Name features
- Species and months that did not show the virus were removed from the dataset
- Retain only latitude and longitude columns for location

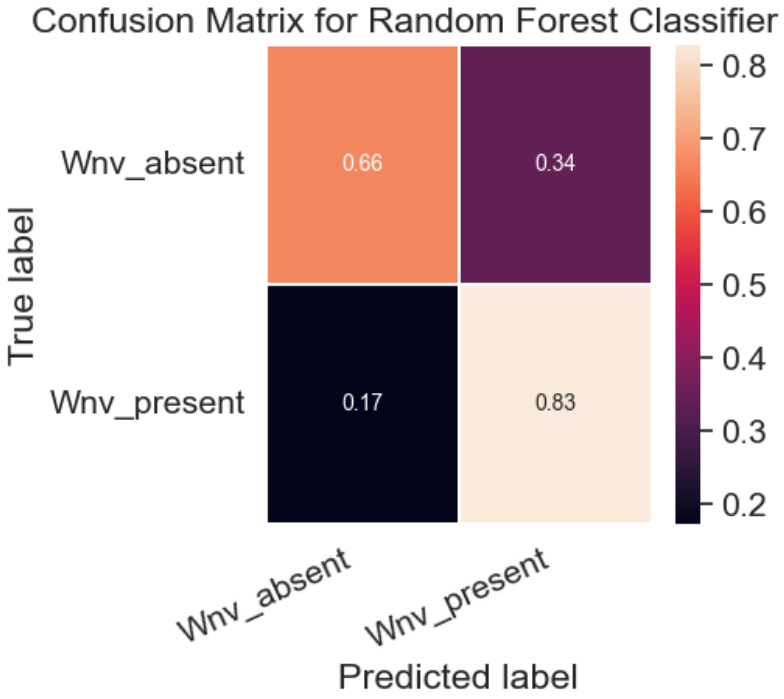
Preprocessing and Training

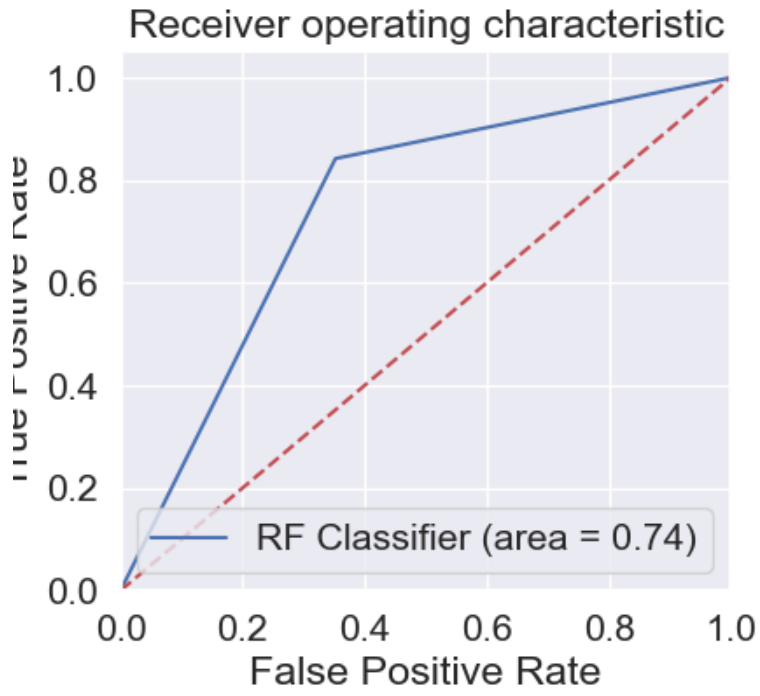
- Undersampling of majority class for a balanced dataset
- 70-30 train-test split
- Weight of evidence technique used to select important features
  - IV < 0.1: not useful
  - IV> 0.8: potential biased relationship
- 13 remaining features
  - Time lags, weather, species, date
  - No more features indicating location

Modeling and evaluation

Random forest classifier

- 5 fold cross validation using grid search
- AUC score: 0.74
- Rcall score: 0.83

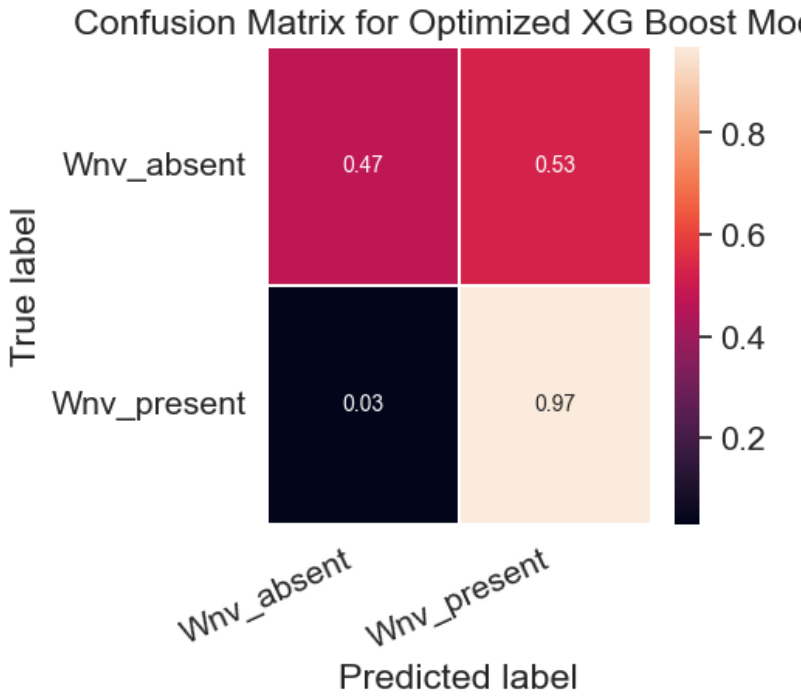


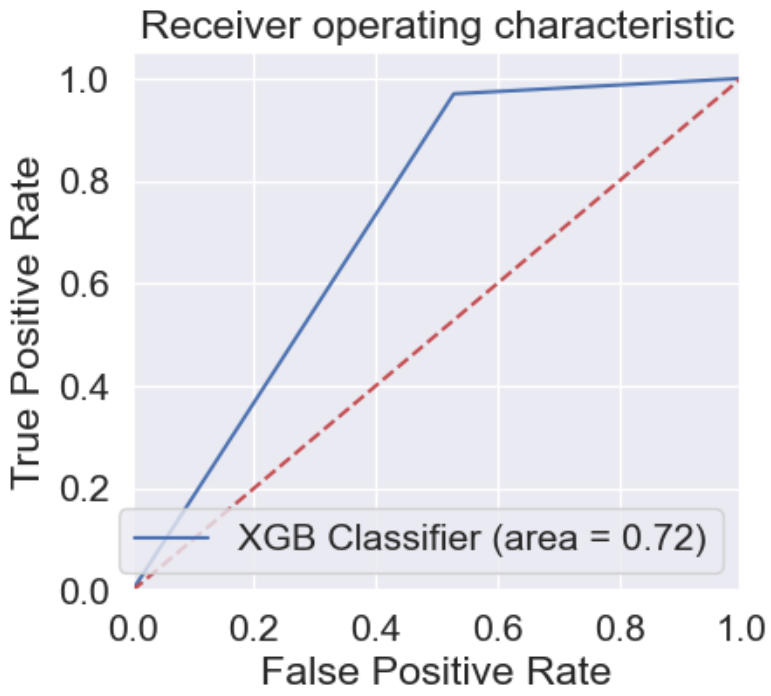


Modeling and evaluation

XGBoost classifier

- 5 fold cross validation using grid search
- AUC score: 0.72
- Rcall score: 0.97
  - Very effective at determining the presence of the virus
- More effective than random forest model





Results

- Great risk in misclassifying the absence of the virus as it could lead to an outbreak
- XGBoost model was most effective at determining the presence of the virus
  - AUC score: 0.72
  - Recall score: 0.97
- Fourteen and seven day lags were the most important in determining the presence of the virus
- CDPH should modify control program to take preventive measure as soon as fourteen day lag value is positive.
- There are no indications that location was a strong indicator for the presence of the virus