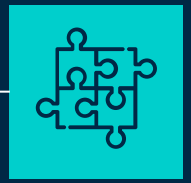
West Nile Virus Classification

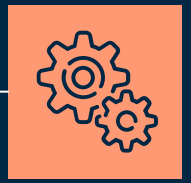
By Group 3

TABLE OF CONTENTS



01

Background/Problem Statement & EDA



02

Model selection and features

03

Conclusion & Recommendations

Background/Problem Statement & EDA

01

Background

- Epidemic of West Nile Virus in Chicago, a virus transmitted by certain species of mosquito.
- Over 37,000 (underestimated) WNV disease cases have been reported since 1999.
- One in 150 of those infected develop a serious nervous system illness that typically requires hospitalization.
- Hospitalisation cost median $7,500 (less serious) / $22,500 (serious) cases.

Source

Problem Statement

- Department of Health wants to reduce the spread of virus which in turn reduces the hefty healthcare cost.
- As data scientists, we will build model to predict whether particular location having WNV-carrying mosquitos.
- Evaluate effectiveness of annual spray program.

Datasets

- Train set (Odd years - 07, 09 etc...)
- Test set (Even years - 08, 10 etc...)
- Spray data (for year 2011 and 2013 only)
- Weather data (daily)

Data Cleaning Process

Removed station 2 weather data

Station 1 weather Data

Missing Data

Replaced “M” and “-” with None. “T” with 0.00001 (indicates very little rainfall)

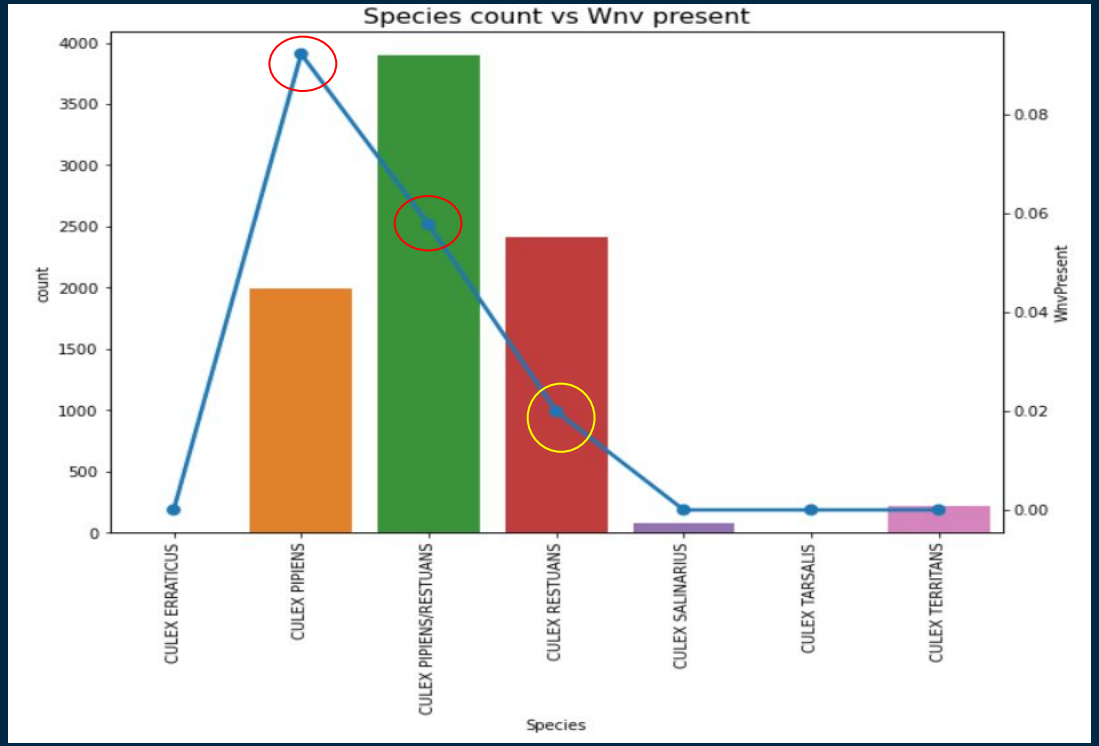
Dropped columns with more than 90% missing data / mostly 0 / 1 unique value

Dropped (Water1, Depth, Snowfall)

Combined with Train data

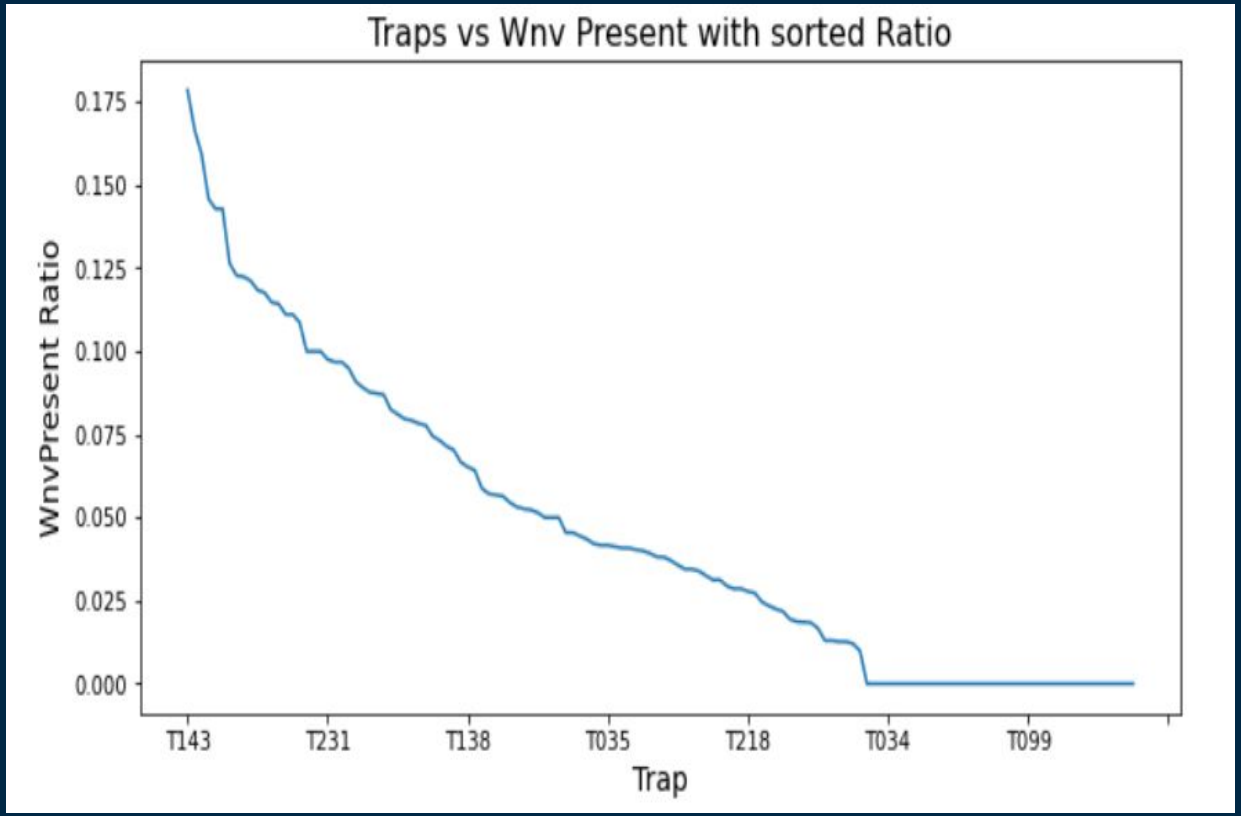
Merged the cleaned weather data with the Train data on date

EDA: Species vs. West Nile virus present



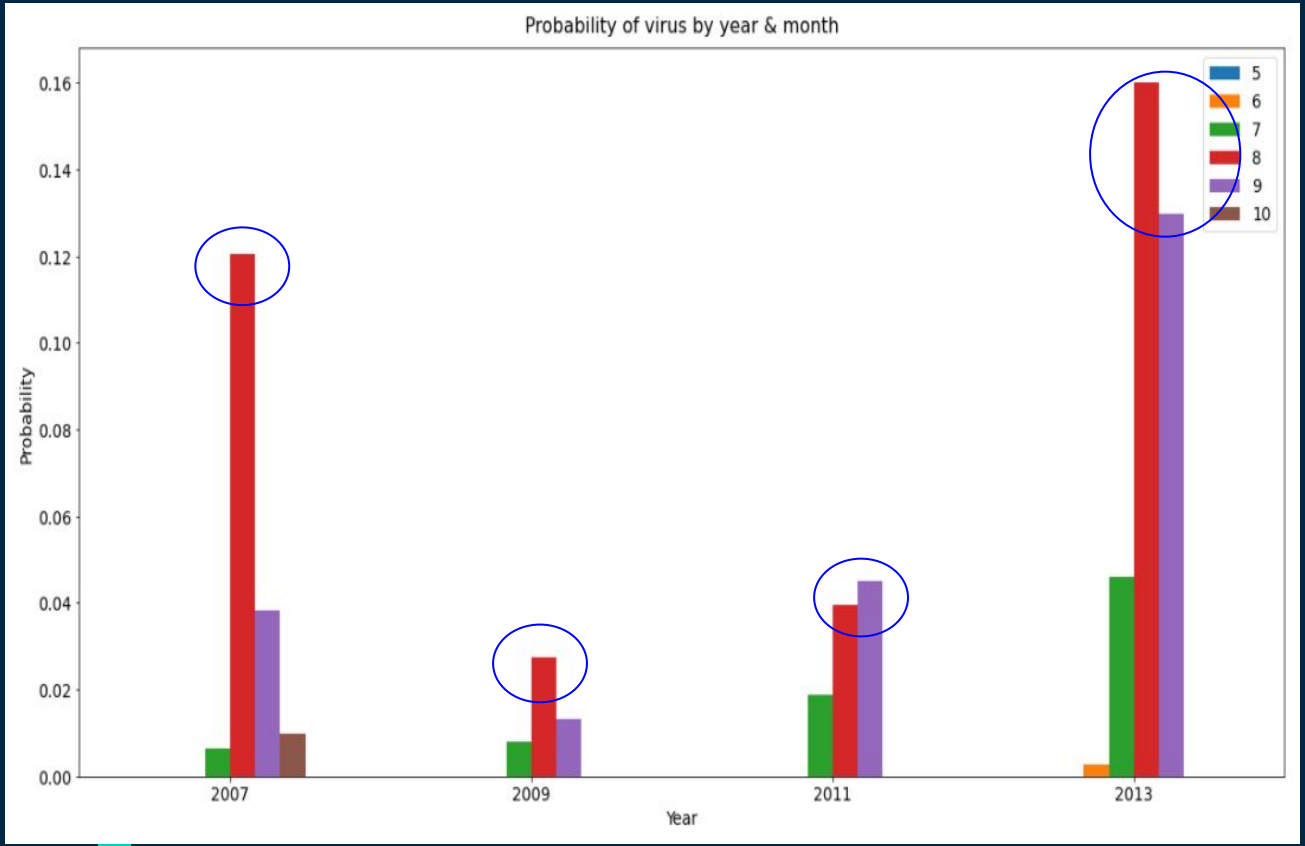
- Only 2 species with Wnv presence
- Pipiens > Pipiens / Restuans > Restuans
- Encode species with ordinal values

EDA: Trap vs. West Nile virus present



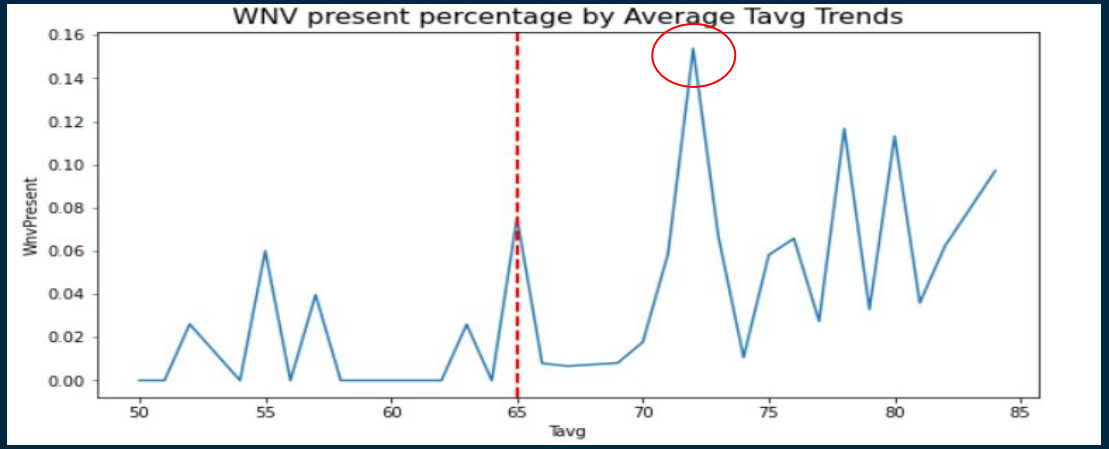
- Certain trap locations with higher Wnv presence
- Some traps have 0 Wnv presence
- Encode Trap with ordinal values

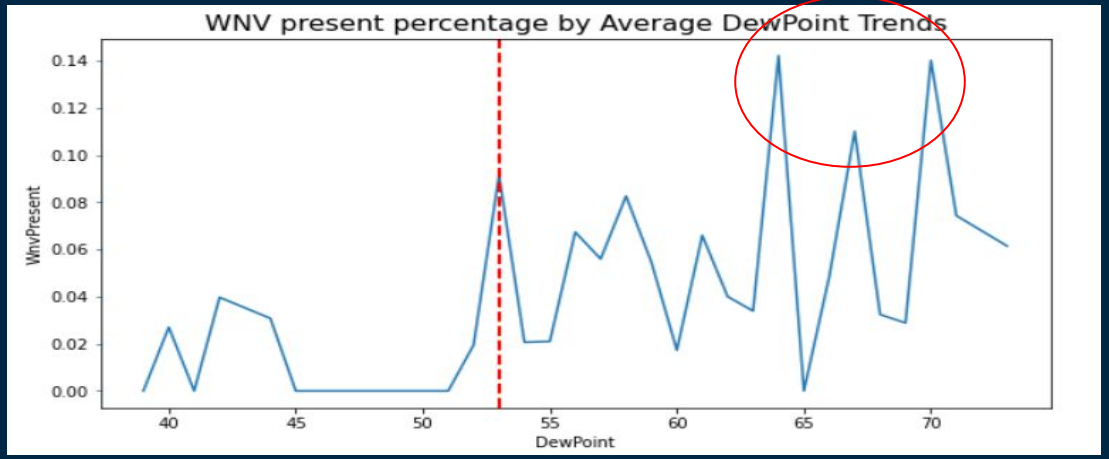
EDA: Month vs. West Nile virus present



- Months of Aug & Sep - highest Wnv presence
- Summer in Chicago (June - September)
- Year no clear trend
- Spike in Year 2007 and 2013
- Encode only Month with ordinal values

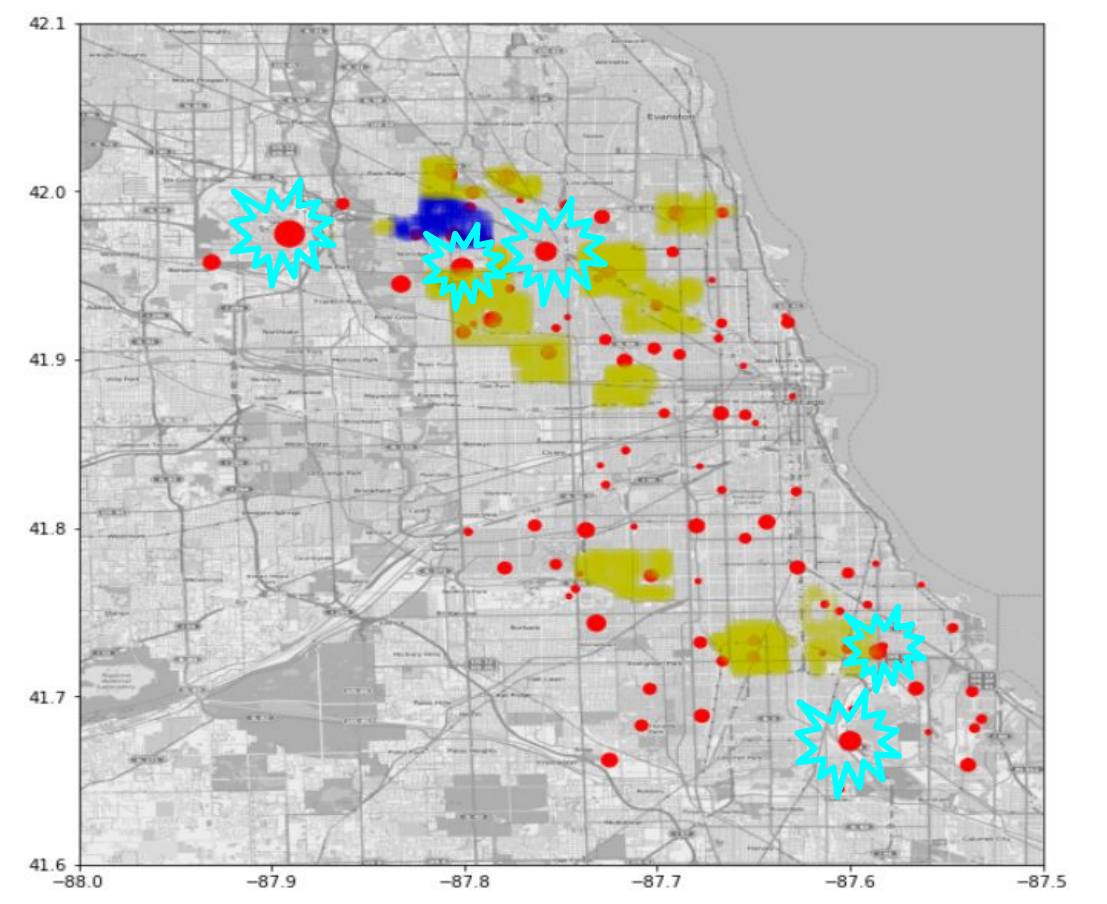
EDA: Tavg/DewPoint vs. West Nile virus present





- Temperature around 72°F highest Wnv presence
- Too high or too low temperature reduces Wnv presence
- Humidity around 64-70°F Td highest Wnv presence
- Low humidity reduces Wnv presence

EDA: Map plot for WNV present and spray



- Areas sprayed not on areas with high Wnv presence (2011 - yellow, 2013 - blue)
- Locations with high Wnv presence (Teal - star)

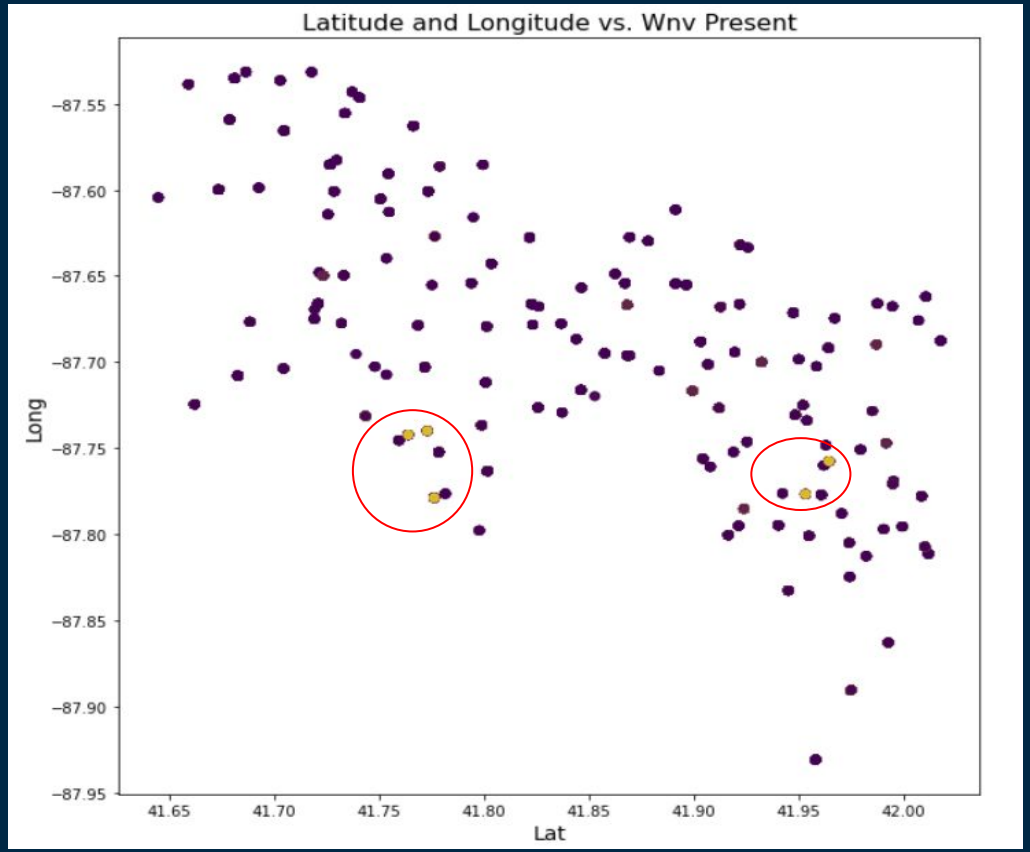
Feature Engineering

- Encode rain related CodeSum to **IsRain** (1 if rain else 0)
  - Rain > puddles > stale water > breeding conditions
  - #TS THUNDERSTORM, #GR HAIL, #RA RAIN, #DZ DRIZZLE, #SH SHOWER

- **DayInMins** by taking time of Sunset minus time of Sunrise and convert it into minutes.
  - Higher temperature results in higher Wnv presence
  - Longer daylight increases Wnv presence

- **Weather lag** by 1, 7, 14 days
  - Follows mosquito life cycle which is 7-14 days
  - Stronger correlation than original date weather?

Feature Selection



- Wnv presence no obvious trend with Lat/Long
- Correlation close to 0
- Drop address related features.

Feature Selection

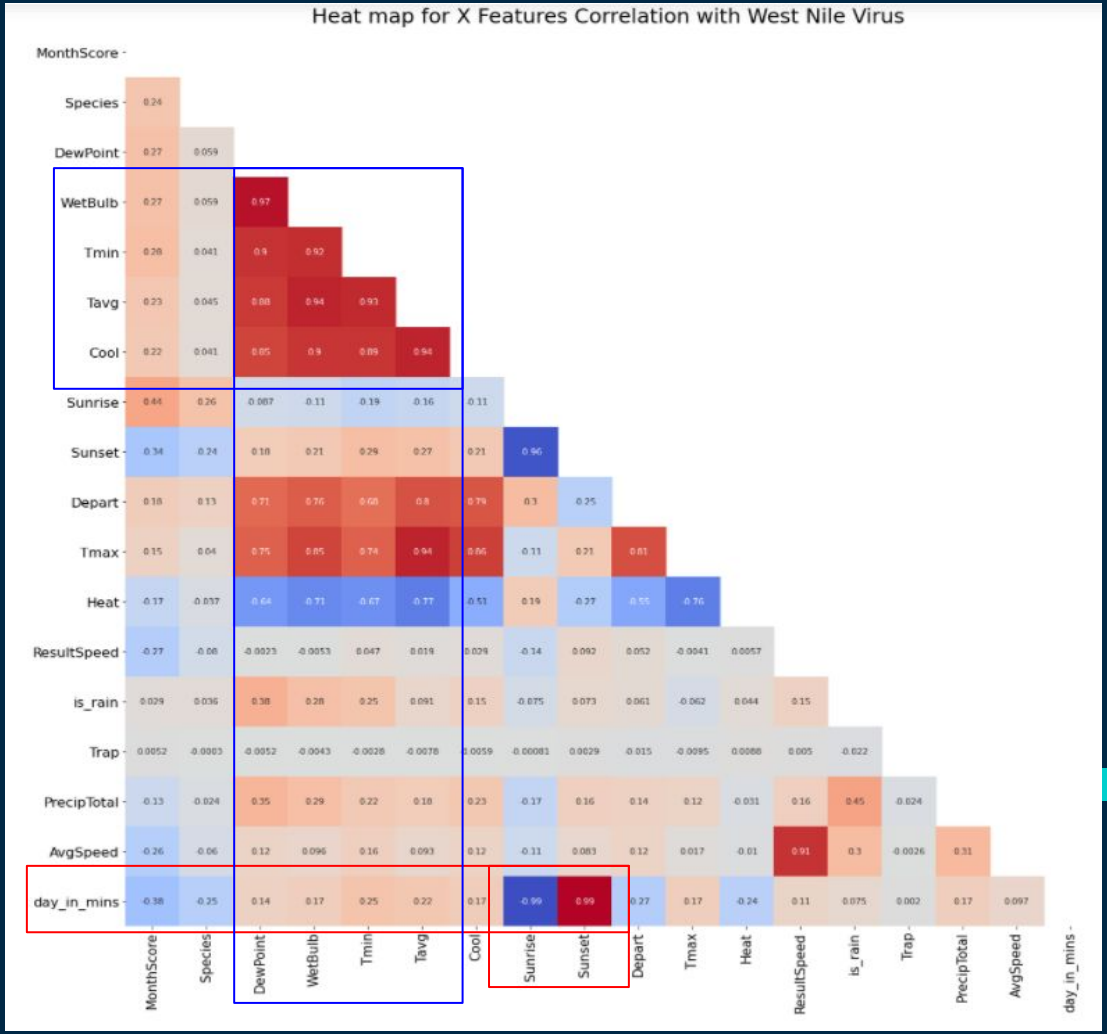
- Selected features based on highest correlation
- Did not include "lag" features as they do not shower higher correlation

	WnvPresent
WnvPresent	1.000000
MonthScore	0.186121
Trap	0.171282
Species	0.123341
Sunrise	0.096179
DewPoint	0.096124
DewPoint_lag_1	0.095385
DewPoint_lag_14	0.094724
WetBulb	0.094166
WetBulb_lag_1	0.093698
WetBulb_lag_14	0.092466
DewPoint_lag_7	0.092253
WetBulb_lag_7	0.089764
Tmin	0.086730
Tmin_lag_1	0.084910
Tmin_lag_14	0.083847
Tmin_lag_7	0.081475
Tavg	0.079215
Tavg_lag_1	0.078781
Tavg_lag_14	0.076129
Cool	0.075605
day_in_mins	0.074981
Tavg_lag_7	0.073997
Depart	0.063704

Tmax_lag_1	0.062129
Tmax	0.061218
Heat	0.058976
Sunset	0.058570
Tmax_lag_14	0.058232
Tmax_lag_7	0.056318
ResultSpeed	0.055551
Year	0.042496
AvgSpeed	0.035324
is_rain	0.030431
PrecipTotal	0.025936
ResultDir	0.009709
StnPressure	0.003302
SeaLevel	0.002164

Final Features - multicollinearity

- Collinearity between our features
- Will not be removing them as our model will be able to deal with multicollinearity



Model selection and features

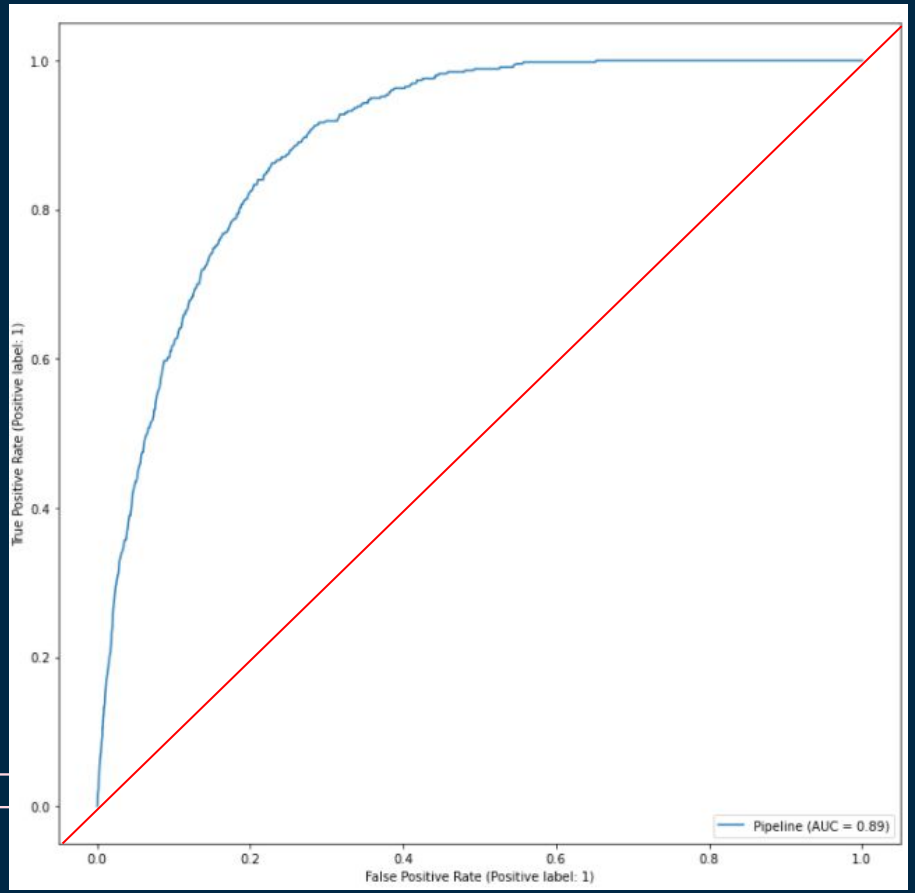
02

Summary of Model score (baseline ROC-AUC 0.5)

Models	ROC-AUC score for train data	ROC-AUC score for val data	ROC-AUC score for Kaggle
Logistic Regression	0.78	0.74	0.70
Random Forest	0.79	0.75	0.71
SVM	0.8	0.78	0.71
XGBoost	0.81	0.80	**0.76**

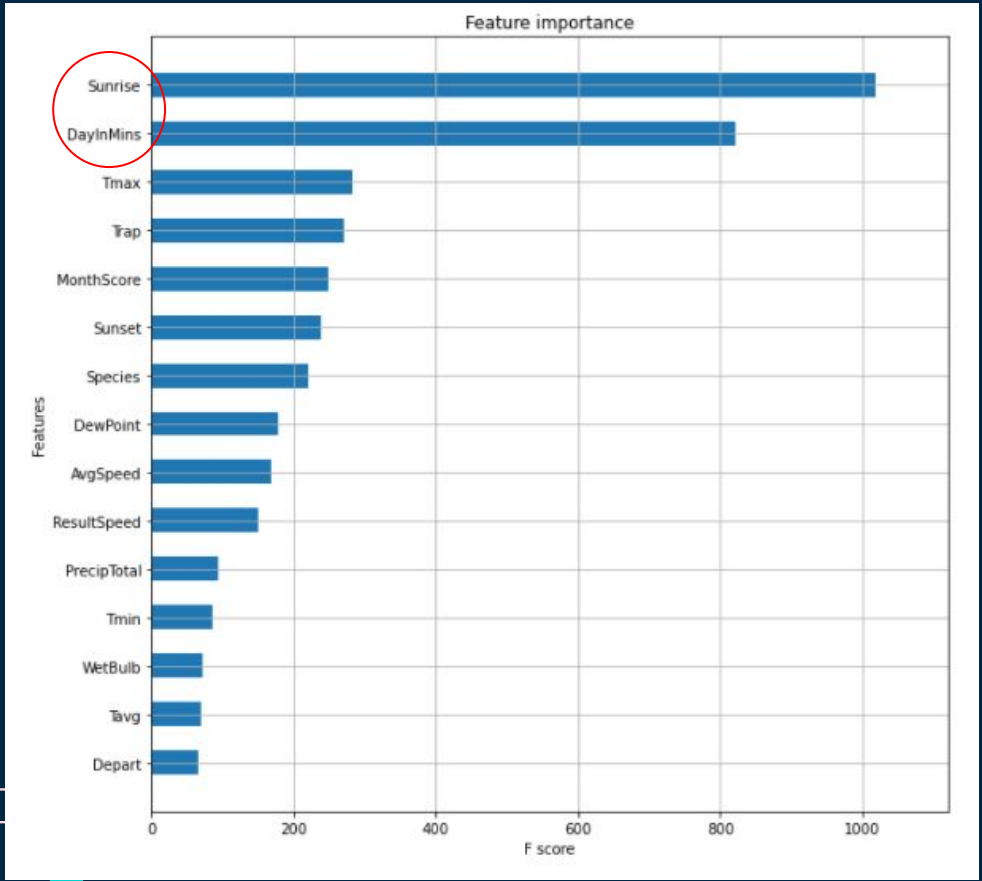
- XGBoost model handles colinearity better than logistic regression

Model selection (baseline ROC-AUC 0.5)



- XGBoost model highest Kaggle ROC AUC of 0.75
- Signs of overfitting: 0.85 on train, 0.75 on validation set
- SVM with RBF kernel second best Kaggle ROC AUC of 0.71

XGBoost – Top features for Predicting



- Sunrise and DayInMins strongest features
- Longer daylight more likely to have Wnv presence
- Features with multicollinearity have been dropped (Heat, Cool)

Conclusion & Recommendations

03

Cost Benefit Analysis of Spraying

Financial Cost of Spraying:

- Total area of Chicago is 606.1 km2.
- Cost $140,409 to cover the entire area.

Environmental Cost of Spraying:

- Biodiversity issues especially when dealing with natural protected areas.

Cost Benefit:

- Have to prevent at least 19 non-serious case of Wnv to justify the cost. ($7,500 hospitalisation cost).
- No noticeable decrease in WNV occurrences after spraying.

Conclusion and Recommendations

- Top features contributing to likelihood of virus are:
  - **Sunrise**
  - **Daylight**
  - **Max temperature**

- Current model predicts relatively well (0.76 ROC-AUC score), can possibly predict areas outside Chicago with similar seasons

- Improve features:
  - Learn more about mosquitoes behavior
  - Collect more information about problematic breeding areas

- Reduce spraying, focus on spraying on hot and humid days; the cost does not justify the benefits derived

- More effort in educating citizens to
  - Reduce mosquito breeding habitats: e.g. clear stale water, clean rain gutters
  - Wear protective clothing (long sleeve) during mosquito season

Do you have any questions?

THANKS