Predicting the West Nile Virus in Chicago

Jade Lam

Mar 2020

CONTENTS

Executive Summary

What is West Nile Virus?

The West Nile virus (WNV), according to CDC, is the leading cause of mosquito-borne disease in the continental United States. It is most commonly spread to people by an infected mosquito. WNV usually occurs during summer/ fall. There are no vaccines or medications for treatment.

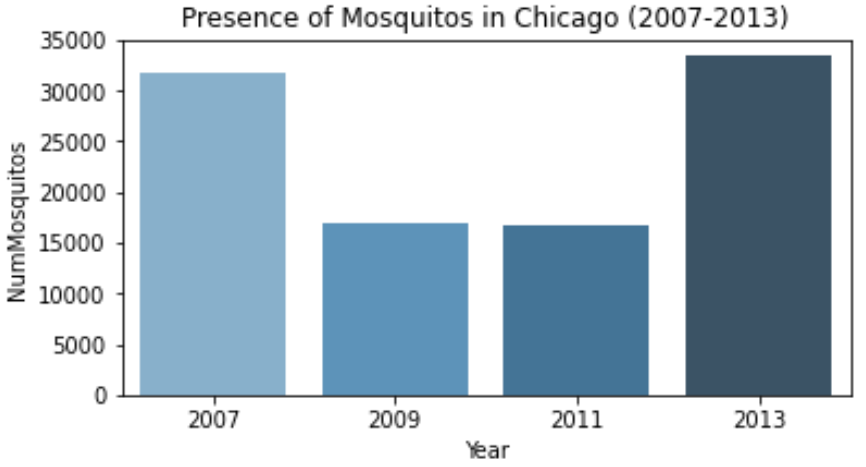
Project Objective

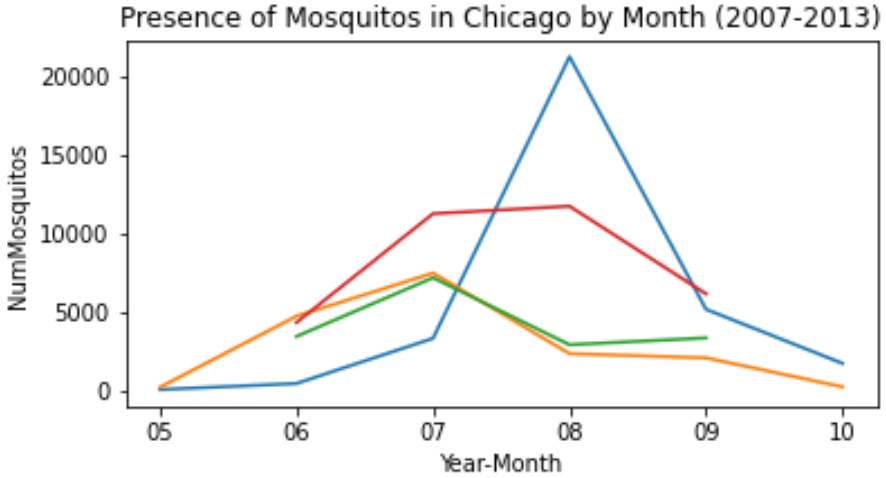
Through a methodical analysis of Chicago's West Nile Virus data, build a model to predict the presence of West Nile Virus in given a set of conditions in Chicago.

Project Conclusion

- Number of mosquitos, location in Chicago, time lagged number of mosquitos, wind speed/ direction and temperature are key predictors to West Nile Virus
- XGBoost with parameters of {learning_rate': 0.01, 'max_depth': 5, 'min_child_weight': 10, 'n_estimators': 500} gives the best AUC (0.86) and cross-validation scores among all models

Understanding West Nile Virus in Chicago

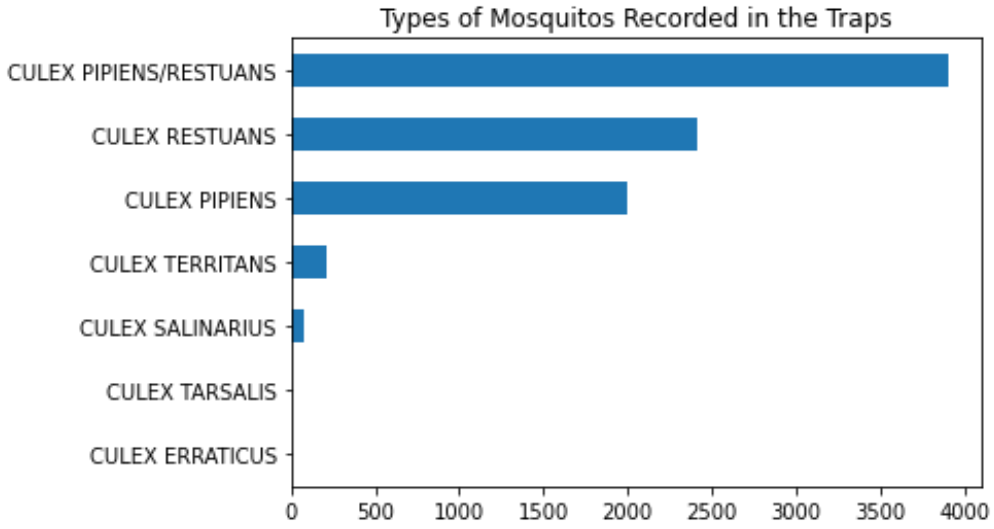


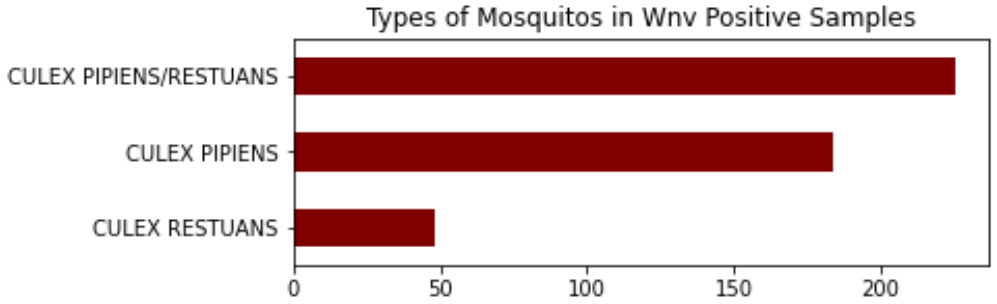


The Mosquito Problem in Chicago

- Mosquito traps were placed in 136 locations in Chicago, with data recorded between the months between May and October
- On selective years between 2007 to 2013, on a per annum basis, there are between ~16,000 to 33,000 mosquitoes recorded in the city
- Mosquitos are most prevalent starting June and generally decreases in amount after September
- Mosquito count usually peaks between July and August

Understanding West Nile Virus in Chicago





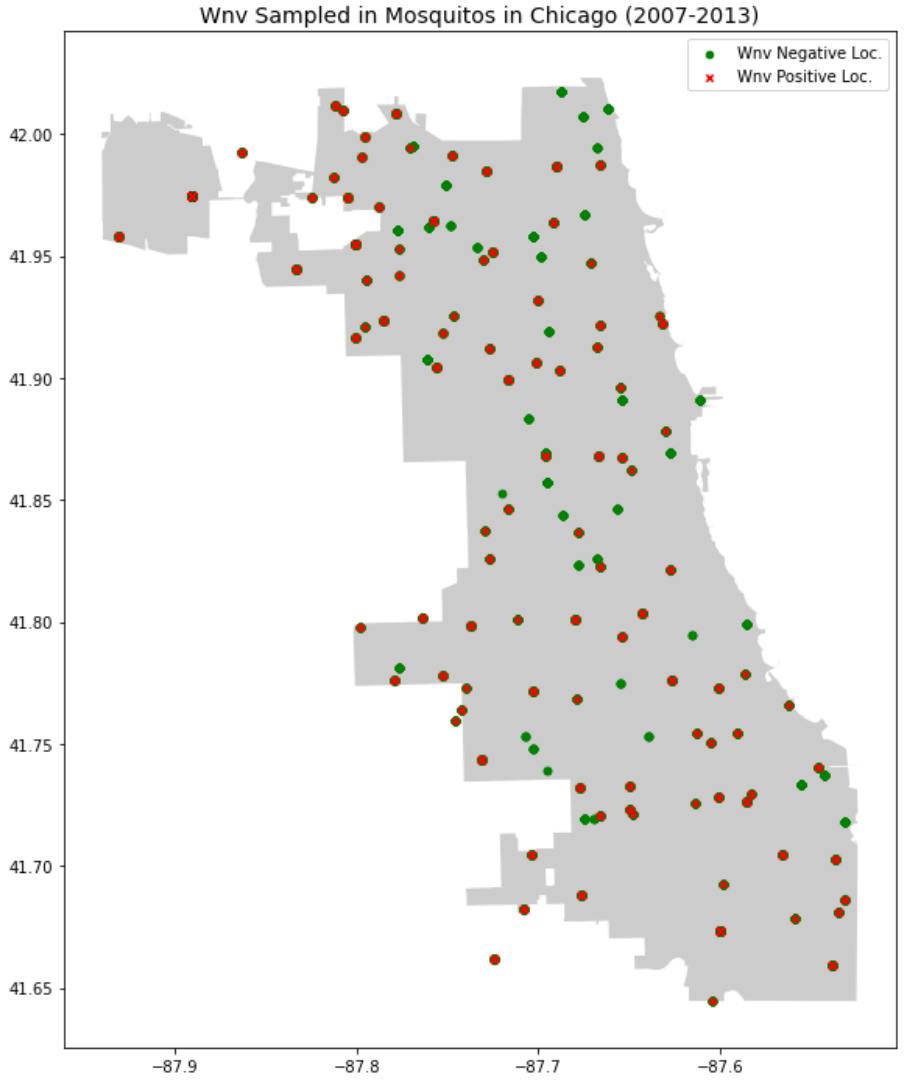
The breeds causing West Nile Virus (Wnv)

- Among traps and across years, 7 different breeds of mosquitos are being identified in Chicago
- However, only Culex Pippens and Culex Restuans were carriers of Wnv

Fig. Culex Pippens

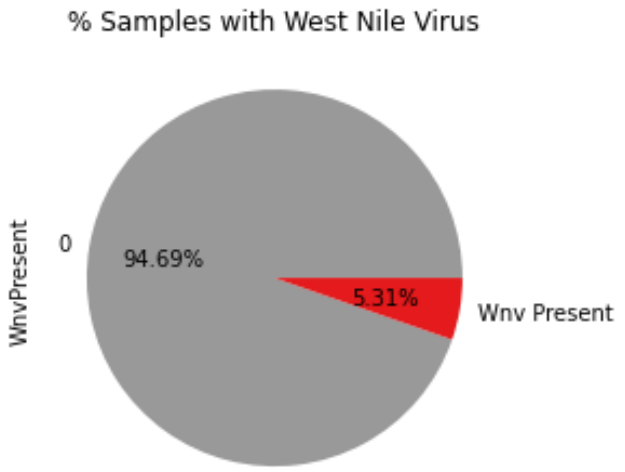
Fig. Culex Restuans

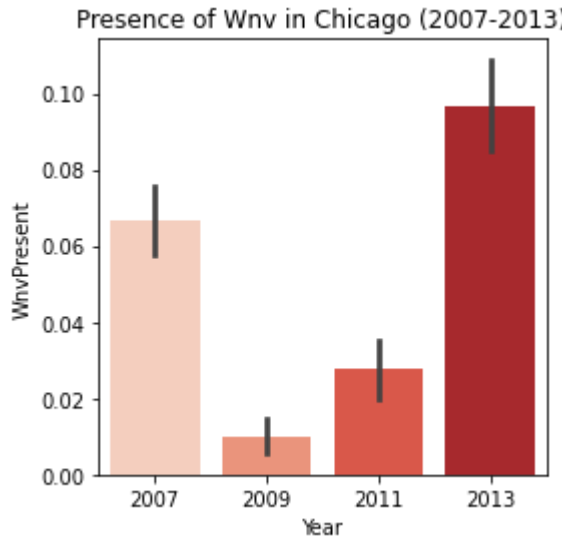
Understanding West Nile Virus in Chicago



Occurrences of Wnv in Chicago

- Map on the left detailed mosquito traps across Chicago, with segregation on trap locations where Wnv is found
- In all the mosquito samples collected across years, 5.31% of the sample is identified as Wnv positive. This also indicates the data is very imbalanced
- The overall trend on presence of Wnv is similar with the prevalence of mosquitos across the same years in Chicago. Years with more mosquitos found, Wnv is also more prevail





Pre-Processing & Features Engineering

Issues addressed in pre-processing:

01

HOT ENCODING

- One Hot-Coded mosquito breeds from categorical data to numeric features
- Encoded mosquito traps into unique numeric features

02

UNDERSAMPLING

- Given only 5.31% of the data is Wnv positive, the data is very imbalanced and requires re-sampling to be balanced
- Mosquitos not contributing to Wnv were first removed. 30% of the remaining records with no Wnv were being randomly sampled
- After under-sampling, Wnv present samples improved to 16%+

New Features Created:

DAYLIGHT DURATION

- Transforming sunrise and sunset timings into useful features by converting the two features into daylight duration

ADDITIONAL WEATHER FEATURES

- MaxTemp & Precipitation
- MaxTemp & Pressure
- Precipitation, Wind Speed & Wind Direction
- Relative Humidity (calculated using Dew Point and average temperature)

10 DAYS TIME LAGGED FEATURES

For these features:

- MaxTemp & Precipitation
- Wind Speed & Direction
- Number of Mosquitos collected in Trap

Pre-Processing & Features Engineering

Features Selection

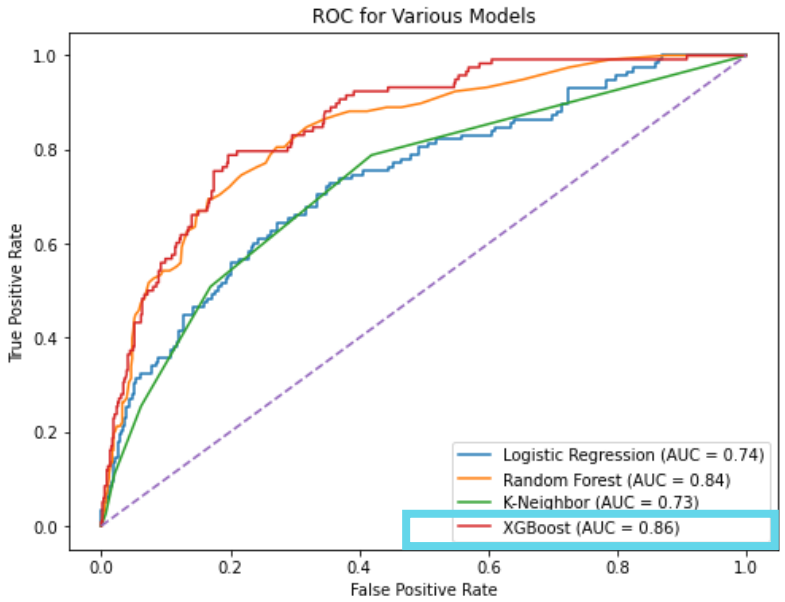
- Variance Inflation Factor (VIF) is used to test for collinearity among features.
- Features that were less important were being removed in the process
- Final top 6 features were being selected for modelling

	VIFactor	features
0	2.537111	Trap
6	2.040188	Speed_dir +10
5	1.472204	Species_CULEX RESTUANS
4	1.356172	Species_CULEX PIPIENS
1	1.328338	NumMosquitos
7	1.304704	NumMosquitos +10
3	1.204422	PrecipTotal
2	1.108977	Heat

Modelling & Parameters Tuning

Models Examined

1. Logistic Regression
2. Random Forest
3. K-Nearest Neighbor
4. XGBoost Classifier



Initial Results

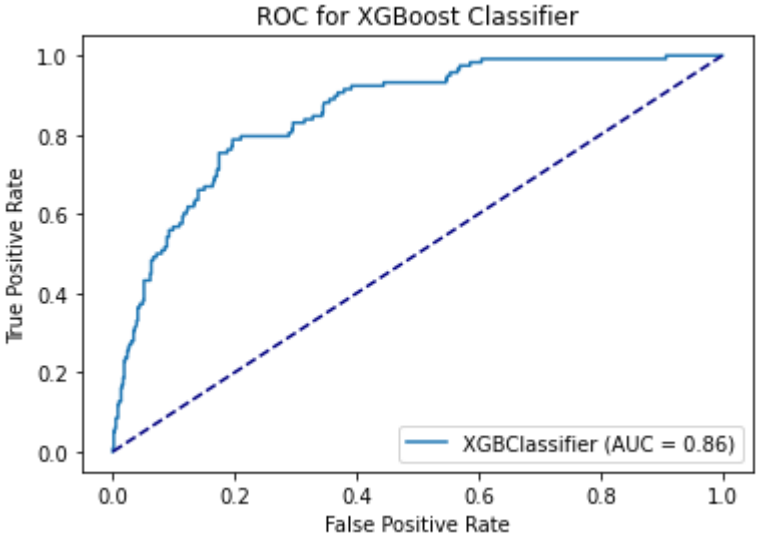
	Metrics/ Model	Logistic Regression	Random Forest	K-Nearest Neighbor	XGBoost
0	Accuracy Score (Test data)	0.835	0.846	0.824	0.858
1	ROC Score	0.519	0.634	0.596	0.664
2	F1 Score	0.079	0.407	0.326	0.468
3	Precision Score	0.625	0.578	0.455	0.629
4	Recall	0.042	0.314	0.254	0.373

- Initial modelling results are suggesting XGBoost is producing the best result (AUC score) among all models being attempted
- XGBoost will be selected as the model for further parameters tuning for better optimized results

Modelling & Parameters Tuning

Grid Search for optimized parameters

- Grid Search result suggests below are the best XGBoost parameters for the model
- {learning_rate': 0.01, 'max_depth': 5, 'min_child_weight': 10, 'n_estimators': 500}

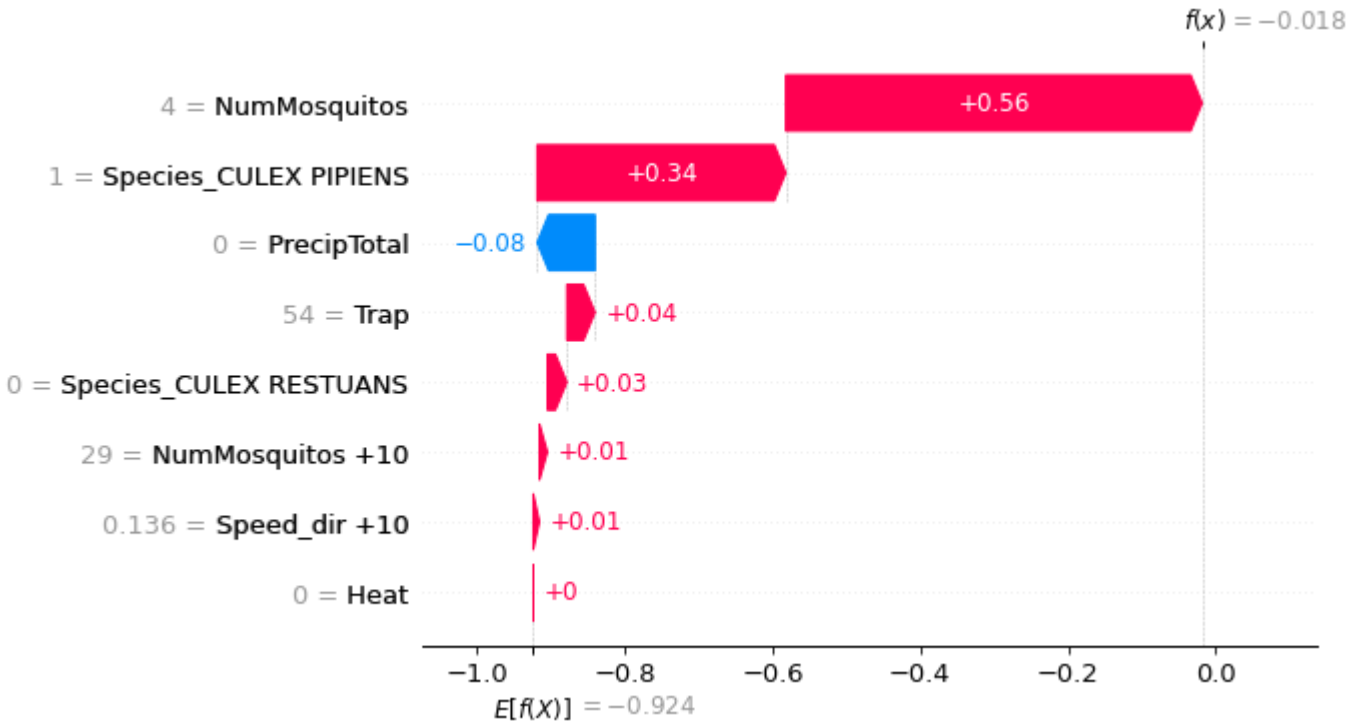


Re-Running the Model with Optimized Parameters

```
Score - XGBoost Classifier (optimized):
On Testing data: 0.858
On ROC Score: 0.664
On F1 score: 0.468
On Precision score: 0.627
On Recall score: 0.373
```

- 5 fold cross validation has been performed to this model with revised parameters
- Accuracy of the model with cross-validation is at 83.53 %

Interpreting the Results



Key Contributors to Wnv

- SHAP analysis was performed to the final model, to understand the importance of certain features in the model prediction
- From the waterfall chart on the left, with the exception of total precipitation, all other variables contributed positively to the model's prediction
- Numbers of mosquitos in the are and the existance of Culex Pipens mosquitos are the most important features contributing to the prediction
- Though not significant from a scoring perspective, the location, time lagged factors in number of mosquitos, wind speed/ direction and temperature remains key predictors to Wnv prediction

Model Limitations/ Future Phases

Limitations/ Other Considerations

The model was built based on data collected in 2007, 2009, 2011, 2013. Changes in environment may affect the model's predictive power.

Future Scope

Possible areas for further studies

- Test the model on other years in Chicago to further improve the model's predictability
- Use the spray data to evaluate the cost effectiveness of eradication of West Nile Virus
- Perform cost and benefit analysis of spraying to Chicago