

01	MOTIVATION	What do we want to solve?
02	EXPLORATORY DATA ANALYSIS	What does the data look like?
03	FEATURE ENGINEERING	What else can we cook?
04	MODELING	Which model makes sense?
05	CONCLUSION	Are we successful in predicting?

01 MOTIVATION

What drives the increase in West Nile Virus?

Are mitigating efforts (spraying) effective?

 Can we predict the prevalence of West Nile Virus based on our data so that health authorities can act preemptively to save lives? 02 EDA DATASET

Weather Data:

Contains daily weather info from two weather station at two airport in Chicago area

Train Data

 Nearly 10K rows with information on location, num and type of mosquitos and prevalence of WNV

Test Data

Nearly 100K rows of data but does not have WNV info so for this project we will ignore this data

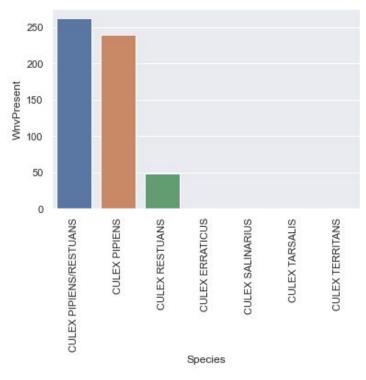
Spray Data

Contains date, time and location of spray done to kill mosquitos in the city

02 EDA

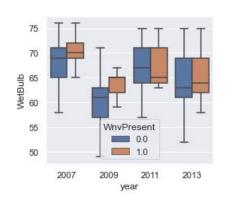
4000 3000 2000 1000 0 CULEX PIPIENS/RESTUANS CULEX PIPIENS CULEX TERRITANS CULEX TARSALIS CULEX ERRATICUS Mosquito Species Count

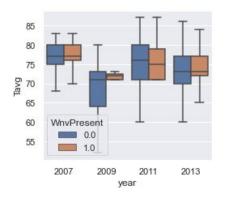
MOSQUITO TYPES

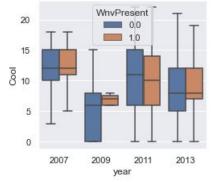


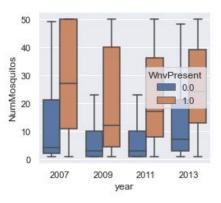
Mosquito Species Vs WNV

WHAT CONTRIBUTES TO WNV?









02 EDA

SOME CORRELATIONS

- 0.2

- 0.0

-- 0.2

--0.4

NumMosquitos	1	0.2	0.15	0.19	0.18	0.16	0.18	-0.13	0.18	-0.0053	-0.031	-0.048	-0.0033	-0.029	0.019
WhyPresent		- 1	0.048	0.065	0.06	0.085	0.08	-0.053	0.053	0.021	-0.0046	0.0093	-0.045	-0.0012	-0.032
Tmax		0.048	1	0.79	0.95	0.74	0.87	-0.73	0.89	0.043	-0.27	-0.33	-0.056		-0.025
Tmin		0.065	0.79		0.94	0.88	0.93	-0.68	0.91	0.17	-0.32	-0.43	0.017	0.082	0.11
Tavg		0.06	0.95	0.94		0.85	0.94	-0.75	0.95	0.11	-0.31	-0.4	-0.025		0.039
DewPoint		0.085	0.74	0.88	0.85		0.97	-0.6	0.83		-0.44	-0.5	-0.05	0.077	0.071
WetBulb		0.08	0.87	0.93	0.94	0.97		-0.69	0.91	0.25	-0.39	-0.46	-0.06	0,1	0.035
Heat	-0.13	-0.053	-0.73	-0.68	-0.75	-0.6	-0.69	1	-0.51	-0.012			0.034	-0.11	0.03
Cool		0.053	0.89	0.91	0.95	0.83	0.91	-0.51	1	0.14	-0.31	-0.4	-0.016	0.11	0.064
PrecipTotal	-0.0053	0.021	0.043		0.11			-0.012		1	-0.33	-0.29		-0.014	
StnPressure	-0.031	-0.0046	-0.27	-0.32	-0.31	-0.44	-0.39		-0.31	-0.33	1	0.95	-0.31	-0.26	-0.37
SeaLevel	-0.048	0.0093	-0.33	-0.43	-0.4	-0.5	-0.46		-0.4	-0.29	0.95		-0.28	-0.26	-0.37
ResultSpeed	-0.0033	-0.045	-0.056	0.017	-0.025	-0.05	-0.06	0.034	-0.016		-0.31	-0.28	1	0.23	0.9
ResultDir	-0.029	-0.0012		0.082		0.077	0.1	-0.11	0.11	-0.014	-0.26	-0.26	0.23	1	
AvgSpeed	0.019	-0.032	-0.025	0.11	0.039	0.071	0.035	0.03	0.064		-0.37	-0.37	0.9	0.31	1
	NumMosquitos	WhyPresent	Ттах	Tmin	Tavg	DewPoint	WetBulb	Heat	Cool	Precip Total	SinPressure	SeaLevel	ResultSpeed	ResultDir	AvgSpeed

Merged train data to the weather data from the closest weather station.

• Given the life cycle of mosquito, temperature and precipitation can have time delayed impact on the presence of WNV so created time lag data on these variables at 1,2, 7 and 14 days.

 Calculated correlation with target feature and it indicates lag data has higher correlation.

- Checked for multicollinearity using
 - Using VIF (Variance Inflation factor)

- Ignored columns that had VIF score of greater than 15
 - Guidebook said ignore above 10 but that would make my feature list too small.

- # of Total features
 - 0 17

- The data had target feature heavily skewed (19:1)
- Balanced the data using resampling

04 MODELING

MODEL SELECTION

Considered Models and their AUC Score:

Logistic Regression: 0.7128

o Random Forest: 0.8324

Decision Tree: 0.7040

o KNN: 0.7203

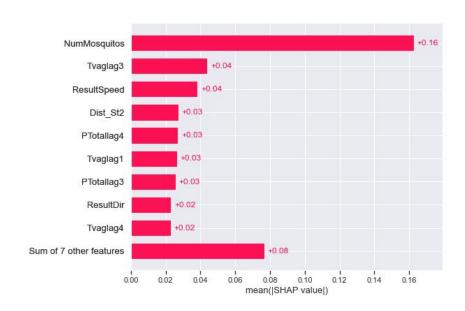
AdaBoost: 0.7358

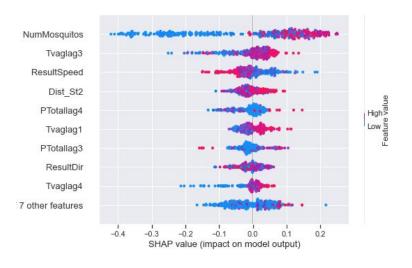
XGBoost: 0.7943

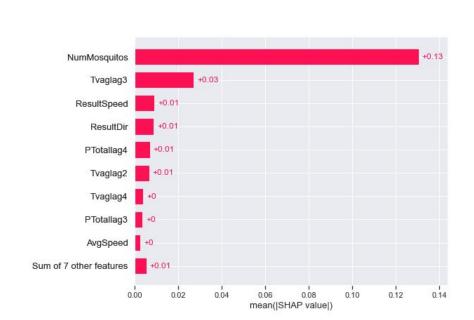
Picked two top model:

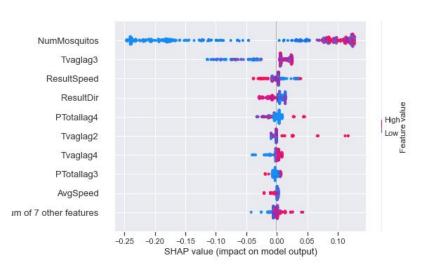
- Random Forest
- XGBoost

RANDOM FOREST









- Number of Mosquitoes is the biggest indicator of WNV
- Lag Data shows that WNV peaks:
 - A week after the hot temperature and
 - two weeks after precipitation
- The city should proactively spray few days after the rainfall or hot temperature.