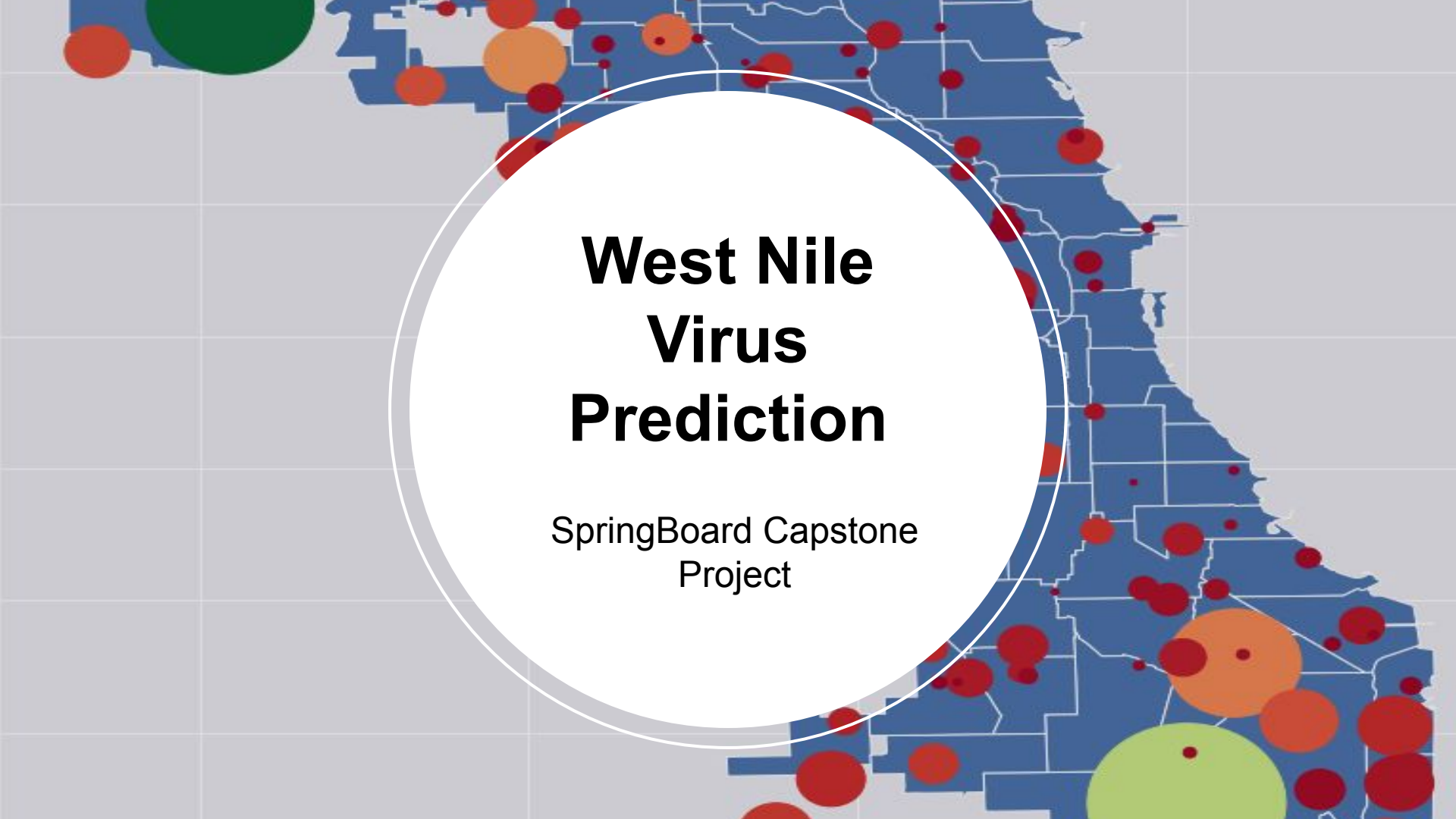


West Nile Virus Prediction

SpringBoard Capstone Project


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?

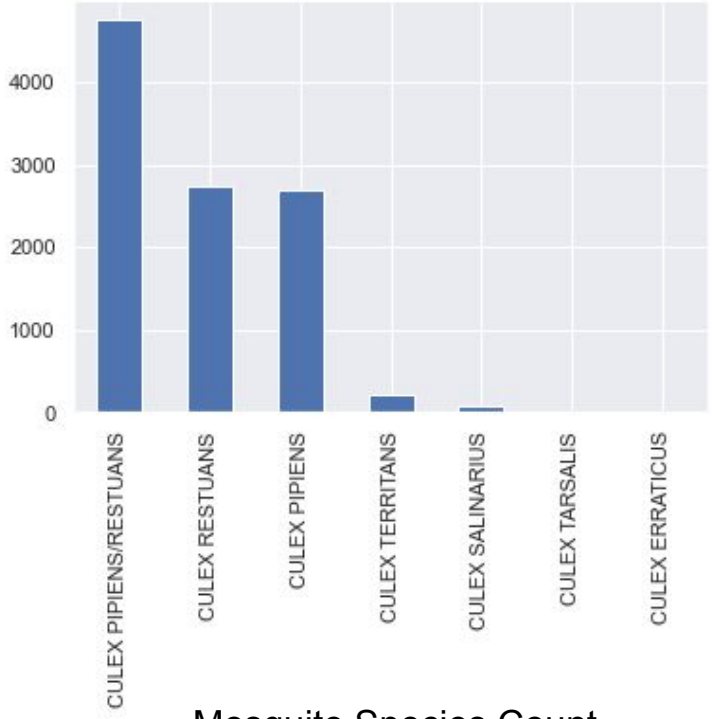
02 EDA

DATASET

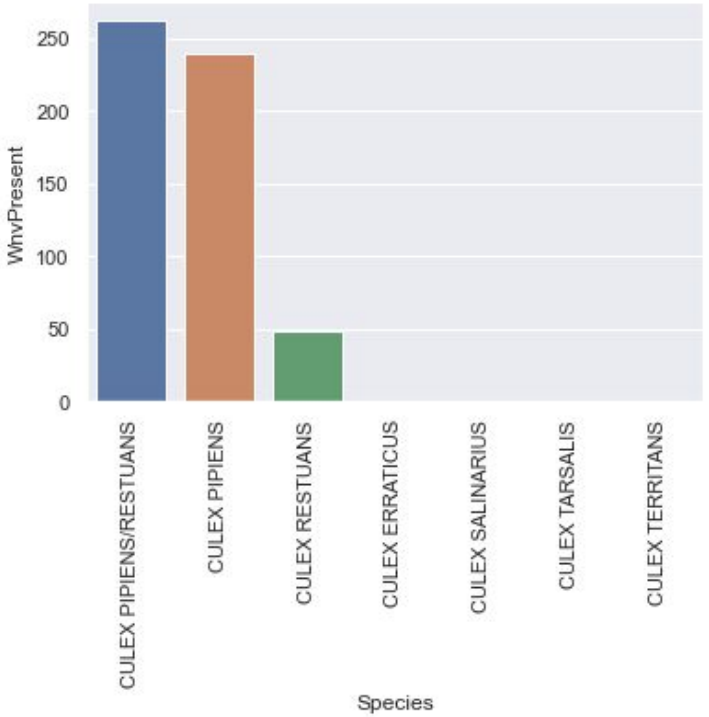
- Weather Data:
  - Contains daily weather info from two weather station at two airport in Chicago area
- Train Data
  - Nearly 10K rows with information on location, num and type of mosquitos and prevalence of WNV
- Test Data
  - Nearly 100K rows of data but does not have WNV info so for this project we will ignore this data
- Spray Data
  - Contains date, time and location of spray done to kill mosquitos in the city

02 EDA

MOSQUITO TYPES



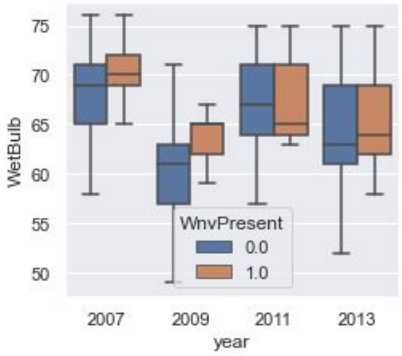
Mosquito Species Count

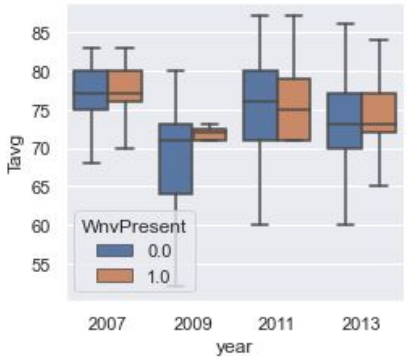


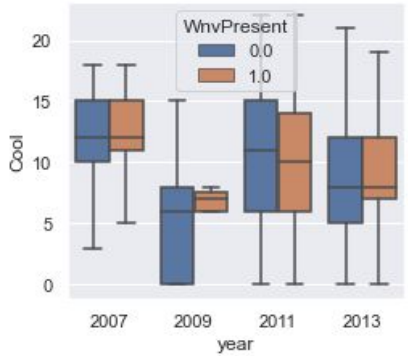
Mosquito Species Vs WNV

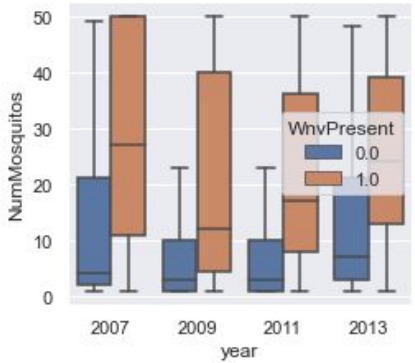
02 EDA

WHAT CONTRIBUTES TO WNV?



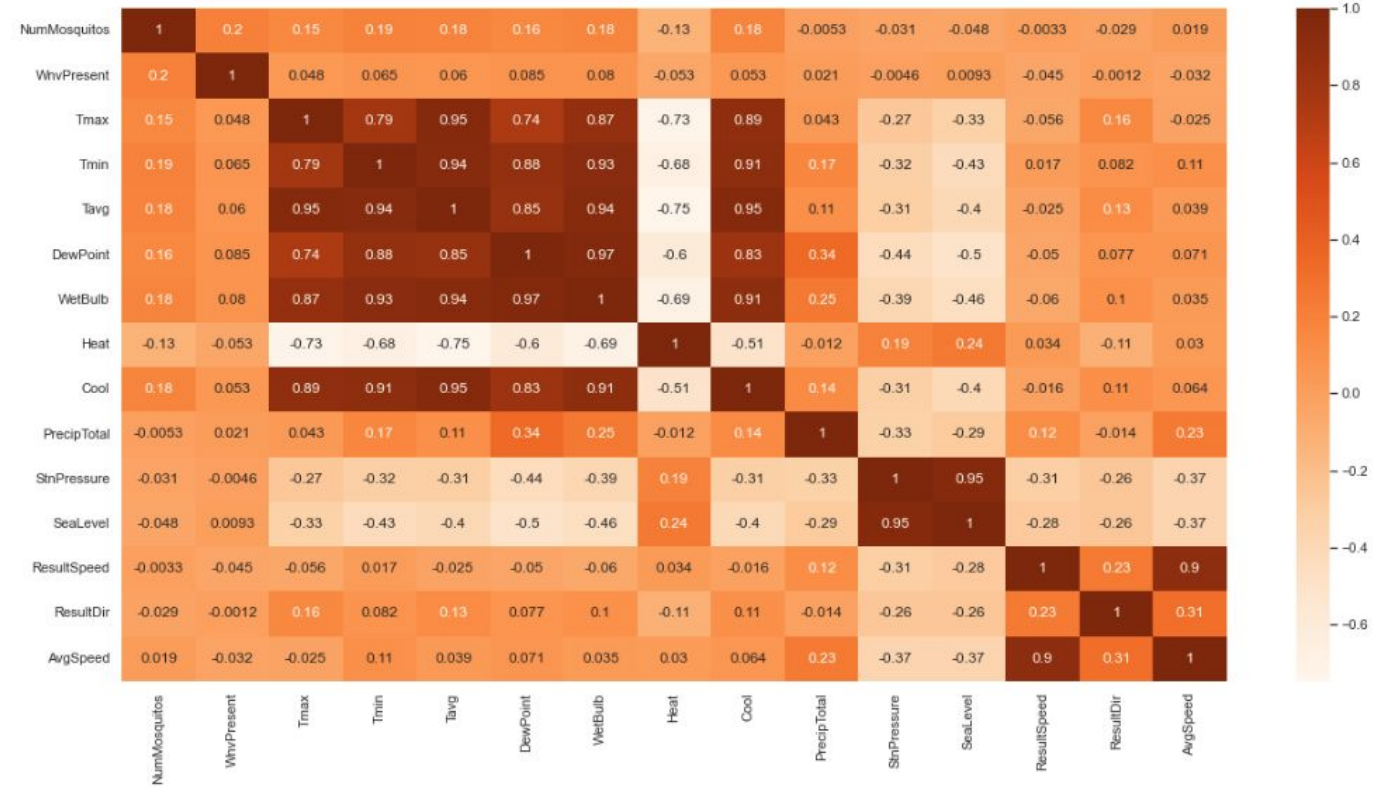






02 EDA

SOME CORRELATIONS



03 FEATURE ENGINEERING

TIME LAG DATA

- 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.

03 FEATURE ENGINEERING

MULTICOLLINEARITY

- 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
  - 17

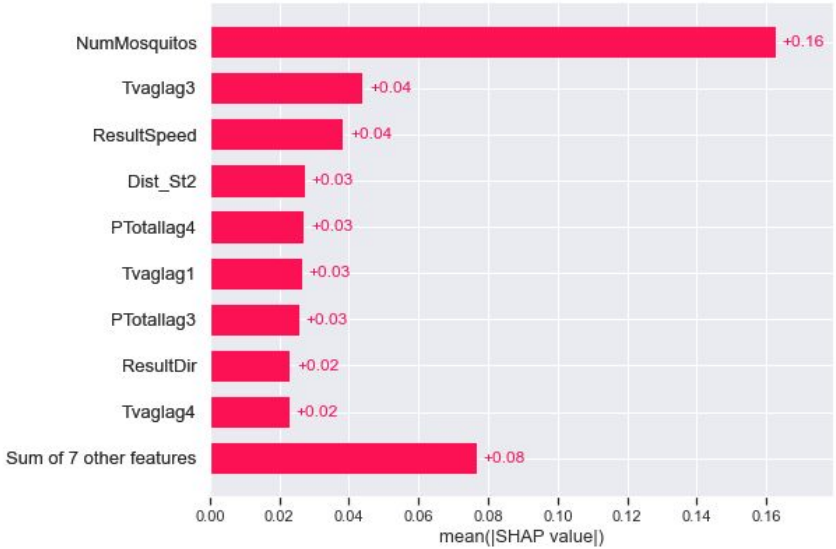
BALANCED DATA

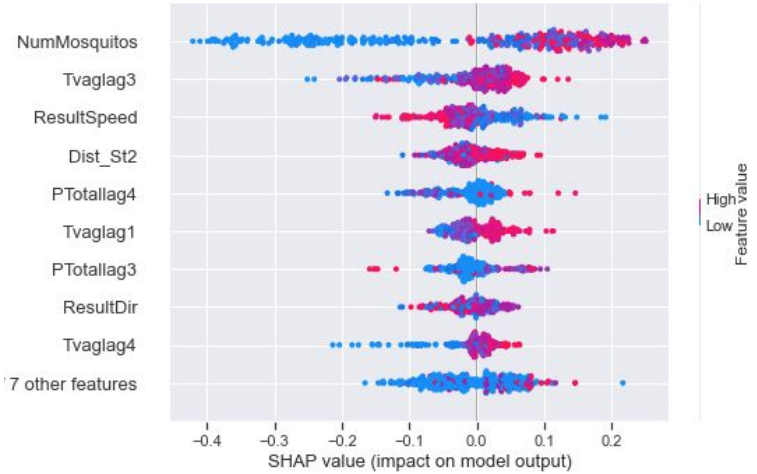
- The data had target feature heavily skewed (19:1)
- Balanced the data using resampling

- Considered Models and their AUC Score:
  - Logistic Regression: 0.7128
  - Random Forest: **0.8324**
  - Decision Tree: 0.7040
  - KNN: 0.7203
  - AdaBoost: 0.7358
  - XGBoost: 0.7943
- Picked two top model:
  - Random Forest
  - XGBoost

04 MODELING

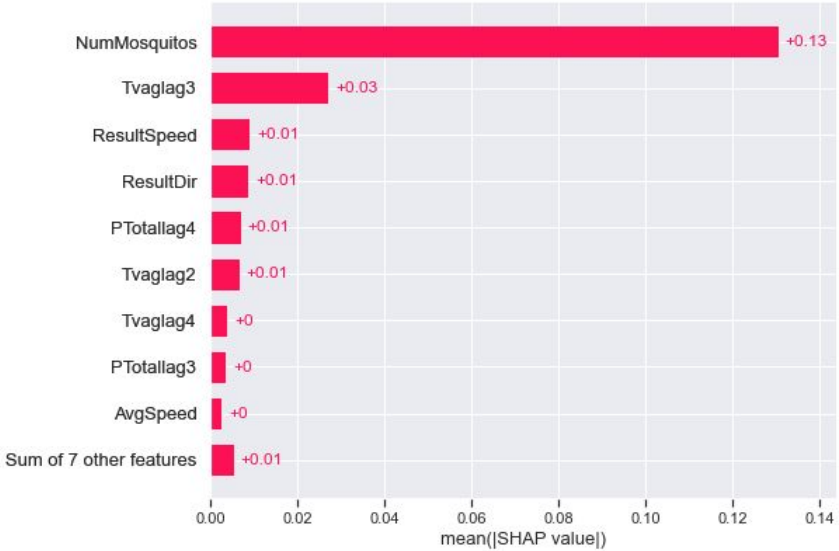
RANDOM FOREST

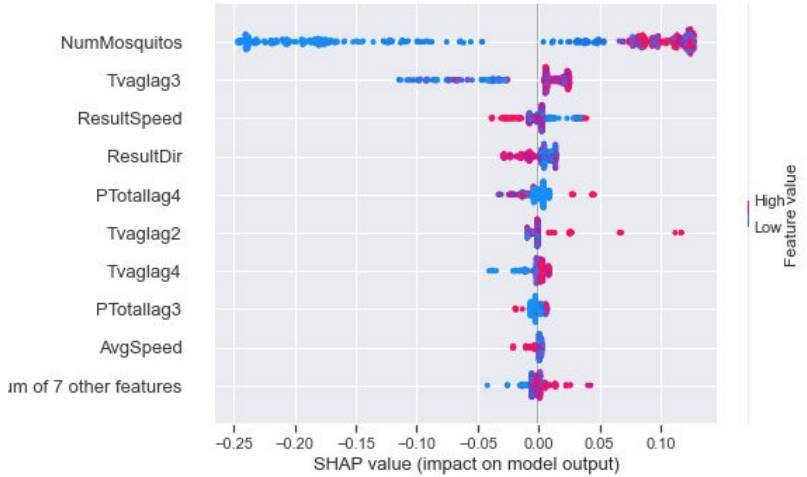




04 MODELING

XGBOOST





05 CONCLUSION

OBSERVATION & RECOMMENDATION

- 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.