West Nile Virus Prediction

Dataya Pte Limited

Meeting Participants

Officials from the Department of Public Health

Team from Dataya Pte Ltd

Agenda

1

Introduction

Background,
& Data Cleaning

2

EDA

Exploratory data Analysis

3

Modeling

Model Evaluation

4

Conclusion

CBA & Recommendations

Introduction

Background

- West Nile virus (WNV) is the leading cause of mosquito-borne disease in the continental United States.
- Most commonly spread to people by the bite of an infected mosquito.
- On average, 2,622 people gets infected every year and there are 122 deaths due to WNV every year.
- Every week from late spring through the fall, mosquitos in traps across the city are tested for the virus.
- Results of these tests influence when and where the city will spray airborne pesticides to control adult mosquito populations.

Problem Statement

- The Department of Public Health has engaged us, an independent Data Science company, to derive an effective plan using data science methods to deploy pesticides across Chicago.

- We would have to make recommendations on where pesticides should be sprayed and the cost vs benefit of deploying these pesticides.

Data Cleaning

Data Cleaning

- 3 different datasets: Train, ‘Spray’, ‘Weather’

Dataset	Test	Spray	Weather
No. of Columns	12	4	22
No. of Rows	10,506	14,835	2,944
No. Null values	0	584	10,689

- Impute all the missing values in the dataset
- Remove duplicates

EDA

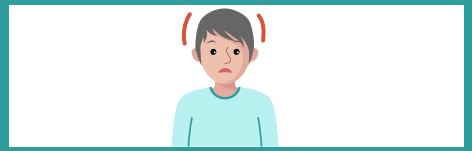
EDA : Mosquitoes Count

- 2 / 6 species found containing WNV

EDA: Features

Dew Point

Average dew point in Degrees Fahrenheit.



Wet Bulb

Average wet bulb in Degrees Fahrenheit.

T-Average

Average temperature in Degrees Fahrenheit.

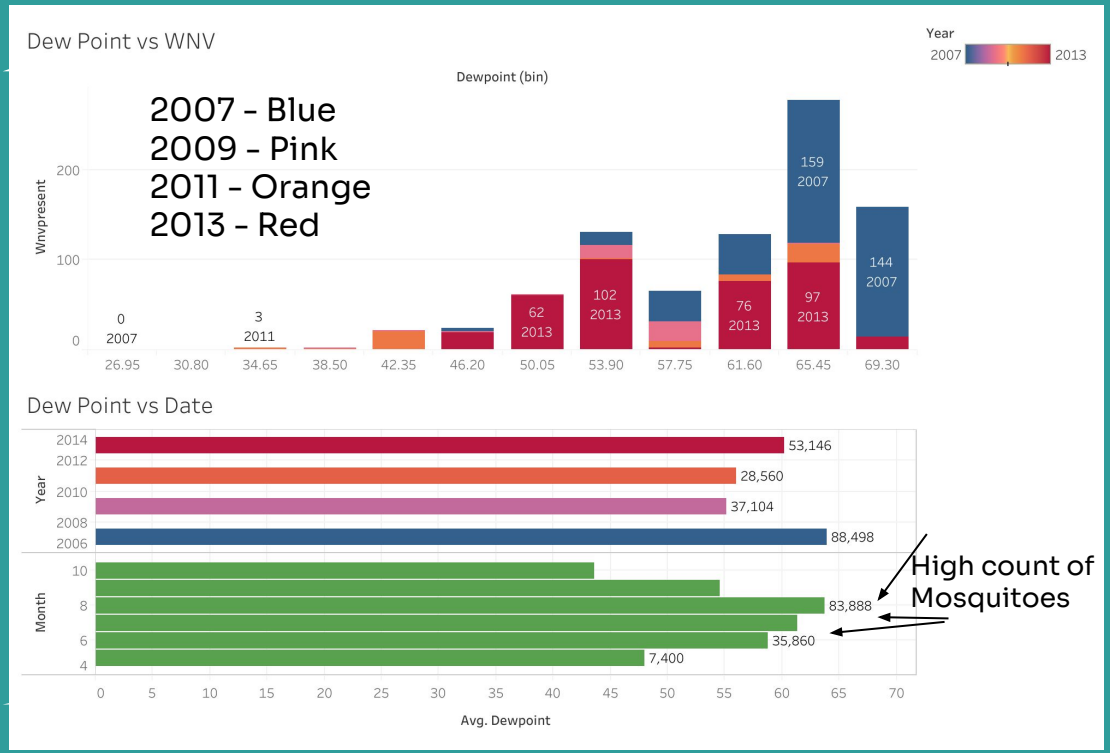
Wind

Average Speed Wind

Precipitation

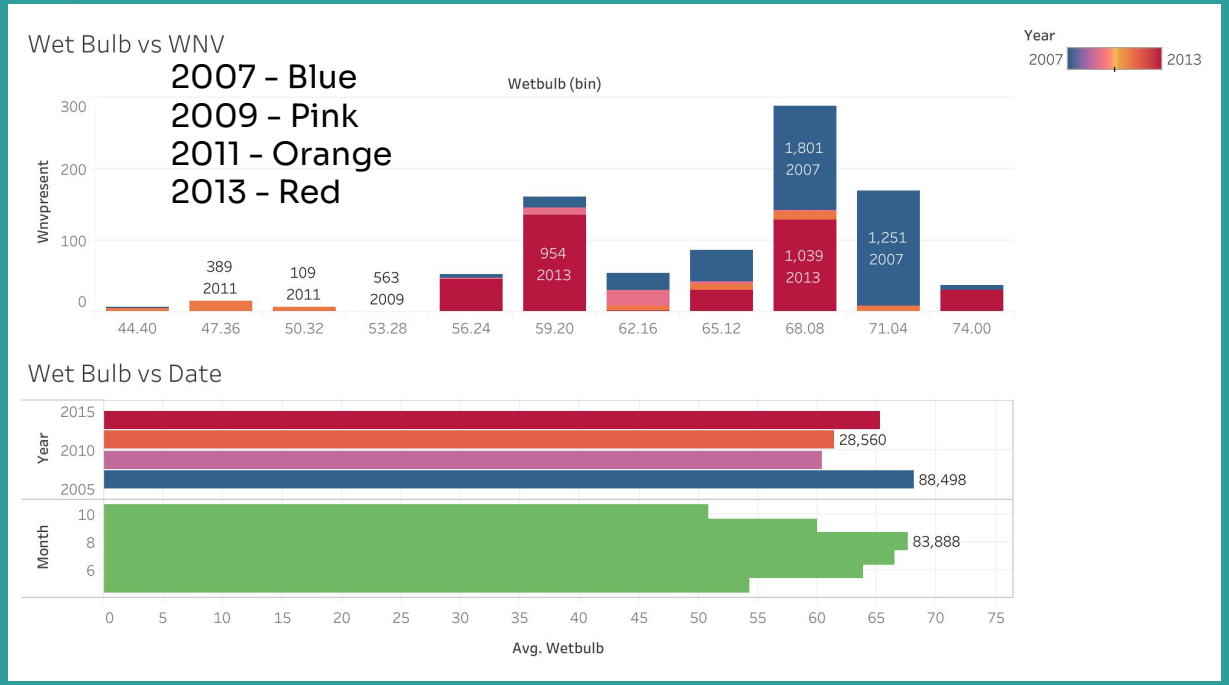
Water equivalent(Inches & Hundredths

EDA : Dew Point



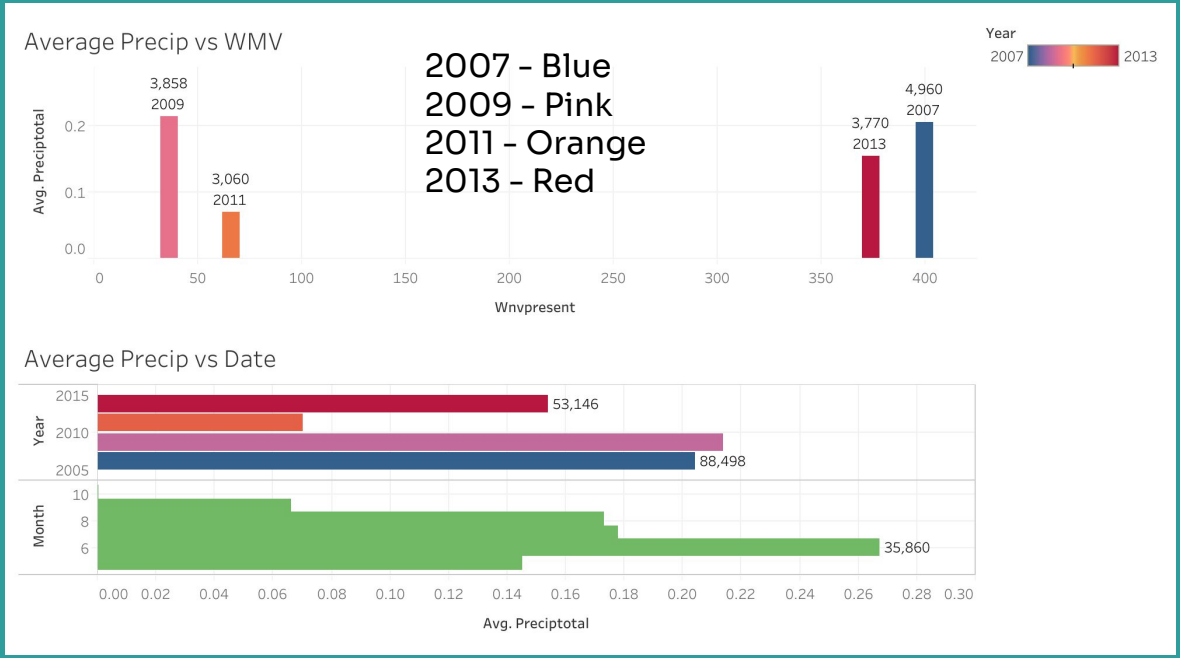
- Number of mosquitoes peak between June - Aug
- High dew points counts are correlated with WNV.

EDA : Wet Bulb



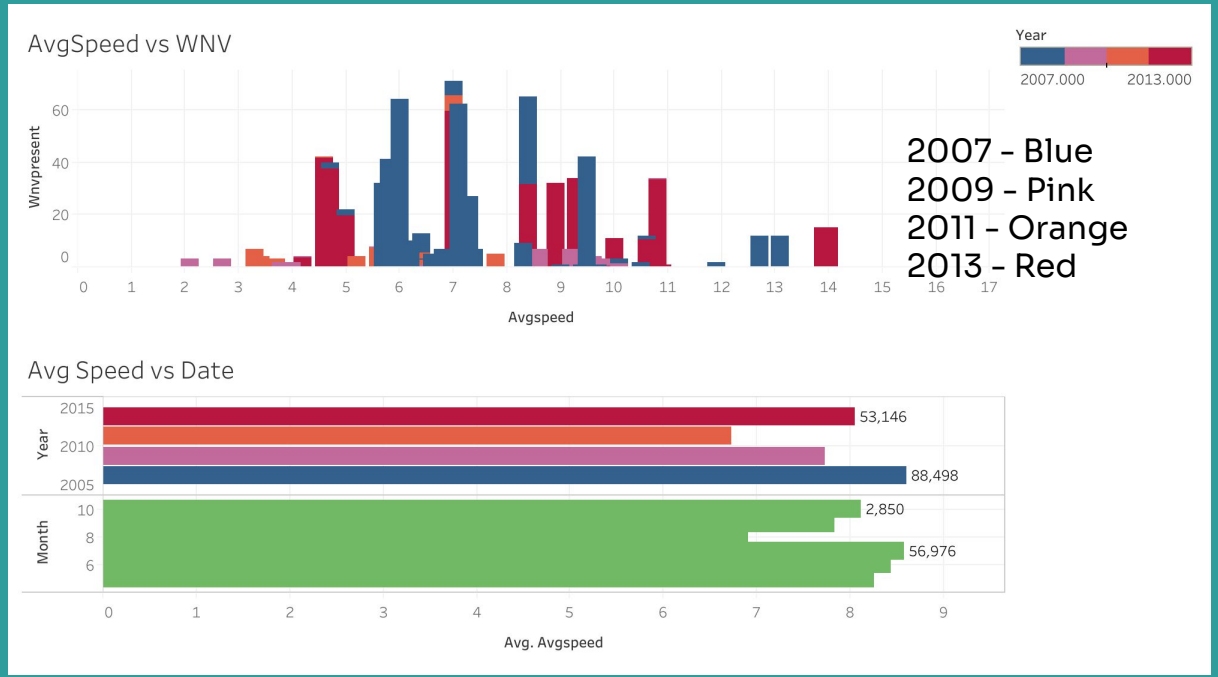
- Peak between July to Aug
- High humidity offsets higher temperatures
- Wet Bulb - humidity relationship

EDA : Precipitation



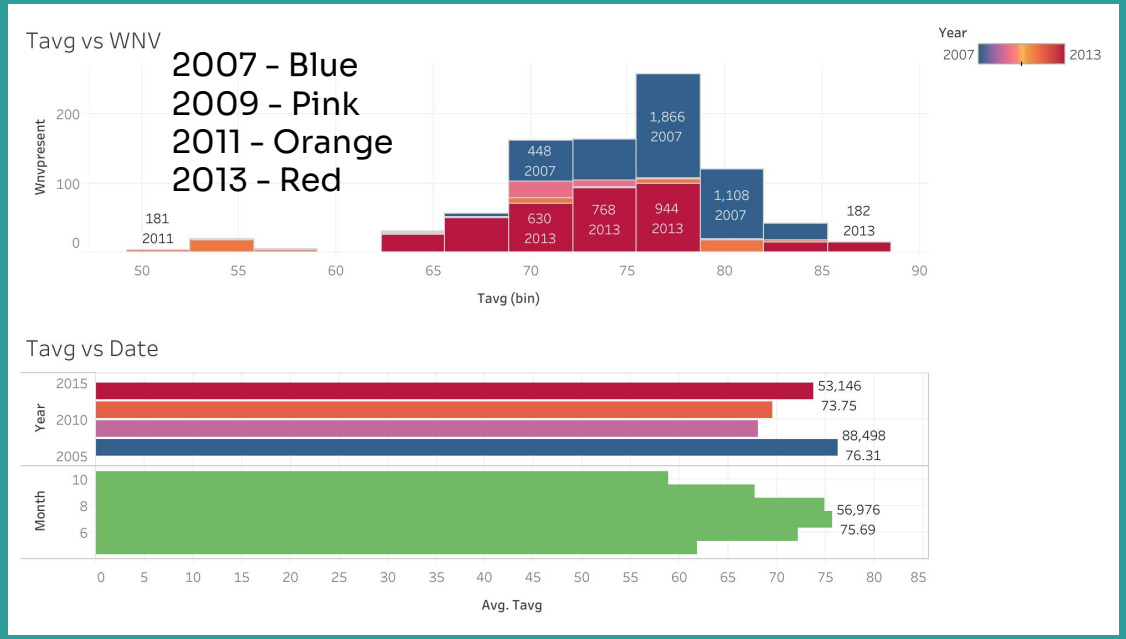
- Breeding is encouraged by precipitation.
- In June, we can clearly detect a rise in mosquito populations.

EDA : Wind Speed



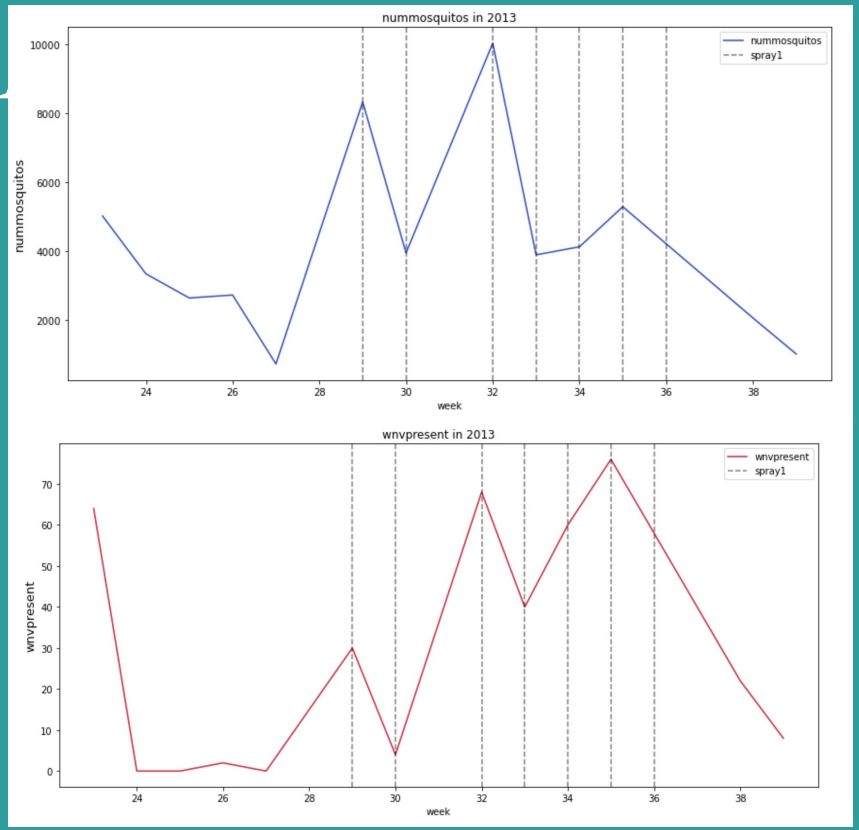
- Higher WNV numbers are indicated by slower wind speeds.
- The spray and trap's effectiveness may be impacted by wind speed.

EDA : T - Average



- An increase in temperature during July - September Period.
  - Positively correlated to Number of mosquitoes
- Impact of Temperature

EDA : Spray



- Spraying seems to reduce mosquito populations but not the spread of viruses.
- Potential causes for this could be:
  - Improper spray targeting
  - Incorrect timing of sprays

Featuring Engineering

Relative Humidity

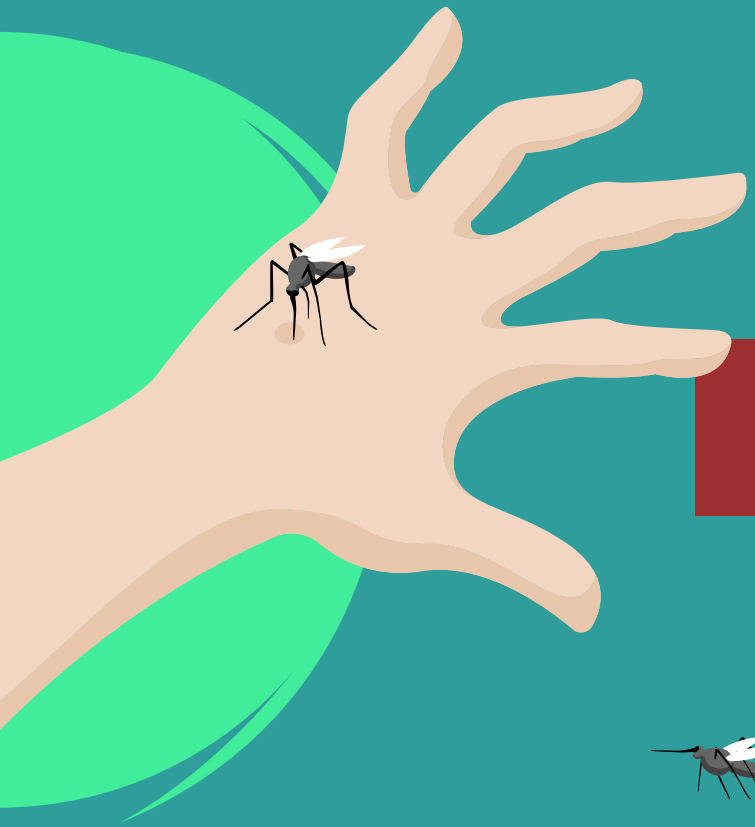
Average Weekly Temperature

Traps (One Hot)

Average Weekly Humidity

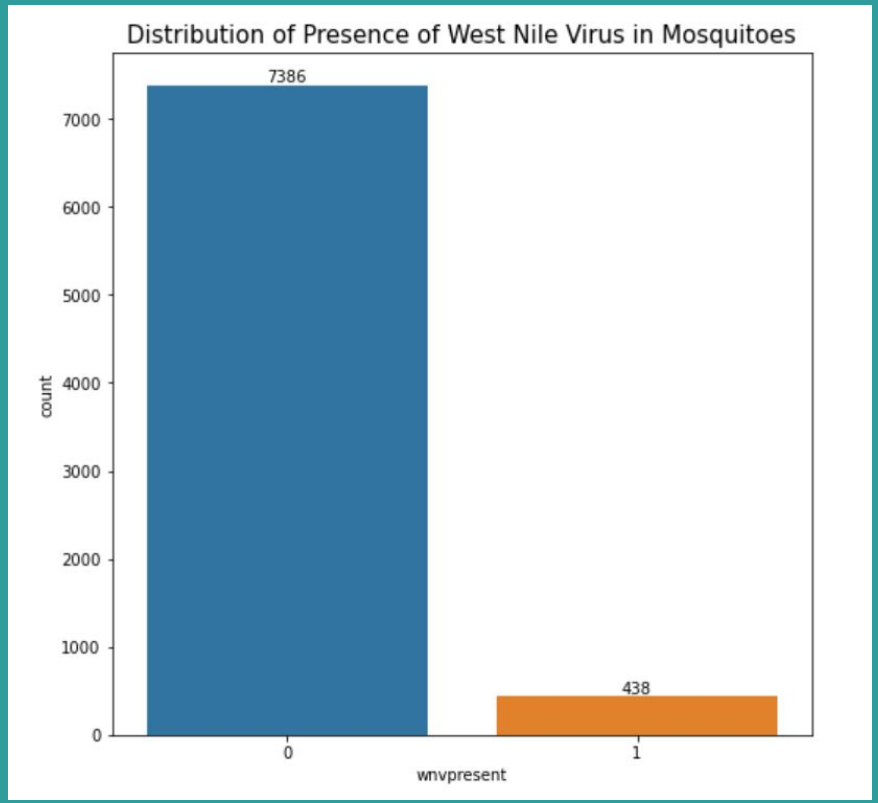
Species (Re-map)

Cumulative Weekly Precipitation



Modelling

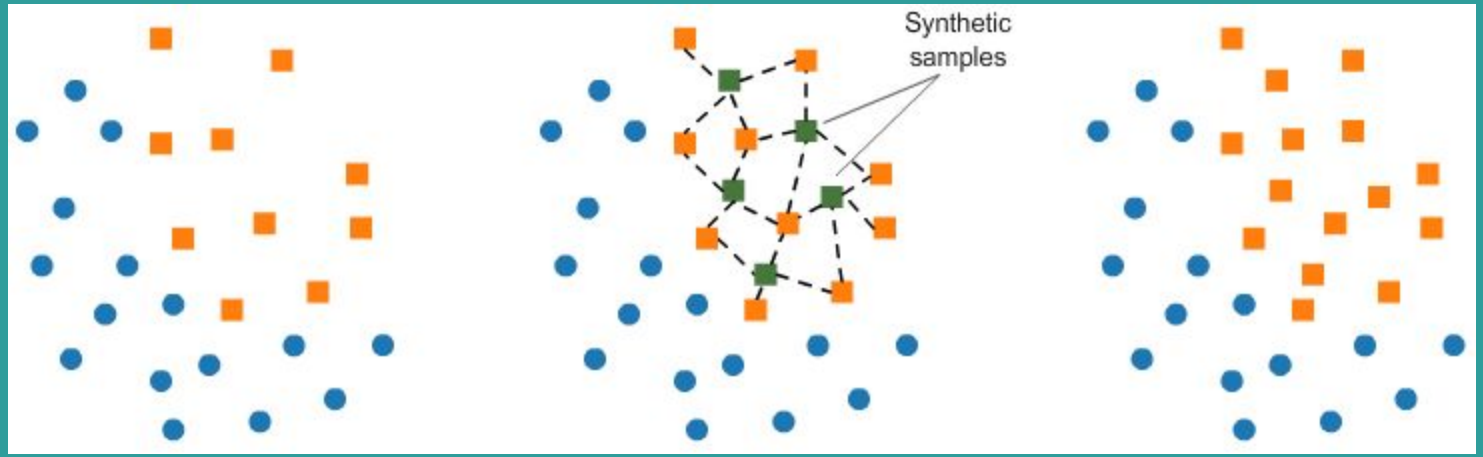
Highly Imbalanced Dataset



- 94.4% of the mosquitoes do not have West Nile Virus.
- Only 5.6% of the data shows the presence of the West Nile Virus.
- Highly imbalanced classes so we need to balance them out for better model performance.

SMOTE

Synthetic Minority Oversampling Technique (SMOTE) is an oversampling technique where the synthetic samples are generated for the minority class. This will help to balance our data in order to aid our machine learning algorithms.



Source:
https://medium.com/@asheshdas.ds/oversampling-to-remove-class-imbalance-using-smote-94d5648e7d35

Logistic Regression (Baseline Model)

After applying SMOTE on our dataset, we are able to achieve a better ROC AUC score.

Train Accuracy	0.66341
Test Accuracy	0.63781
ROC AUC Score	0.63781 (vs 0.5 before SMOTE)

GridSearchCV (Hyperparameter Tuning)

GridSearchCV is the process of performing hyperparameter tuning in order to determine the optimal values for a given model.

We will run this for our 5 models below and evaluate their performance:

- Logistic Regression
- K-Nearest Neighbors
- Random Forest
- Ada Boost
- Gradient Boosting

Performance Metrics

ROC-AUC score: The ROC curve is a plot of the True Positive Rate (Sensitivity) vs. the False Positive Rate for all possible decision thresholds.

The higher the AUC, the better the model is at distinguishing between whether there is WNV or not.

Sensitivity: Sensitivity is a measure of the proportion of actual positive cases that got predicted as positive (or true positive).

High sensitivity means there will be few False Negative results and thus fewer cases of WNV-positive missed out

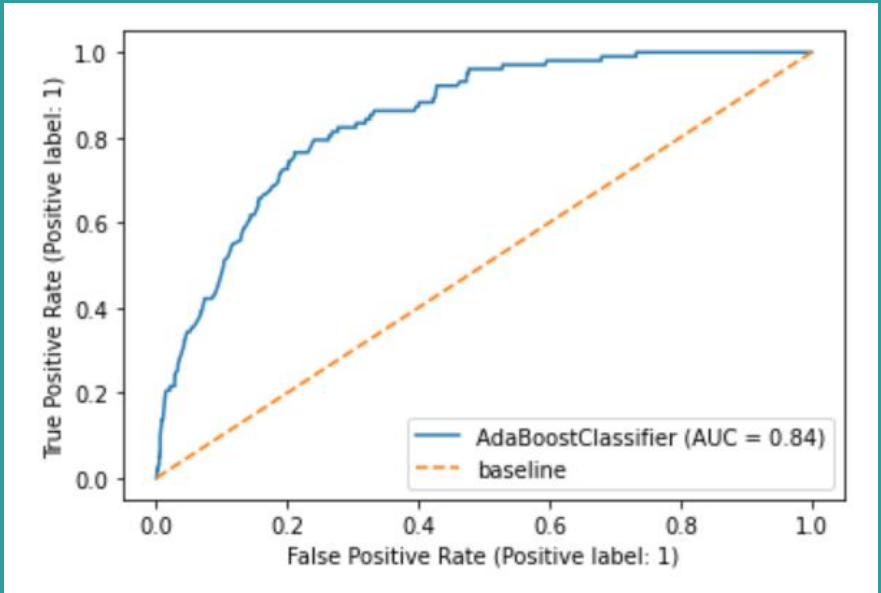
Model Evaluation

model	train_accuracy	test_accuracy	sensitivity	specificity	precision	f1_score	roc_auc_score
K-Nearest Neighbors	0.95074	0.84439	0.78263	0.90615	0.89292	0.83415	0.84439
Ada Boost Classifier	0.83704	0.83792	0.90669	0.76915	0.79706	0.84835	0.83792
Random Forest Classifier	0.94297	0.83279	0.82470	0.84088	0.83827	0.83143	0.83279
Logistic Regression	0.78507	0.79504	0.86462	0.72546	0.75900	0.80837	0.79504
Gradient Boosting Classifier	0.78651	0.77697	0.85491	0.69903	0.73962	0.79309	0.77697

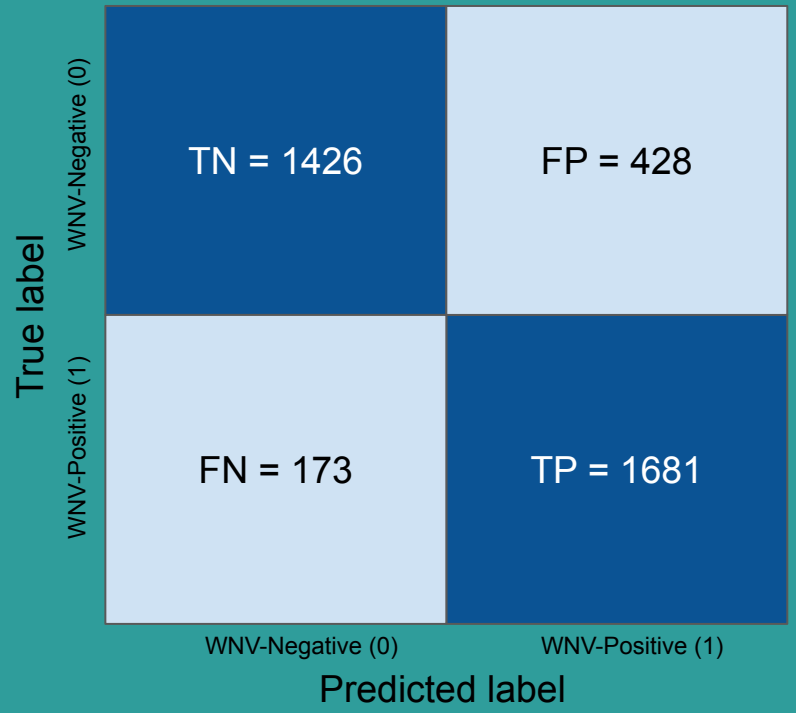
Ada Boost is the best model as it has high sensitivity and ROC AUC score.

Ada Boost

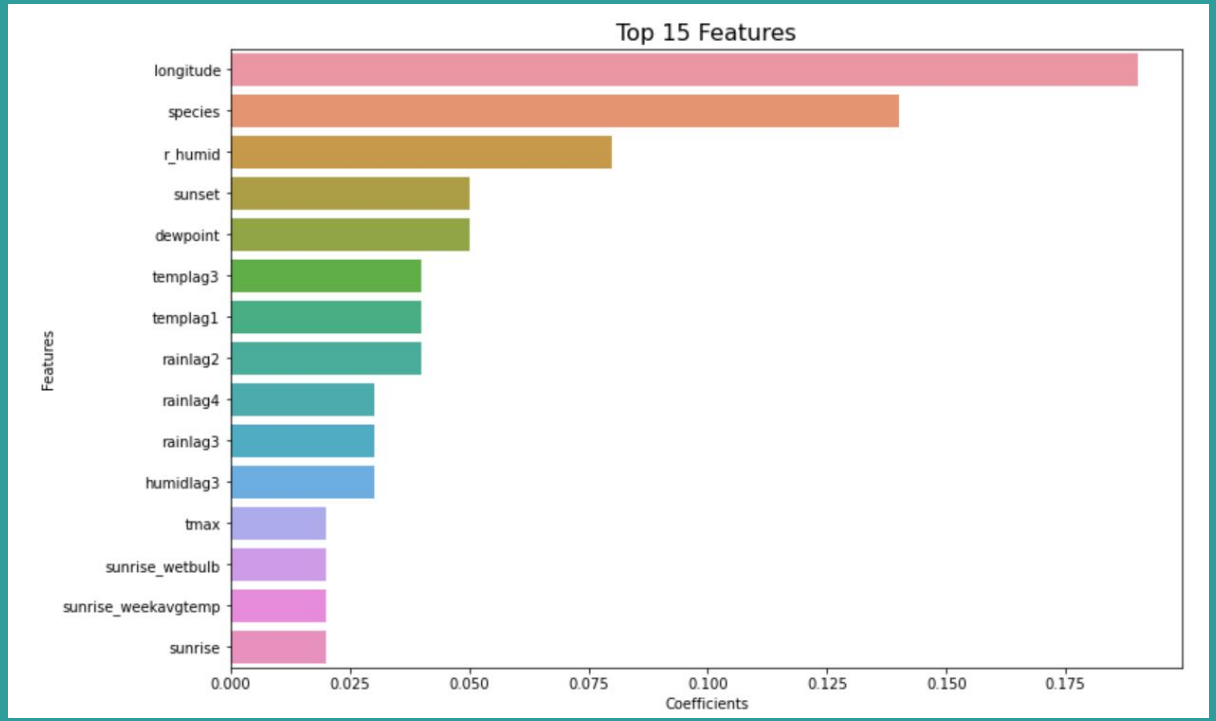
ROC Curve



Confusion Matrix



Top 15 Predictors



Limitations

- AdaBoost model is sensitive to noise and outliers
- Relatively slow and computationally expensive

Future Improvements

1. Use Principal Component Analysis to reduce dimensions and noise in the data
2. Apply more advanced models and compare results
    a. eg. Support Vector Machines, XGBoost
3. Further hyperparameter tuning

Cost-Benefit Analysis & Conclusion

Benefits - Overview

Difficult to quantify the results of spraying

- Mosquito Control is conducted in a preventive manner
  - Pre-emptive, not reactive!

Lack of control groups

- No identical control groups
- Unable to experiment due to fatal nature of virus

Benefits - Quantitative

Based on a study (West Nile Virus Outbreak) found in 2005:

Costs/Severity	West Nile Neuroinvasive Disease	West Nile Fever
Count	46	117
Inpatient Treatment	$2,188,768*	-
Outpatient Treatment	$884,137*	$28,051*
Productivity Loss	$696,009*	$145,727*
Average	$81,392*	$1,485*

** values adjusted for inflation @ 2.15% p.a*

Costs - Quantitative

Based on the same study,

- Aerial spray would cost $861,937* for an area of 477 km²

Assuming a same distribution of the virus severity, there is a net benefit as long as a spray of similar area (477 km²) prevents **36 cases or more**.

Benefits - Qualitative

1. Inevitably reduces other mosquitoes borne diseases
2. Higher well-being by those who consider mosquitoes pests
3. Regular spraying creates jobs for the state
4. Visible effort to residents of steps taken to prevent fatalities

Costs - Qualitative

1. Lower well-being by those who dislike the smell
2. Reports of pesticides causing health effects
    a. eg. itching, convulsions, swelling, asthma
3. Impairs vision
    a. Spraying often done in the morning

Key Takeaways

1. **2 out of** the **6** mosquito **species** constituted all observed presences of the WNV

2. The location's **longitude** is the variable with the **largest impact** on our model's predictions

3. Although the use of pesticides in the 2013 spraying data shows some effectiveness in controlling the virus,
    a. Unable to directly compare cause and effect but there are suitable measures to take

Recommendations

- Focus efforts on deploying pesticides in community areas with higher density since spraying costs proportional to area

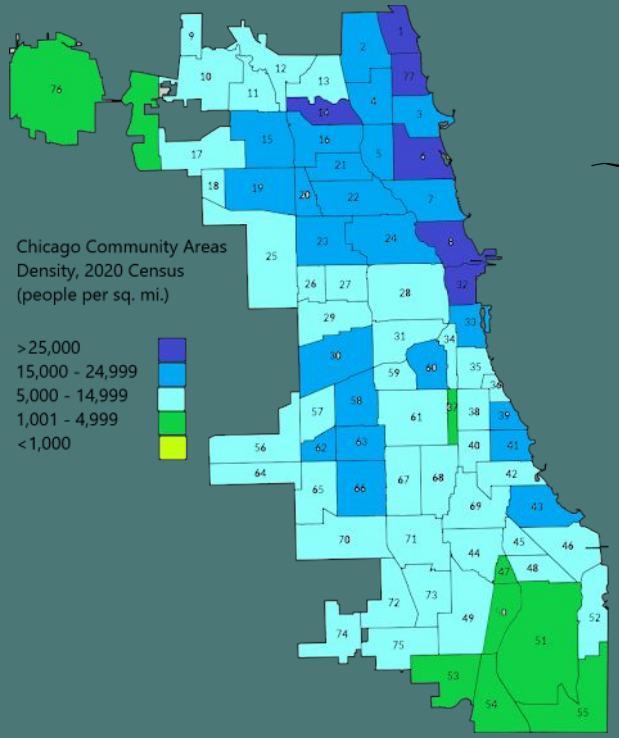


Image courtesy of: newgeography.com

Recommendations

- Run campaigns to educate public
  a. Best practices for prevention
  b. Most effective approach is a joint effort between the community and government
- Alert system
  a. Inform residents in areas when up-to-date data suggests a higher likelihood of the virus