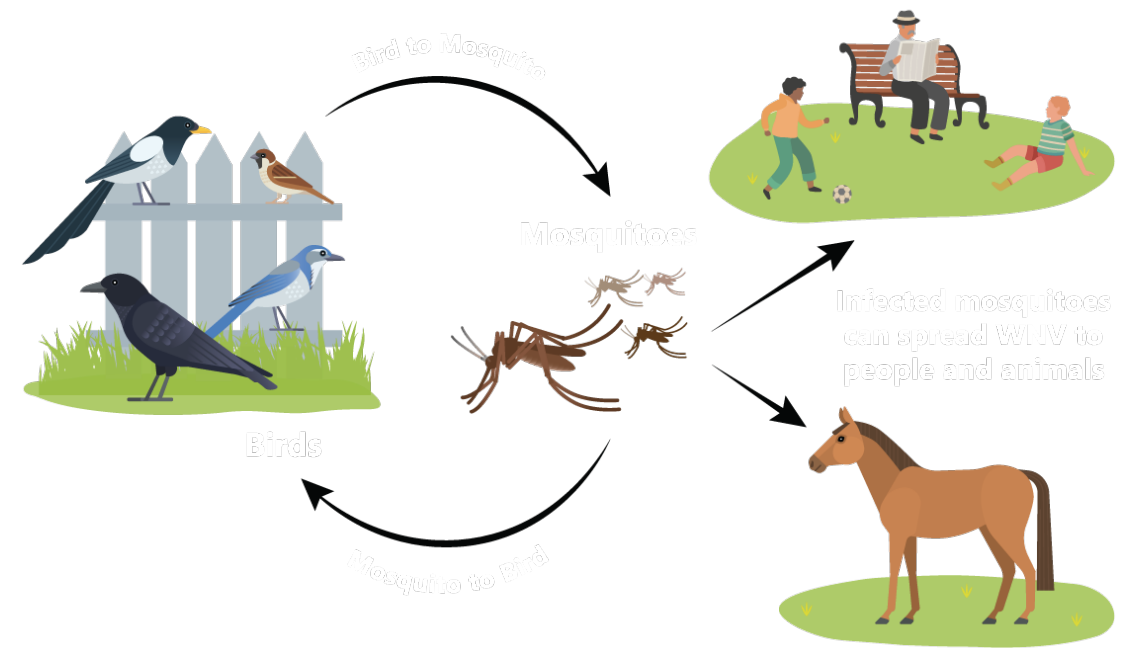# West Nile Virus Analysis in Chicago

Introduction

- West Nile Virus is an illness that spreads through mosquitoes and is prevalent in the United States

- Symptoms are minimal but some can be fatal, potentially resulting in death

- In 2004, The city of Chicago started collecting mosquitoes to asses the virus presence in their city and spraying pesticide to reduce it

Problem:

How can the city of Chicago predict which times of the year and at what temperature the West Nile virus is positive in a collections of mosquitoes?

Data:

Weather Data containing different weather measurements in the city of Chicago and GIS Data containing information about the mosquito collections provided by Chicago Department of Public Health (CDPH)
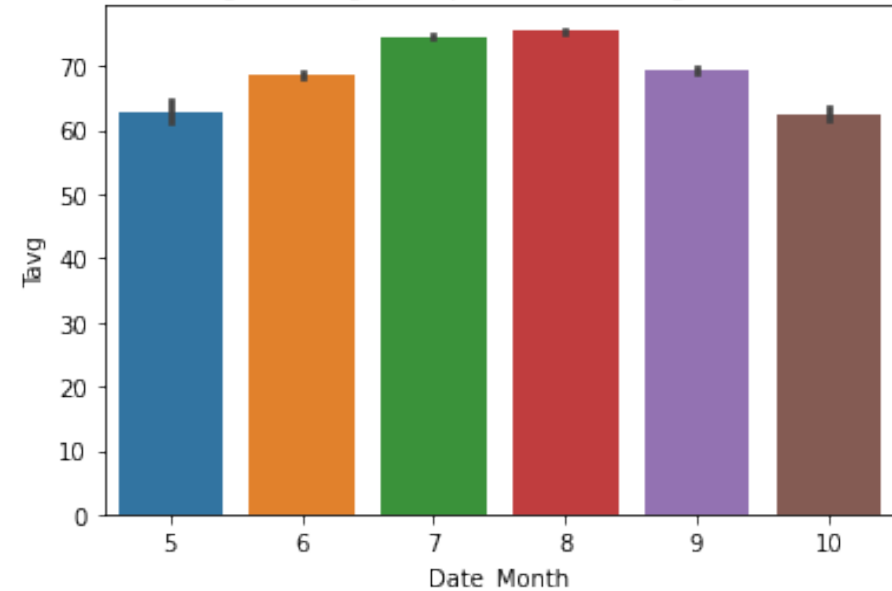
Exploratory Data Analysis

- Majority of the collections took place in the summer months, with August having the most positive mosquitoes

- Samples were collected from 2007-2013

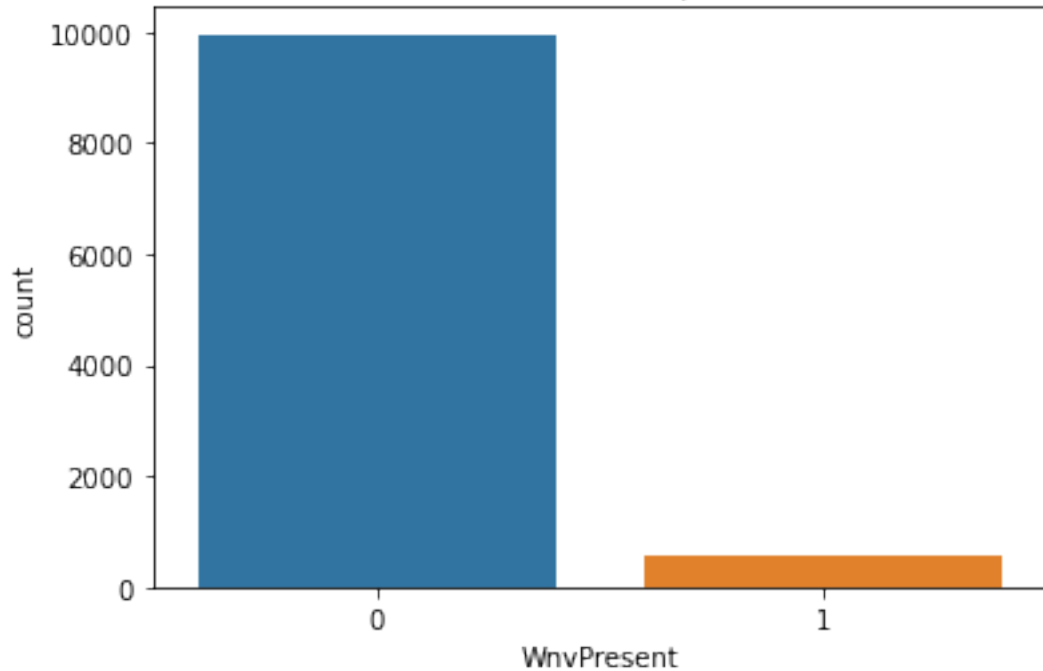- August shows to have the highest average temperature

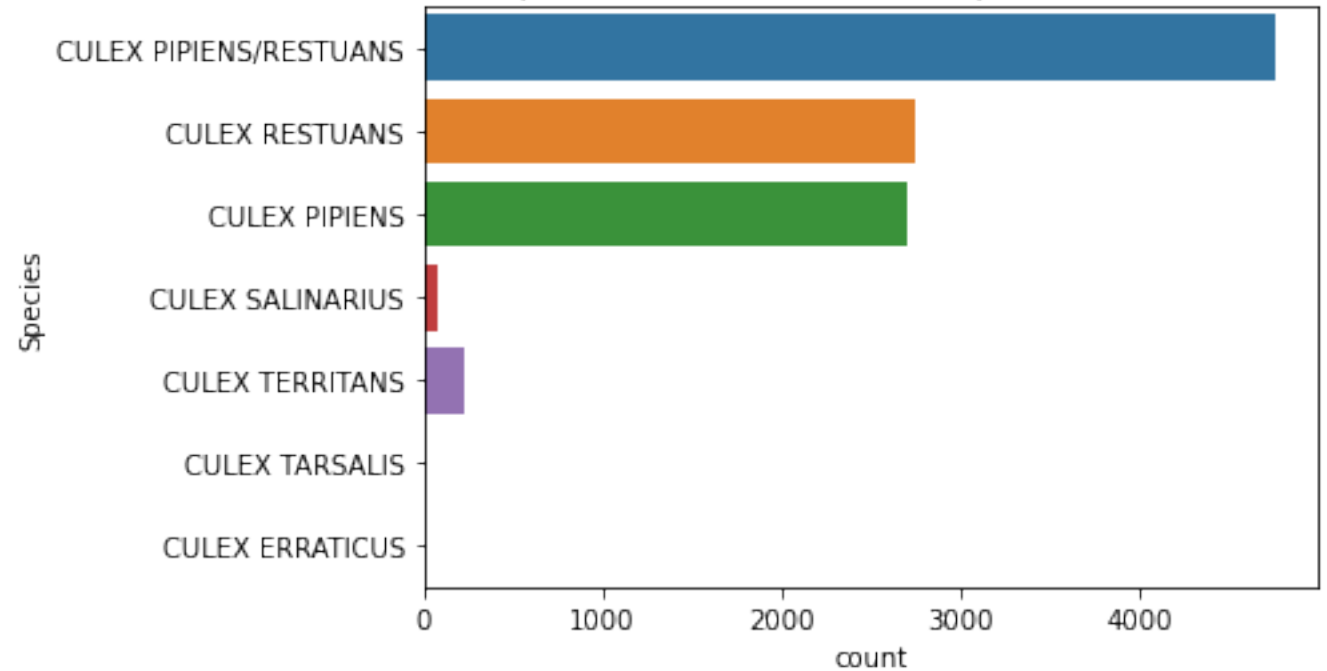Virus Presence amongst the Collected Mosquitoes during the Months

Chicago Average Temperatures during the Months
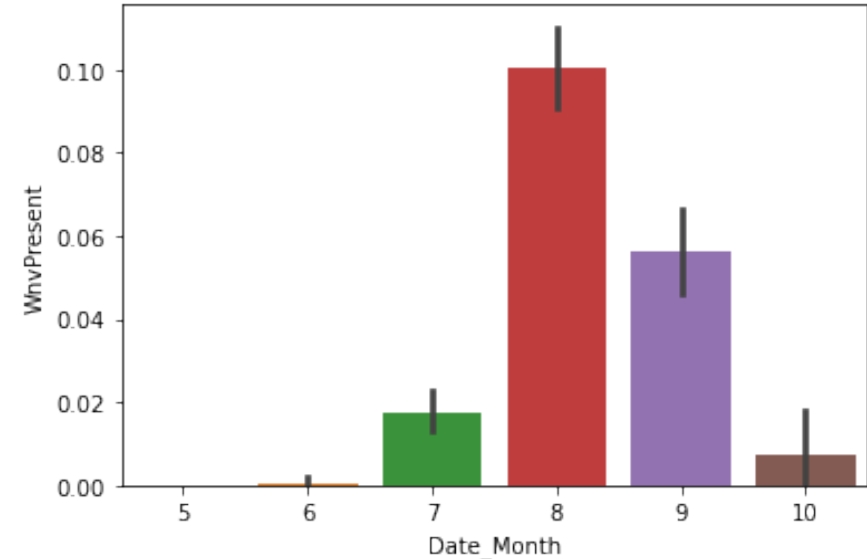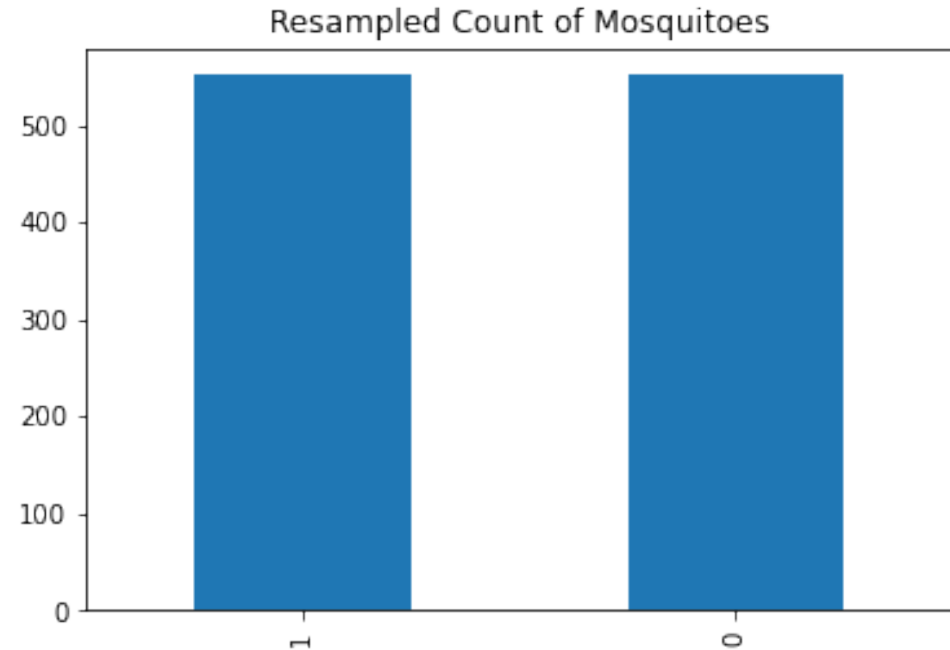
- Only 5.2% of the collections actually contained the West Nile Virus
- Majority of the mosquitoes collected were Culex Pipens and Culex Restuans, and of that, 16.3% of them were positive for the virus.
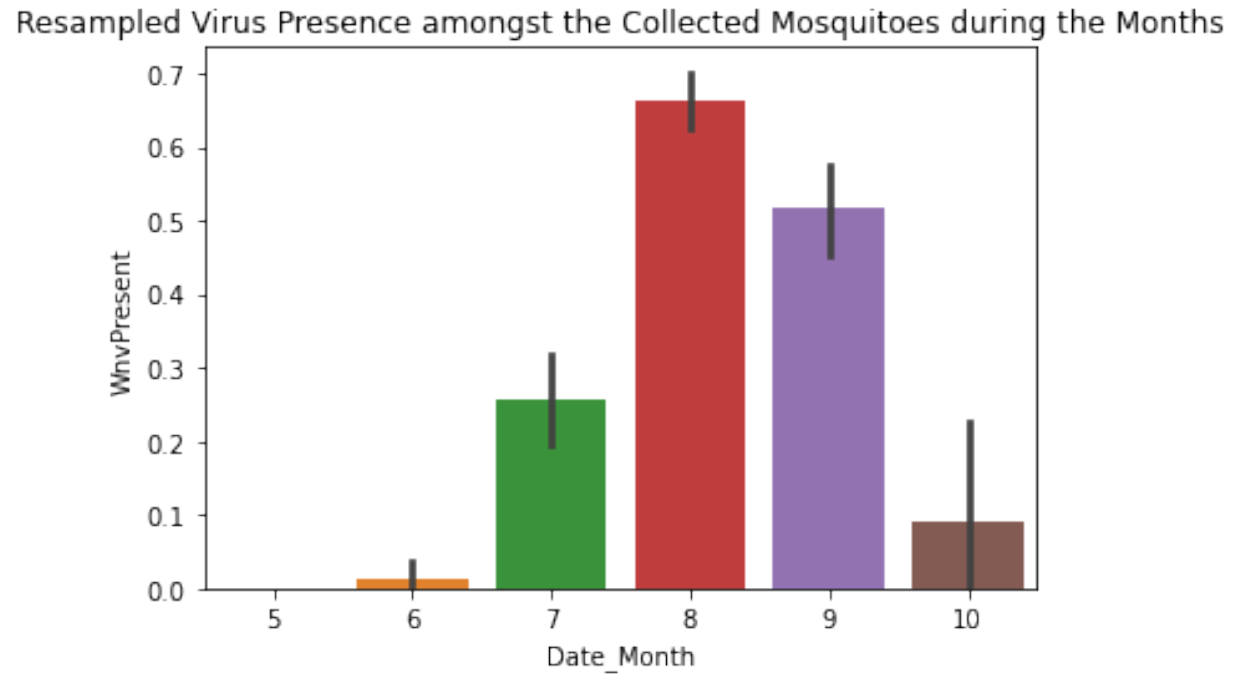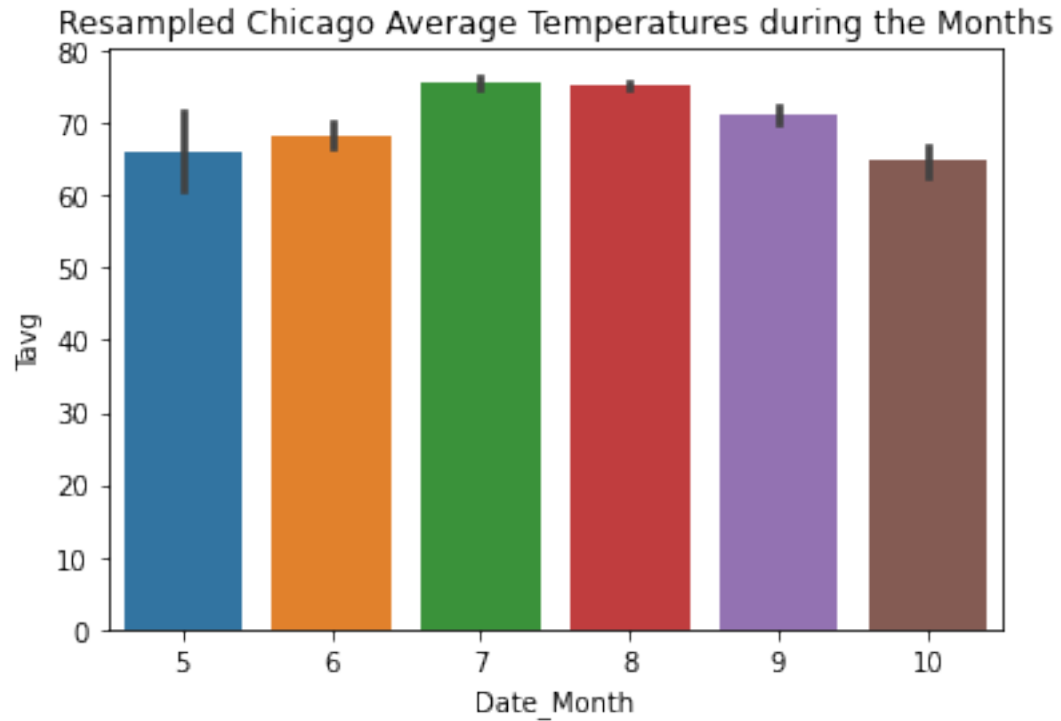
- Station 1 is Chicago O'Hare International Airport

- Station 2 is Chicago Midway International Airport

- Both stations have relatively the same temperatures; therefore, temperature collections at one station can be used to generalize both locations
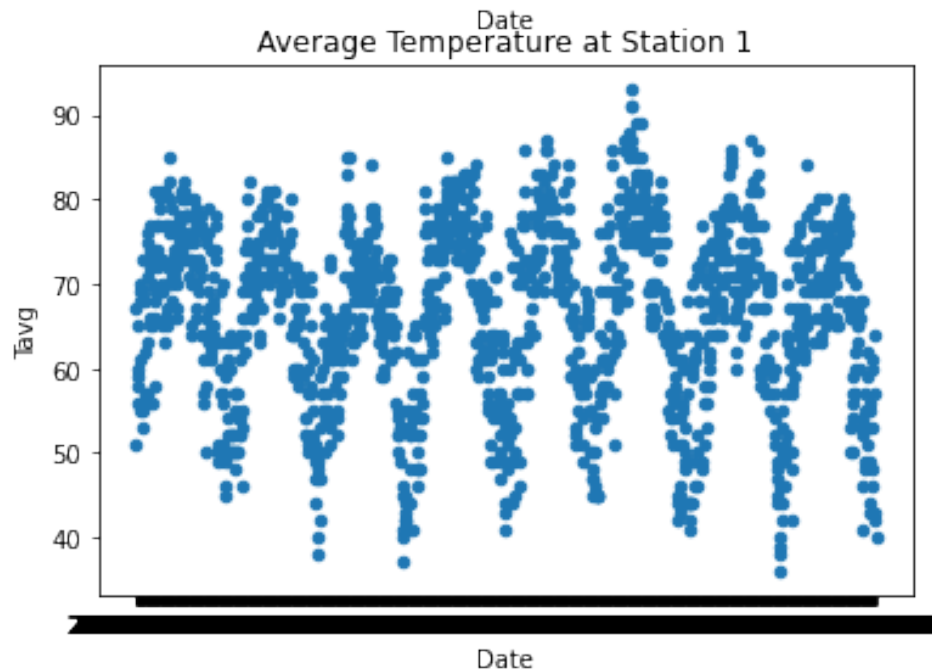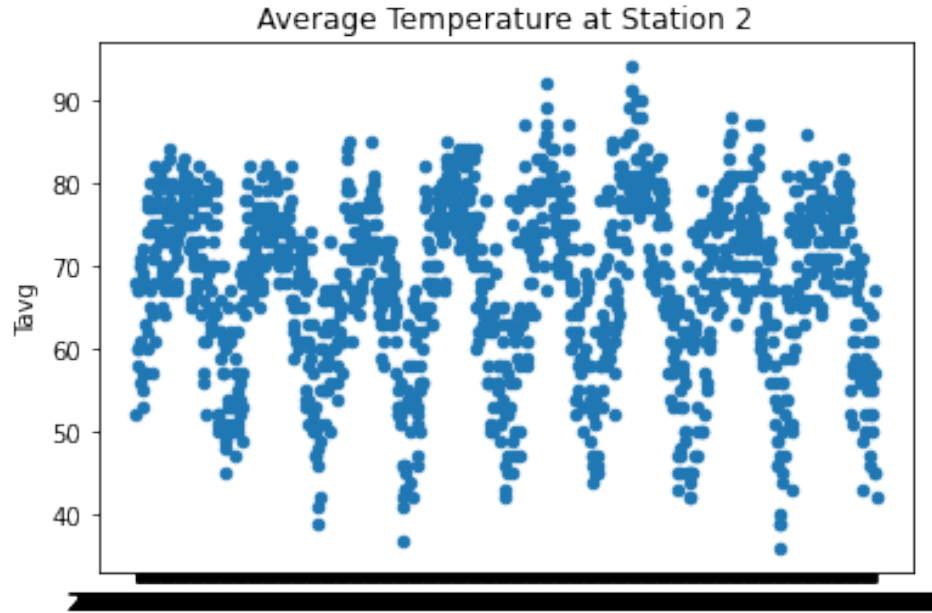
Pre-processing

- Since only 5.2% of the data contained the virus, the data was undersampled

- The distribution contained an equal number of viruses that either contained or did not contain the virus

Resampled Count of Mosquitoes

Resampled Chicago Average Temperatures during the Months

Resampled Virus Presence amongst the Collected Mosquitoes during the Months

- The resampled data shows that July and August temperatures are relatively the same

- August, however, remains to have higher number of positive mosquitoes

Modeling

- Training and testing sets was created with a test size of 0.25

- Gradient Boosting algorithm had the highest AUC score of 0.81

- Confusion matrix shows Having the same probability of the true positive and false negatives

Receiver operating characteristic

Confusion Matrx of a Gradient Boosting Model
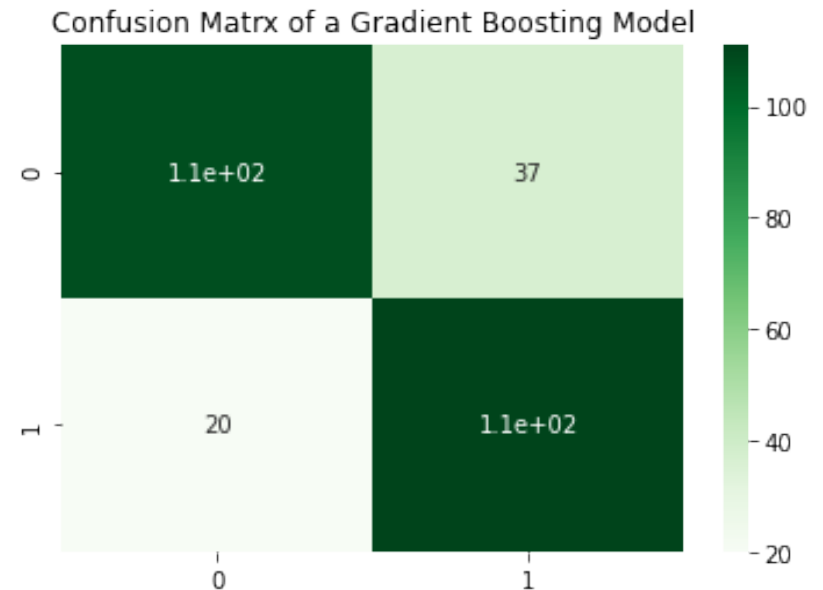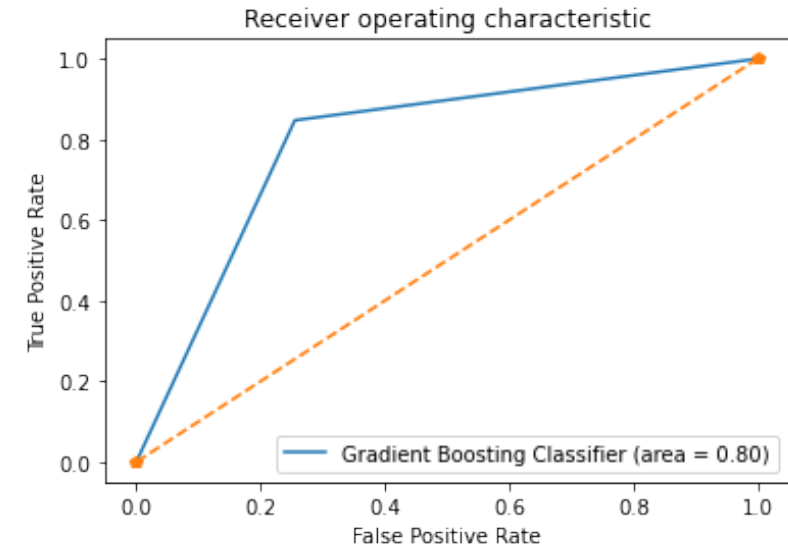
Findings
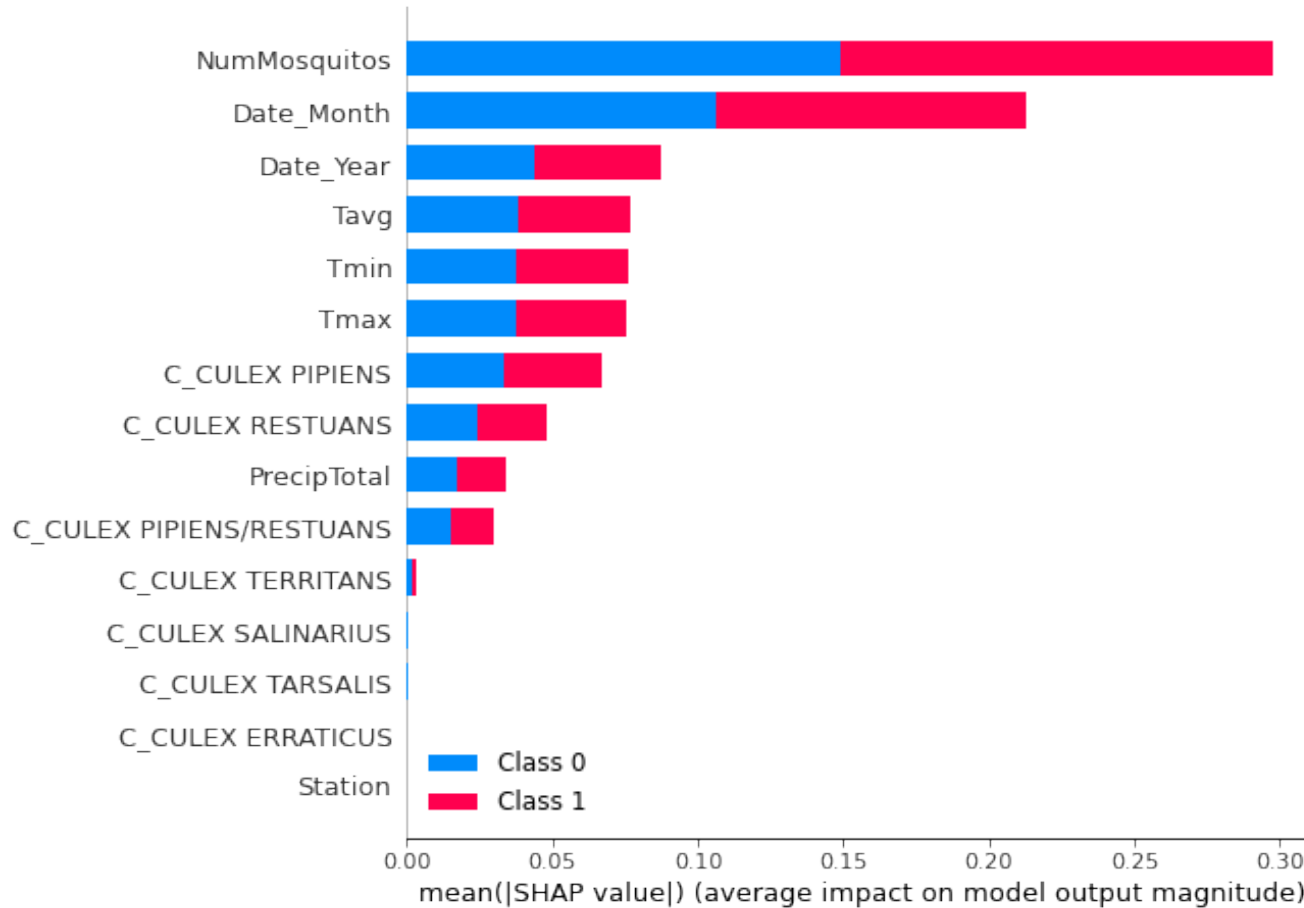
- Information Value (IV) and Variance Inflation Factor (VIF) were used to analyze the different features
  - Preferred to have a VIF value less than 5 and IV value between 0.02-0.8
  - Average Temperature and Minimum Temperature was removed because it had a VIF value of 718 and 98 respectively
- SHAP analysis was done to assess the impact on the model
  - The month in which the collection was taken correlates with the virus predictions
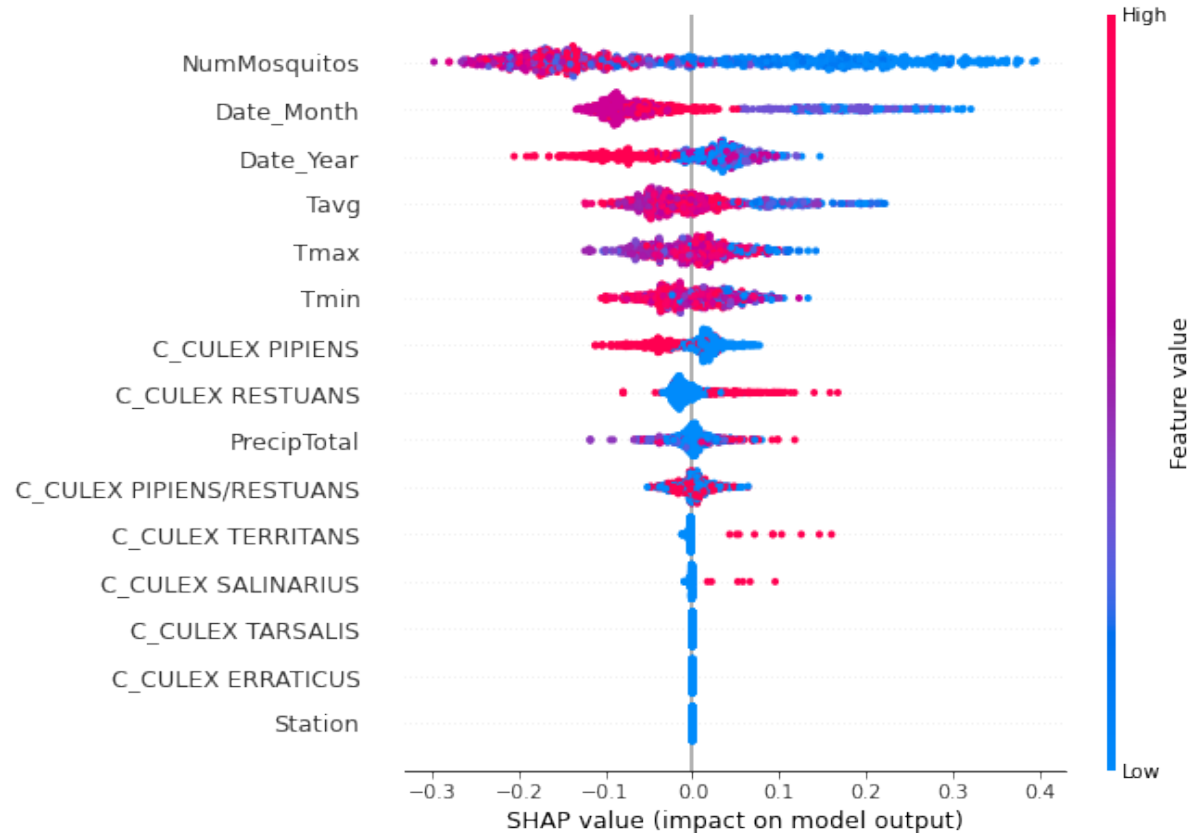
- CDPH should start obtaining more prevention measures in August where the temperatures remain higher for a longer period of time

- Further research can be done by analyzing the data of the people who had the West Nile Virus and understand the impact of the virus