West Nile Virus Analysis in Chicago

Abstract

The West Nile Virus, mosquito-borne disease, been under surveillance in the city of Chicago since 2004 and the Chicago Department of Public Health (CDPH) has made efforts to reduce the virus present by spraying airborne pesticide during different times of the year. Analysis of the provided data is of interest because CDPH would like to determine the optimal date and temperature to spray pesticides so that the city of Chicago can maximize the effectiveness of spraying but still avoid excess spraying so that its people will not be harmed. After running different models, it was found that the Gradient Boosting algorithm with an AUC score of 81% had a better accuracy in determining which features impact on the virus presence give a misquote collected. It was found that in August, temperatures are persistently high, favoring mosquito's presence and increasing viral existence.

Introduction

The West Nile Virus, a single stranded RNA virus, is known to spread amongst mosquitos. According to the Centers for Disease Control and Prevention (CDC), it is the leading cause of mosquito-borne disease in the United States. It spreads through infected mosquitoes by bite. The West Nile Virus is known to spread most efficiently in temperatures that range from 75.2 -77 F. Its presence is seen from the summer and carries throughout the fall. Though fatality is minimal, 1 of 5 people face its symptoms each year., whereas 1 out of 150 people develop serious symptoms. Around 20% of people who become infected develop symptoms ranging from a persistent fever to serious neurological illnesses that can result in death.

Chicago Public Health Workers each year set traps around the city to collect mosquitos and test whether the virus is present in them or not. From 2011-2013, the city of Chicago sprayed its mosquitos with airborne pesticide so that the number of adult mosquitoes and the virus presence could decrease.

The city of Chicago wants to predict whether the West Nile Virus is present for a given date, temperature and species and hopes to use those findings to effectively spray mosquitoes and reduce the exposure of the virus to its people.

Exploratory Data Analysis

The Chicago Department of Public Health had provided weather and mosquito collection data for analysis. The main data gave the exact location and the number of mosquitoes collected at different collections in Chicago. The feature of interest is whether the virus was present on not in the collection. One (1) denotes that the virus is present and zero (0) denotes the virus is not present. Weather data gave various weather information, such as temperature average, dew point and precipitation total at two different stations. A merged dataframe was created with respect to date and used to create more features as well as models. Looking at the data, about 5.2% percent of the collections contain the West Nile Virus. Majority of the collections contain the Culex Pipiens and Culex Restuans which of that, about 16.3% contain the virus. Virus presence becomes predominant during the late summer and early fall months.

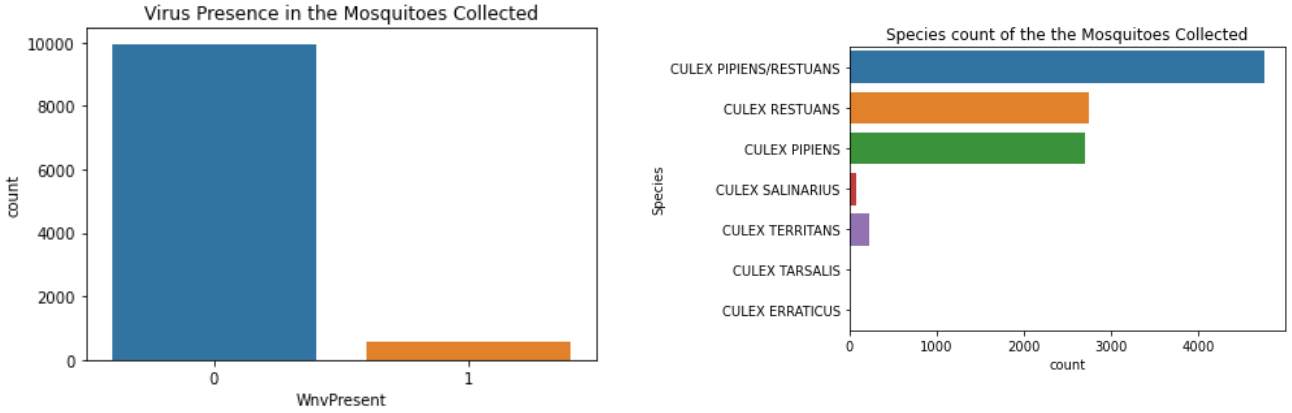


Figure 1: West Nile Presence amongst the collected data. (L) Distribution of the species presence in the mosquitoes collected in the data (R)

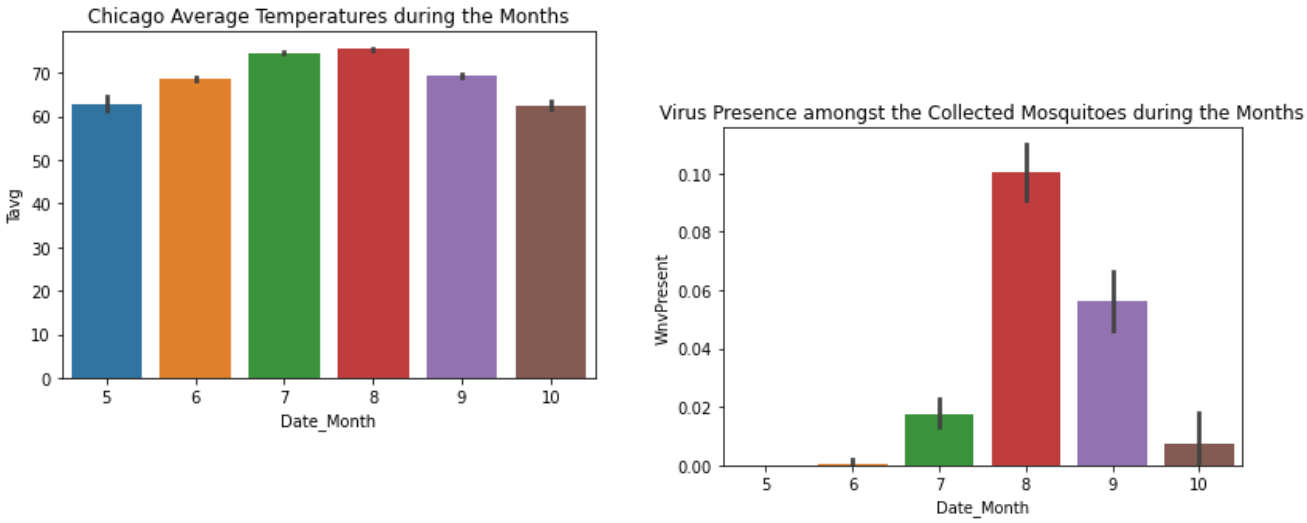


Figure 2: Chicago temperature average during the spring, summer and fall months. (L) West Nile Virus presence amongst the collection during the spring, summer and fall months. (R) Months are denoted with respect to their universal numerical association. (e.g 5=May, 6=June)

The two stations in the weather data are the two main airports in Chicago. Station 1 is Chicago O'Hare International Airport while Station 2 is Chicago Midway International Airport. Both stations appear to have similar temperature ranges during the collections; therefore, temperature collections at one station can be used to generalize both locations.

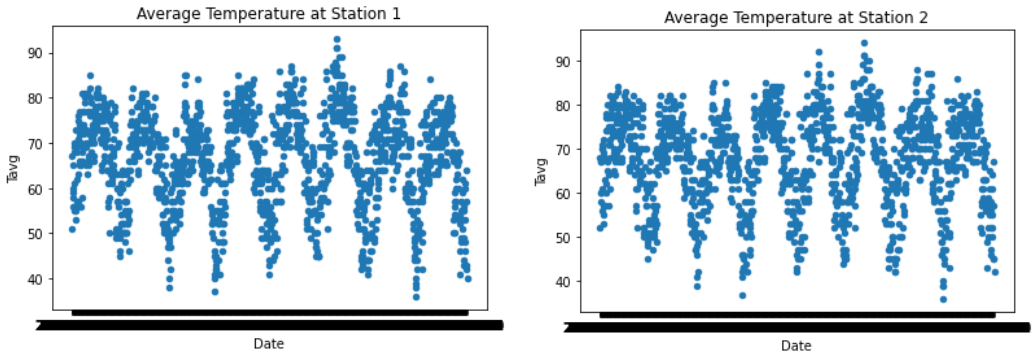


Figure 3: Temperatures recorded at different dates at Station 1 (L) and Station 2 (R)

Pre-processing and Feature Engineering

In order to future analyze the data and time period, features were created to break down the months and years. One hot encoding was used to categorized the species column. Because only 5% of the original data contained the virus, data was resampled to further analyze the virus presence. Resampling the data aids in understanding and quantify the uncertain population. The resampled data contained a sample size of 1102 with each category having a size of 551.

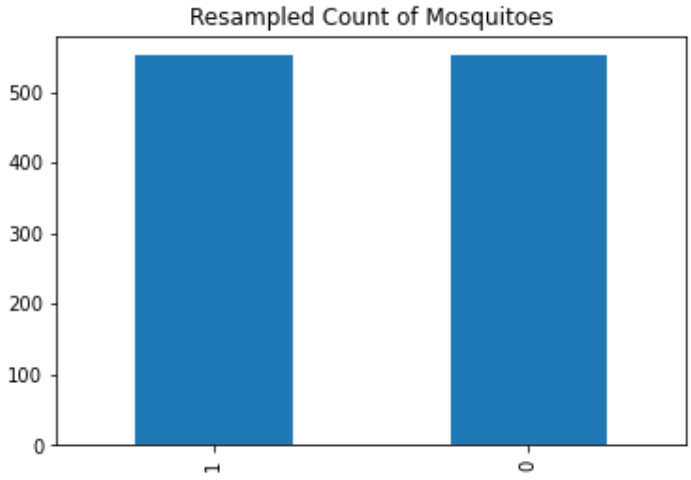


Figure 4: Resampled data containing an equal distribution of West Nile Virus positive and negative mosquitoes. One (1) denotes that the virus is present and zero (0) denotes the virus is not present.

Binary Classification was used to understand if it is a good predictor variable to classify whether the virus is present or not on a given a date and weather. Using the information value (IV), different predictor variables were selected for a binary logistic regression. The range of IV values used to understand the features significance was 0.02-0.80. Variance Inflation Factor (VIF) was also used to analyze the effectiveness of the predictor variable. It provides a measure of multicollinearity among the independent variables in a multiple regression model. A VIF score greater than five is removed to decrease multi collinearity.

Modeling

Logistic Regression, Radom Forrest, and Gradient Boosting algorithms were ran to assess its accuracy and determine which model as a higher AUC score. Test and train data was created with the under sampled data with a test size of 0.25. After fitting the models, its confusion matrix was generated to visually determine the probability of the true positives and the false negatives occurring in the different classifiers.

Grid Search CV was done to perform hyper parameter tuning and determine the optimal values for a given model. This is significant as the performance of the entire model is based on the hyper parameter values specified. The best score for its specific parameters was 0.5 with the best parameters being 'max_depth' and 'n_estimators'.

Shapley Additive Explanations(SHAP) Analysis was performed for each feature to see the change in the expected model prediction. Each SHAP value explains the difference between the average model prediction and the actual prediction of the instance.

Findings

After creating model using the different algorithms, it was found that Gradient Boosting Classifier had the highest AUC score of 0.81. This indicates that there is an 81% chance that the model accurate indicate if the virus is present given a mosquito that is caught during collection. Looking at the confusion matrix, the probability of true positives is relatively the same as the probability of the true negatives.

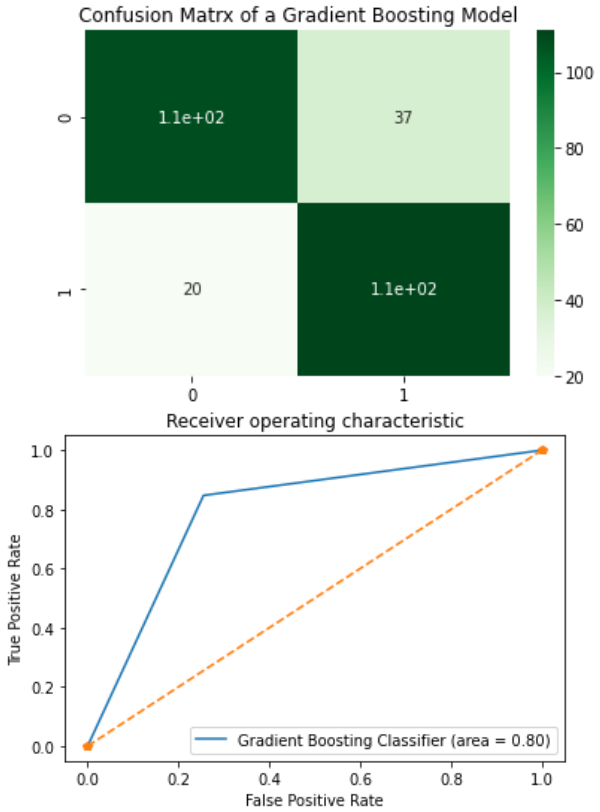


Figure 5: Confusion Matrix of the Gradient Boosting Classifier (T). ROC curve with an AUC of 0.80 (~0.81) (B)

The summer months in Chicago had most virus present; however, during the month of August, temperatures reach its maximum value and stay at those temperatures for time. This increases the virus presence in the collections taken.

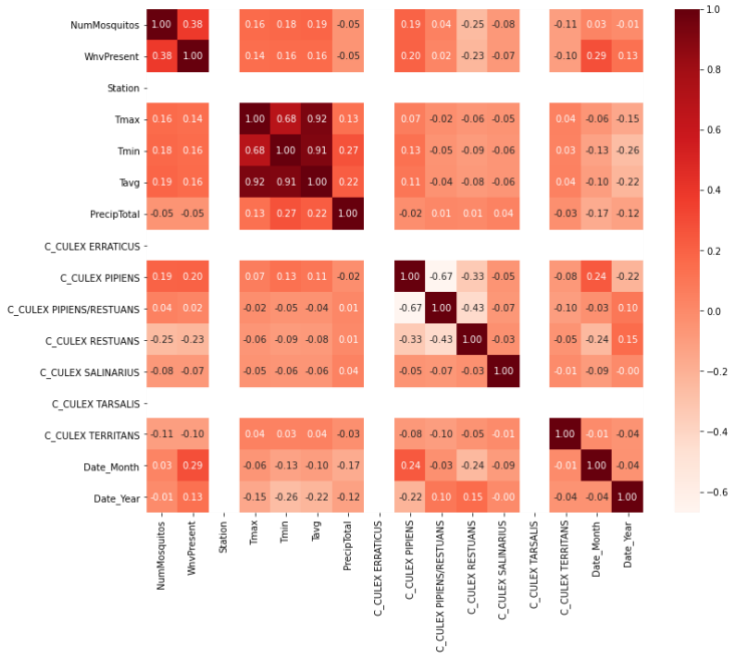
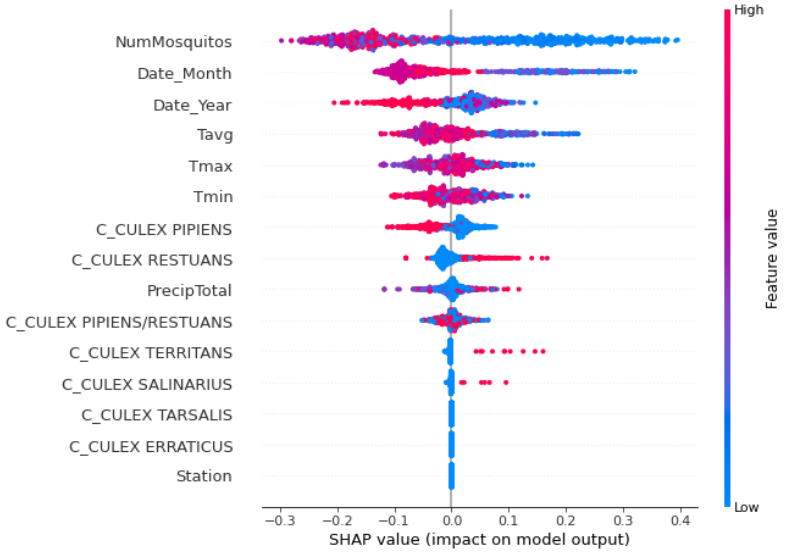


Figure 6: Heatmap of the resampled data

Looking at the SHAP summary plot, the Culex Pipen and Culx Rusten species tend to be the most impactful and an important feature compared to the other species. The month in which the virus is present also has a higher SHAP value, indicating that it is an important feature.



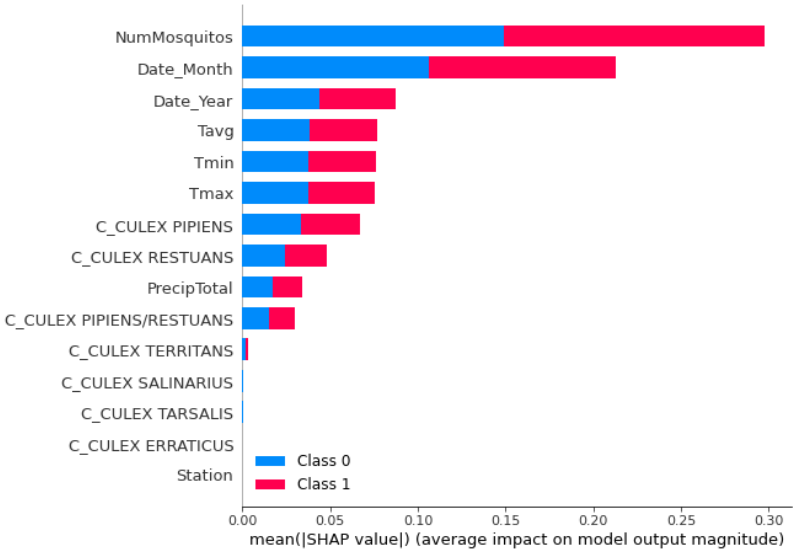


Figure 6: SHAP summary plots of the resampled data

Because the model shows that virus presence increases during August, the city of Chicago should start more preventative measures around August since majority of the virus presence is observed then. Further research can be done by analyzing the data of the people who had the West Nile Virus and see how the presence of the virus impacts Chicago's people.

Resources

1. Rising temperatures could shift US West Nile virus transmission: New model suggests climate change may increase the areas in the US with optimal temperatures for West Nile virus transmission. (2020, September 15). ScienceDaily. https://www.sciencedaily.com/releases/2020/09/200915105932.htm#:%7E:text=Nile%20virus%20transmission,New%20model%20suggests%20climate%20change%20may%20increase%20the%20areas%20in,for%20West%20Nile%20virus%20transmission&text=Summary%3A,)%2C%20a%20new%20study%20shows.

2. West Nile virus | West Nile Virus | CDC. (2020, June 3). Centers for Disease Control and Prevention. https://www.cdc.gov/westnile/index.html