West Nile Virus Report

Mosquitos are an insect that can carry disease with them. In this case, a common and harmful disease is the West Nile virus. Chicago's Department of Health has gathered data with their Chicago Surveillance Program. This data covers many different points in order to help understand when and where mosquitos will be and whether or not they will be carrying the West Nile Virus. Using this dataset I will create a model to help better predict this information about mosquitos in the Chicago area.

Firstly, the dataset needed to be cleaned up. Two separate datasets were given, one with weather data and another with data based on certain locations of weather stations. The weather dataset contains 22 columns while the training set has 11 columns. Within the 10506 data points given in the training dataset, West Nile Virus was present in 551 of those.

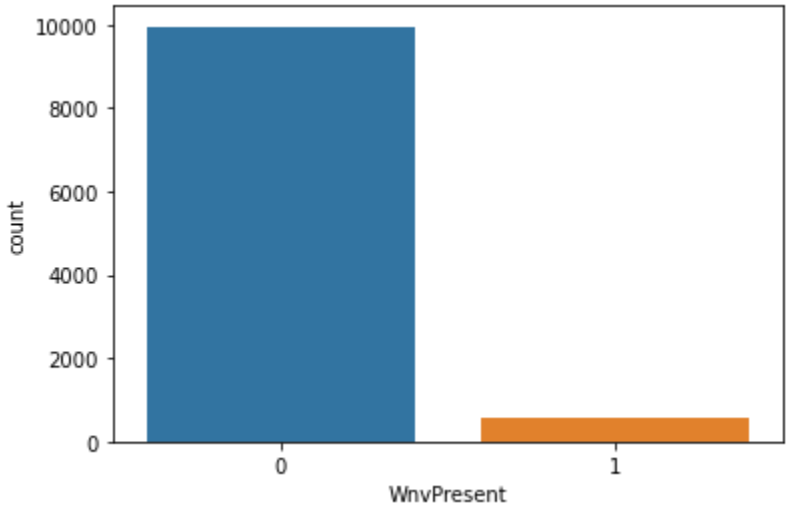

Fig. 1: Count of West Nile Virus Present in Dataset:

Only 3 species of mosquitoes accounted for those 551 cases, leaving three species that have no trace of the West Nile Virus.

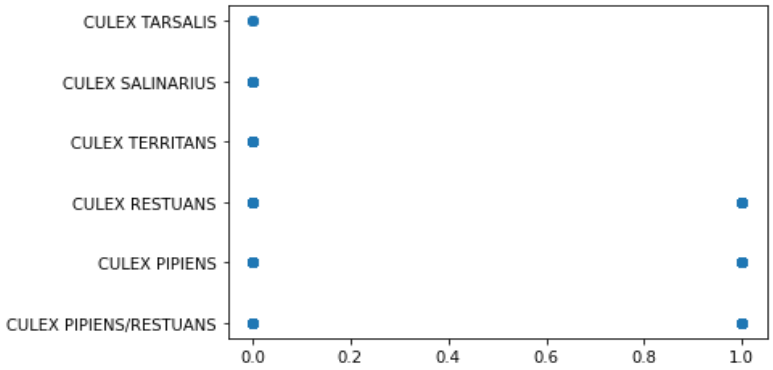

Fig 2: Species and the Presence of West Nile Virus

My next task on cleaning the dataset was dropping columns I thought to be unnecessary such as Depart, CodeSum, Depth, SnowFall, etc. As in most data sets, there were also columns with null values which would interfere with future steps in the modeling process. These values were replaced with 0's to ensure no issues later on and in order to better understand the data, I created some new columns. Three new columns that were created were the month and weekday that the data was given as well as the distances based on weather stations and the traps. The month and weekday will help me understand what time of year is most notorious for mosquitos and the West Nile Virus. The distance will allow me to make the dataset smaller by reducing repetitive data and get a better idea on location of the mosquitos. The month that had the most cases of West Nile Virus was August with 131 cases. It also was the month with the most caught mosquitoes.

	Total_Count	Wnv_Count	Wnv_Percent
8	1363	131	9.611152
9	1143	51	4.461942
7	1093	16	1.463861
6	816	1	0.122549
10	142	0	0.000000
5	59	0	0.000000

Fig. 3: Total Count and WNV count based on Month

To further ensure the success of the modelling process, I joined these tables by Date and then changed the date to the correct datetime format.

Once the data was cleaned and explored, I moved onto pre-processing and creating train and test sets. The data was examined through information value (IV) technique to see which columns were important and which were not. A IV value less than 0.1 is not useful for modelings and a IV value greater than 0.8 is too biased. Thus we are looking for values that are between 0.1 and 0.8. In the end, the technique produced a list of 12 columns that were based on location, species, precipitation, and month. Now that we have the columns that were useful for modeling purposes, I chose to use XGBoost to carry out the rest of the project. The model had an accuracy score of 95.6% and an area under curve score of 0.5. To further understand the data and the remaining columns, I had to understand which columns had the most impact on the modeling and prediction process. For this I created a definition that would select features and see which one had the biggest impact on the prediction. The best predictor was the month of August, as we saw before August had the most mosquitoes captured along with the most positive cases of the West Nile Virus. Next I used SHAP analysis to further enhance and

visualize the data to see the impact of each column on how they predict the West Nile Virus. After the month of August, Precipitation total was the next indicator of whether or not mosquitoes and the West Nile Virus were present. This makes sense as we know mosquitoes like wet and humid environments.

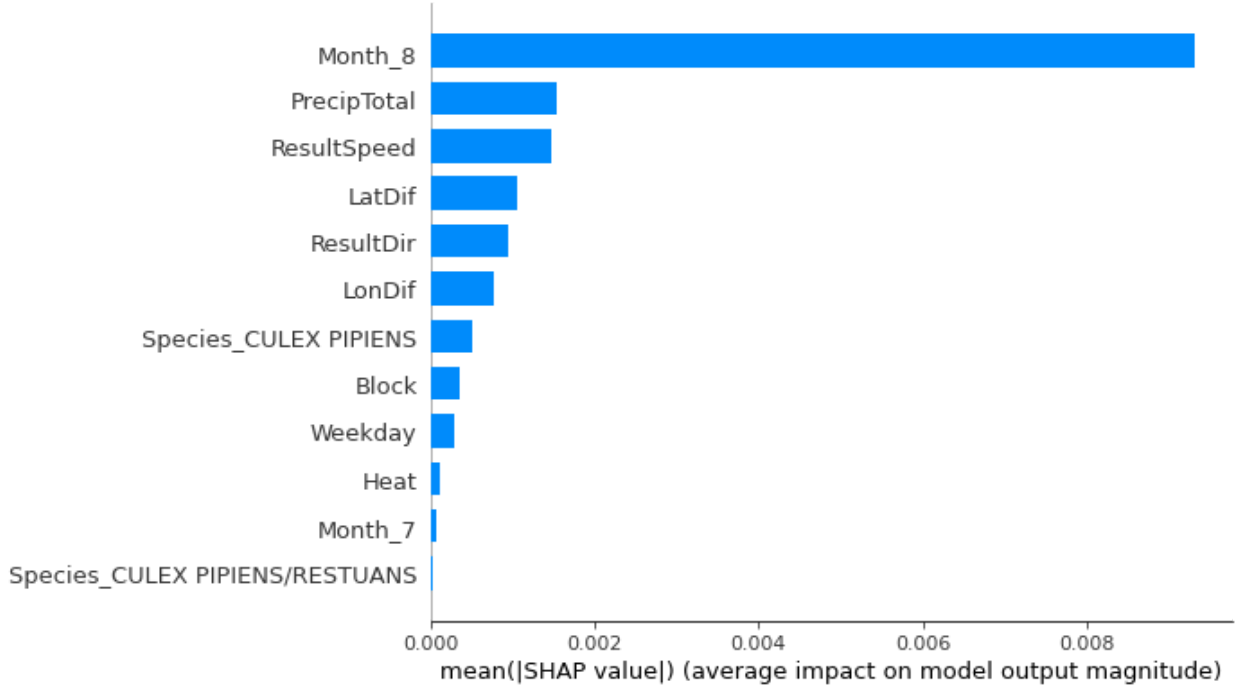

Fig 4: SHAP Summary Plot

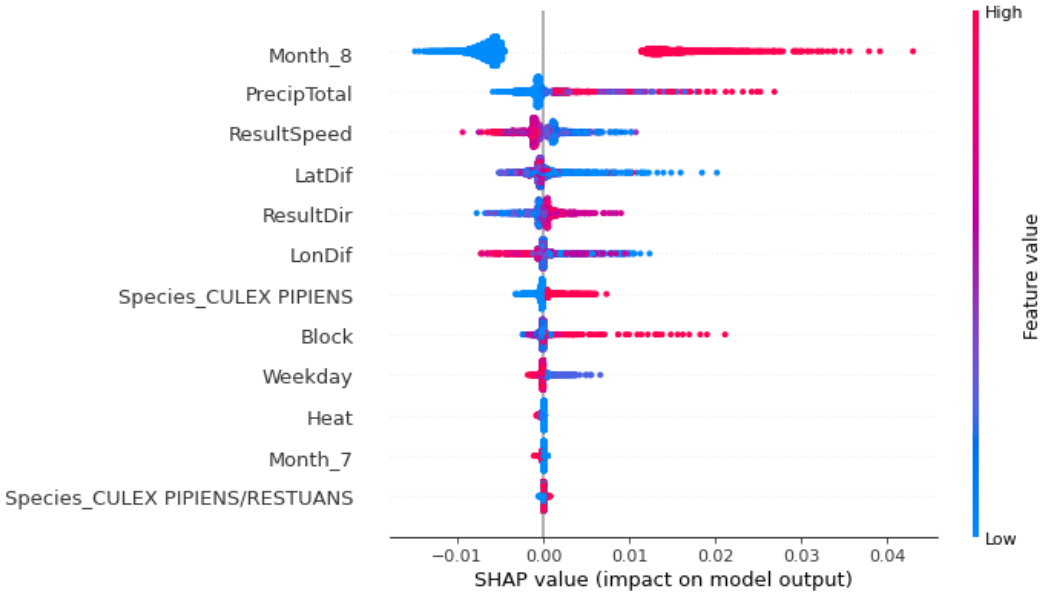

Fig 5: SHAP Summary Plot

In conclusion, I was able to clean the data provided by Chicago's Department of Health and use it to create a predictive model. I had to ensure no null values were within the dataset, that the data was in their correct format, and lastly removing and adding columns in a way to ensure I could get the most information out of the dataset that I could. I used the XGBoost model for which came out to have an accuracy of 95.6% and an auc of 0.5. Lastly using SHAP analysis, I was able to find that the month we were in was the best indicator whether or not West Nile Virus would be present.