Final Project Report On West Niles Virus Prediction

1. Problem Statement

The West Niles Virus(WNV) is the leading cause of mosquito-borne disease in US. It is most commonly spread to people by bite of an infected mosquito. WNV cases usually start to occur in summer and continue through early fall. Every year public health workers in Chicago will set up mosquito traps to test the presence of the virus. These test results can be analyzed to determine effective methods to reduce the active cases of the virus.

In this project we use machine learning techniques to analyze the data. Our goal is to be able to predict the presence of the virus as well as finding the most important factors that contribute to the presence of the virus. Combining the test results with weather data, demographic data and geographical data for Chicago, we are able to build predictive models to solve this problem. Our final random forest model can reach a recall score of 0.94 and AUC score of 0.84, which indicates that our model is sensitive at detecting the presence of the virus.

2. Data Wrangling

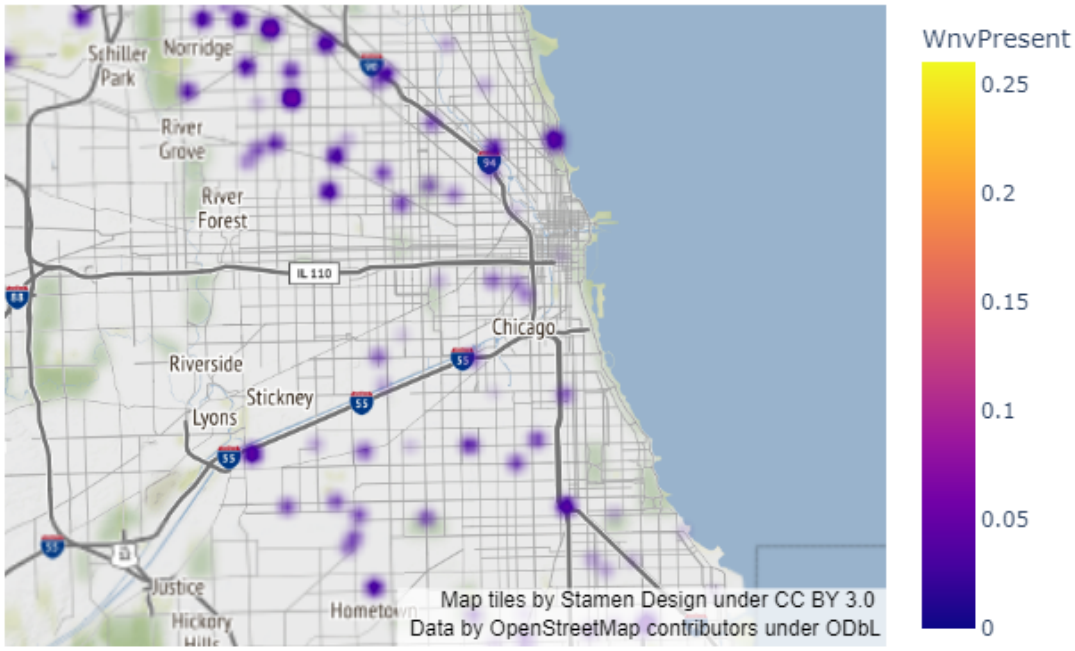
The first step is to clean up the historical weather data of Chicago. It is believed that hot and dry conditions are more favorable for WNV. For the majority of the entries in weather data, they are of string type rather than numeric. So we first transformed the data into correct datatype. Next, the weather data consists of records from both the OHare station and Midway station. Many columns have missing values, but usually only the record in one station is missing, so in that situation we fill it with the value in the other station. Next, in the dataset there are weather codes that stand for special weather conditions. For instance, RA stands for rain. We use one-hot encoding to transform weather codes into features. Lastly we noticed that some of the features are constant, so we dropped those features.

In the second step, we add weather data to the test data. For each test location, we first compute its distance to OHare station and Midway station, then we choose the weather data from the closest station to the test location.

Lastly, we also believe the presence of the virus should be related to the demographic information, so we also added population data, income data and population density data for each test location in our dataset. The demographic data are extracted from online resources.

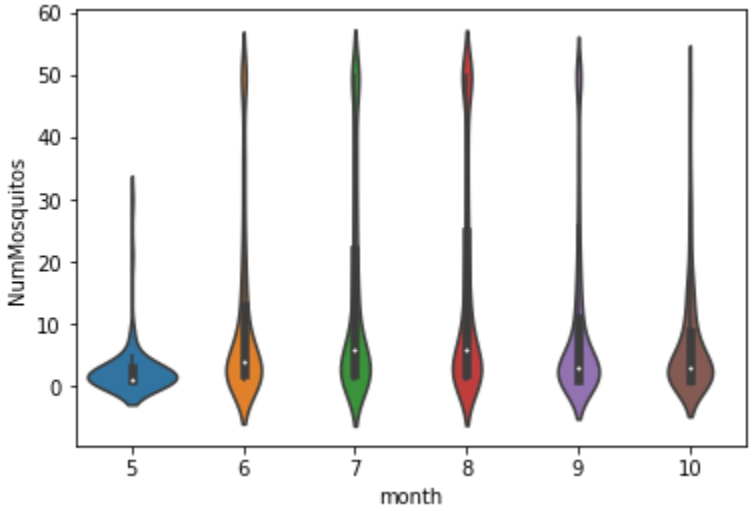
3. EDA Analysis

From our data, first we need to understand the relation between the frequency of virus presence and geographical information. This is provided in the following picture:

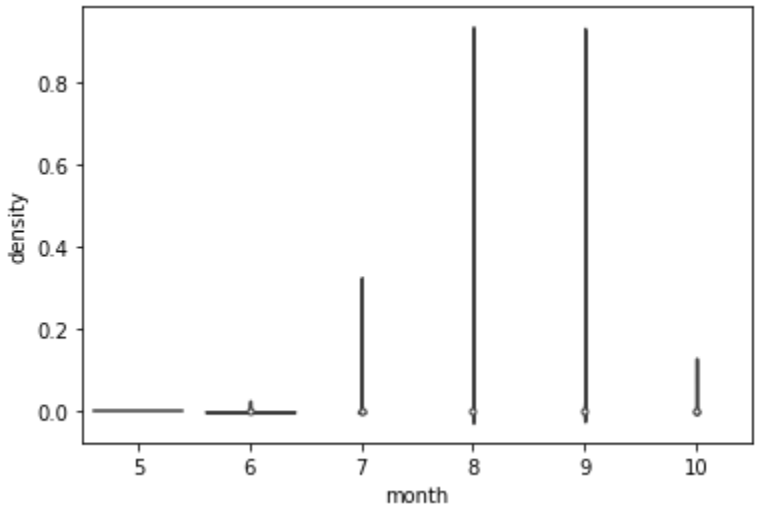


From this we can clearly see that many regions in the north west areas have high frequency for virus presence. Similarly in the south east areas.

Next we analyze the relation between month and virus presence. The following for June through October, the number of mosquitos are similar.

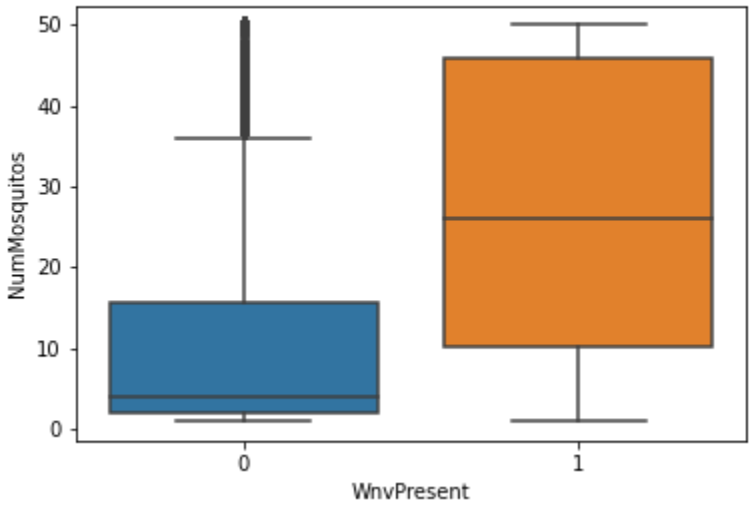


To compare the likelihood of detecting the virus, we take the reciprocal of the number of mosquitos of each record and plot its distribution for each month, we find that the chance of finding virus in July through September are much higher than the rest:

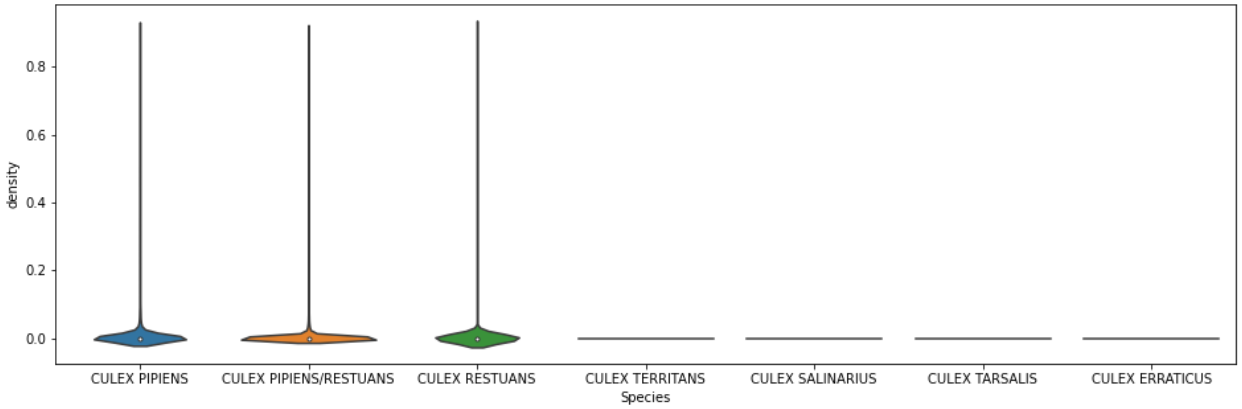


From this one may conclude that month should be an important factor. This agress our domain knowledge that WNV cases usually occur in summer and early fall.

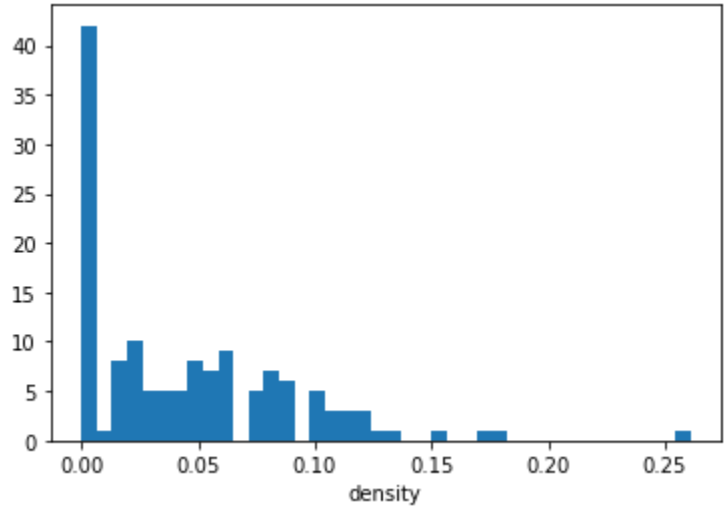
Intuitively, the more mosquitos we catch, the higher the chance that we will detect the virus. So the number of mosquitos is likely to be an important factor in our model. The following picture confirms this:



It turns out that the chance of detecting the virus is different across different mosquito species. So species is likely to be an important factor in our model:



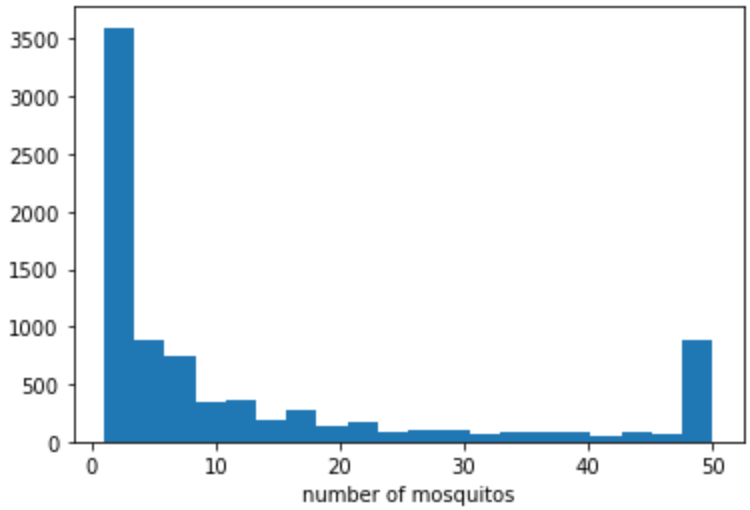
Because different locations have different likelihood of detecting the virus, it is useful to know the distribution of the likelihood for different locations. The following plot shows this. For many locations, the virus is never present, The distribution is right skewed and has a long tail.



One can see a similar pattern for the number of mosquitos for different locations. The maximal number of mosquitos for the traps are capped at 50, so the maximal value is 50.

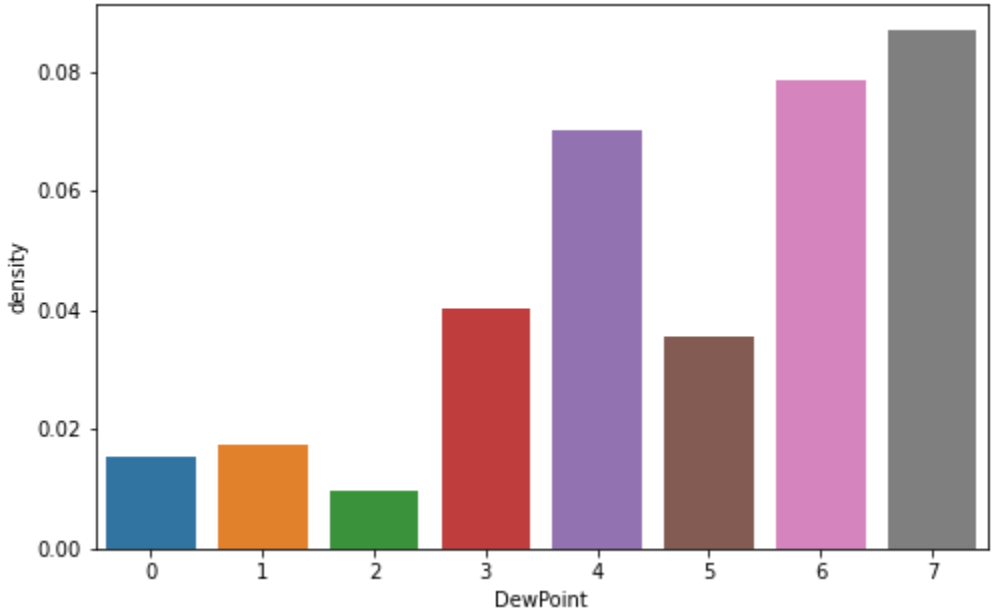
Next we look at the influence of weather on the presence of virus. Since the weather in the past may also influence the number of mosquitos, we also generate lag features for weather data.

First we look at dewpoint vs the chance of detecting the virus. Our domain knowledge is that dry and hot weather is favorable for the virus. But the following plot shows that the situation is more complicate, higher humidity does not necessarily corresponds to higher chance of detecting the virus:

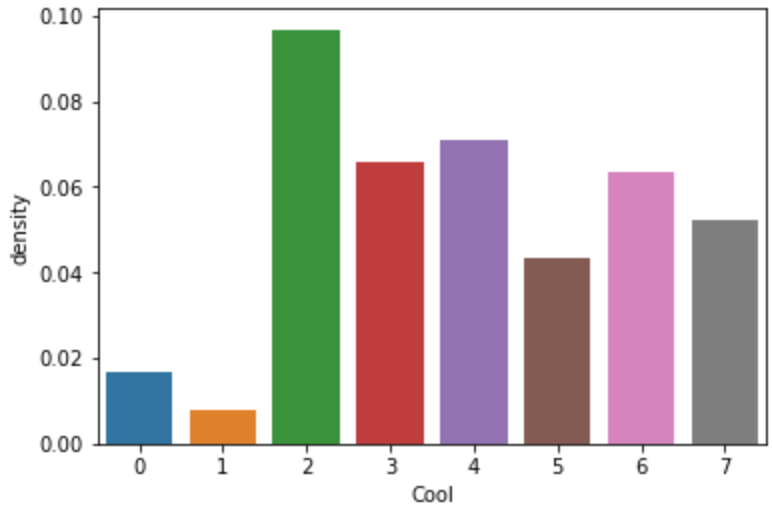


Next we look at the influence of weather on the presence of virus. Since the weather in the past may also influence the number of mosquitos, we also generate lag features for weather data.

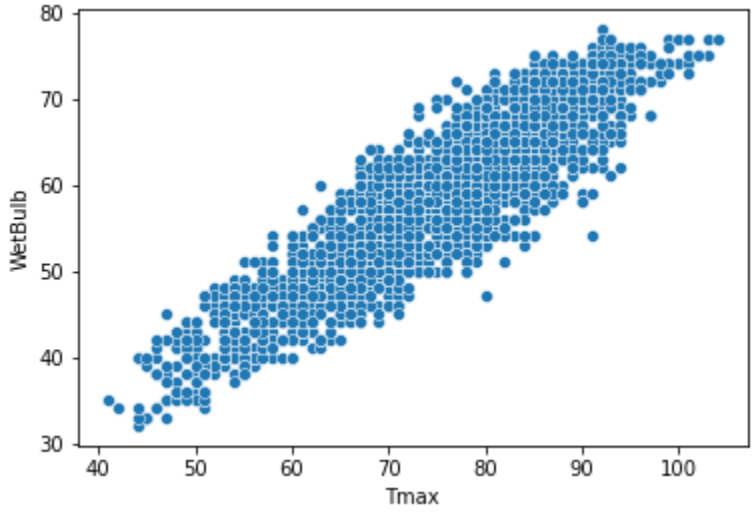
First we look at dewpoint vs the chance of detecting the virus. Our domain knowledge is that dry and hot weather is favorable for the virus. But the following plot shows that the situation is more complicate, higher humidity does not necessarily corresponds to lower chance of detecting the virus:



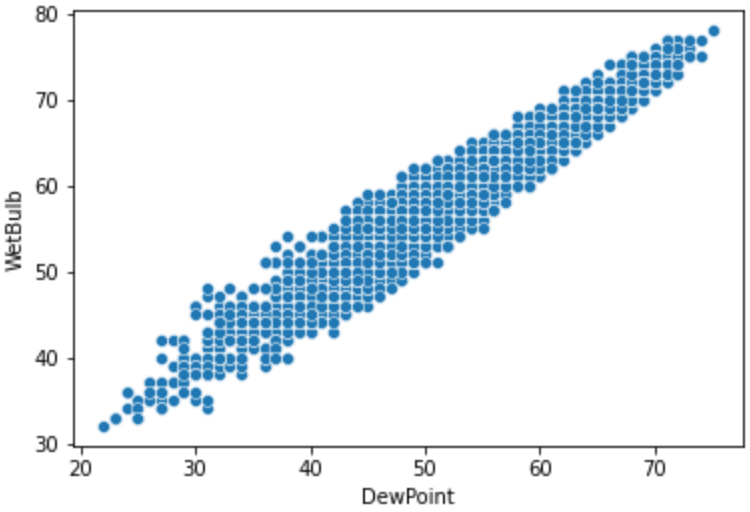
Similarly, higher temperature does not necessarily corresponds to higher chance of detecting the virus:



Next, some of the features are highly correlated with each other, especially among lag features. For instance, Tmax and WetBulb are highly correlated:



DewPoint is also highly correlated with WetBulb:



4. Feature Selection

Since we have many features and we also have high feature multicolinearity, we need to do some feature selection before we finally fit our model. Our first step is to compute the information value for each feature. We compute information score for each feature and then only select features with information value between 0.01 and 0.8.

Next, we also need to reduce feature multicolinearity. To do this we use the variance inflation factor. We do this iteratively. Each time we select the feature that has the highest variance inflation factor, then we eliminate this feature from existing features. In this way we remove features until all features have variance inflation factor less than ten. In the end, the number of features drop from 79 to 48. It turns out that many weather features are dropped, including those that are supposed to be significant from our background knowledge, such as drewpoint, relative humidity, average temperature and so on. To make a comparison, we will still fit the original dataset into our model later and compare the outcome between the two datasets.

5. Modeling

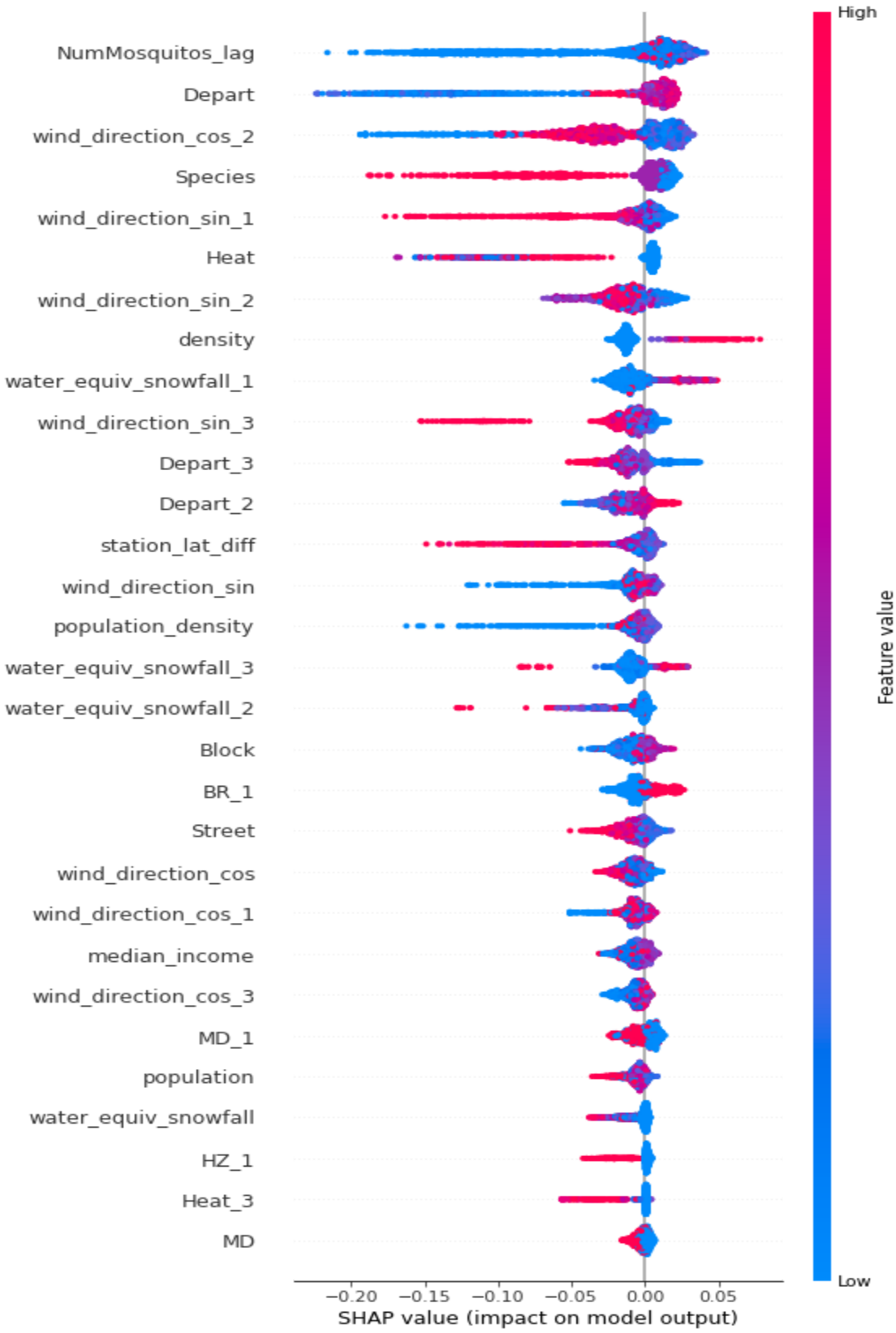
We tried random forest, xgboost, lightgbm, logistic regression and SVM on our dataset. We aims to find model and hyperparameters that maximize the AUC score. In our situation it is more important to be able to detect the presence of the virus, so the recall score for class 1 is also an important factor. Next, our dataset is highly imbalanced. We only have 551 positive class in total, but 10506 data. So we did undersampling for our dataset. When training models, we also search for the optimal class weights that will maximize the AUC score. To make comparison, we train each model on the original dataset as well as on selected features. Here's a summary of the results:

Model	Dataset	AUC	Recall

Random forest	Full dataset	0.85	0.95
Random forest	Reduced dataset	0.84	0.94
Lightgbm	Full dataset	0.84	0.88
Lightgbm	Reduced dataset	0.82	0.84
Xgboost	Full dataset	0.83	0.85
Xgboost	Reduced dataset	0.82	0.84
Logistic Regression	Full dataset	0.82	0.83
Logistic Regression	Reduced dataset	0.80	0.74
SVM	Full dataset	0.83	0.83
SVM	Reduced dataset	0.80	0.93

Based on AUC and recall score for the positive class, the random forest model has the best performance.

Next we use shap analysis to determine the influence of each feature on the model predictions. Since the full dataset suffers from feature multicolinearity, to analyze the influence of each feature we shall use the reduced dataset, where we have removed feature collinearity using variance inflation factor. We have the following observations. First, by looking at Depart, we see that if the temperature is unusually high, then the chance of detecting the virus will be higher. Next, the mosquito species also have significant influence, which agrees with our findings in EDA. From the lag features of wind_direction_sin, we see that we have a higher chance of detecting the virus if the east wind prevails in the last few days. This also makes sense since east wind brings humidity from Lake Michigan, which might be favorable for mosquitoes to survive. By looking at Heat, we conclude that higher temperature appears to correspond to higher chance of detecting the virus. BR_1 implies that if there's mist yesterday, then the chance of detecting the virus will be higher. HZ_1 implies that if there's haze yesterday, then the chance of detecting the virus will be lower. From density and NumMosquitos_lag features we see that if we have a high frequency of virus presence, then it is also likely that we will detect the virus today. From population we see that areas with WNV appear to have lower population. From the shap values for the full dataset, we also see that when the humidity of the previous day is high, then we also have a higher chance of detecting the virus.


NumMosquitos_lag
Depart
wind_direction_cos_2
Species
wind_direction_sin_1
Heat
wind_direction_sin_2
density
water_equiv_snowfall_1
wind_direction_sin_3
Depart_3
Depart_2
station_lat_diff
wind_direction_sin
population_density
water_equiv_snowfall_3
water_equiv_snowfall_2
Block
BR_1
Street
wind_direction_cos
wind_direction_cos_1
median_income
wind_direction_cos_3
MD_1
population
water_equiv_snowfall
HZ_1
Heat_3
MD
−0.20
−0.15
−0.10
−0.05
0.00
0.05
SHAP value (impact on model output)
High
Feature value
Low

6. Conclusions and suggestions

Based on the shap analysis on random forest model, we have the following conclusions.

1. A higher number of mosquitos caught yesterday or the presence of the virus yesterday implies the chance of detecting the virus today is also higher.
2. If the humidity is relatively high in the past few days, then the virus is more likely to be present.
3. If there are east winds in the past few days, then the virus is more likely to be present.
4. When temperature is high, the virus is more likely to be present
5. If there was mist yesterday, then the virus is more likely to be present.
6. If the trap catches more CULEX PIPIENS/RESTUANS species, then the virus is more likely to be present.

We can make the following suggestions. First, if we see an increase in numbers for CULEX PIPIENS/RESTUANS species in a region, then we should also increase the testing frequency in the nearby region. Next, if the past few days have high humidity or east wind prevails in the past few days, then we should increase the testing frequency. Thirdly, when we see an increase in Depart, we should increase the testing frequency.