West Nile Virus Prediction: Final Report

Jonathan Jacobs
Springboard Data Science Career Track
November, 2020

TABLE OF CONTENTS

PROBLEM STATEMENT

Chicago's Problem with the West Nile Virus:

According to the CDC, West Nile Virus (WNV) is the leading cause of mosquito-borne disease in the US. The Chicago area reported 6 cases of WNV in the summer of 2020. Although this figure is small, the disease is dangerous, proving fatal for around 1 in 150 people who become infected. Mosquitos, beyond being a nuisance, are a public health concern, and it's important for densely populated urban areas to control the mosquito population.

The primary method of mosquito control is to spray insecticide over large areas of land. Along with environmental costs, there are significant costs and inconveniences associated with controlling mosquito populations. The city of Chicago Department of Public Health treats 40,000 water basins each year with larvicide and monitors 83 traps around the city each week for mosquitos with WNV. It's costly both in terms of time and resources, and yet there are cases of WNV reported every year. Chicago needs methodology that enables the prediction of WNV so that it may eliminate infections entirely without further straining public health resources.

Incorporating Data into Chicago's Response:

Chicago's approach is reactive, responding only after WNV has been identified in an area. Predictive modeling allows the city to preemptively treat areas that are at the highest risk for WNV virus, rather than responding only once WNV has been observed.

Based on weather, mosquito trap, and insecticide spray data, the WNV can be accurately predicted using machine learning. The weather factors that are most strongly associated with the WNV are identified, as well.

Through the incorporation of predictive modeling into the city's plan for the WNV, infections can be brought down to zero over the course of two years.

DATA: SOURCING AND CLEANING

Data for this project comes from the 2015 kaggle competition sponsored by the Robert Wood Johnson Foundation. Data for the competition is provided by the Chicago Department of Public Health.

Weather Data

Sourcing

Local Climatological Data is sourced from the NOAA. There are 1472 unique dates where weather data is collected at two weather stations, representing several years of summer weather — when mosquitoes are active and present. Daily observations report 20 features at weather station 1 and 19 features at station two. More detailed information on each feature can be found here.

Features

Feature	**Description**
Tmax	Maximum temperature recorded that day
Tmin	Minimum temperature recorded that day
Tavg	(Tmax+Tmin)/2 — the average of the maximum and minimum temperature
Depart	The number of degrees F off of the historical average temperature
Dew Point	The dew point
Wet Bulb	The wet bulb thermometer is a thermometer wrapped in a damp cloth, and is often cooler than Tavg due to evaporative cooling.
Heat	When Tavg < 65, Heat = 65-Tavg, otherwise 0
Cool	When Tavg > 65, Cool = Tavg-65, otherwise 0
Sunrise	The time of sunrise
Sunset	The time of sunset

Code Sum	Codes are recordings of weather events, for example fog, thunderstorm, or tornado
Depth	Depth of snow on the ground
Water1	Water equivalent of snowfall
Snowfall	Snowfall in inches
Precip Total	Precipitation total for the day
Stn Pressure	Pressure at the weather station
Sea Level	Pressure at sea level
Result Speed	Average wind speed, with direction incorporated
Result Dir	Average wind direction
Avg Speed	Average wind speed, regardless of direction

Removing null or sparse features

Three features related to snowfall: Snowfall, Depth, and Water. Since all of the weather data is from summer months, there were only 12 non-zero entries out of the 1472 observations. For this reason, I remove all three features.

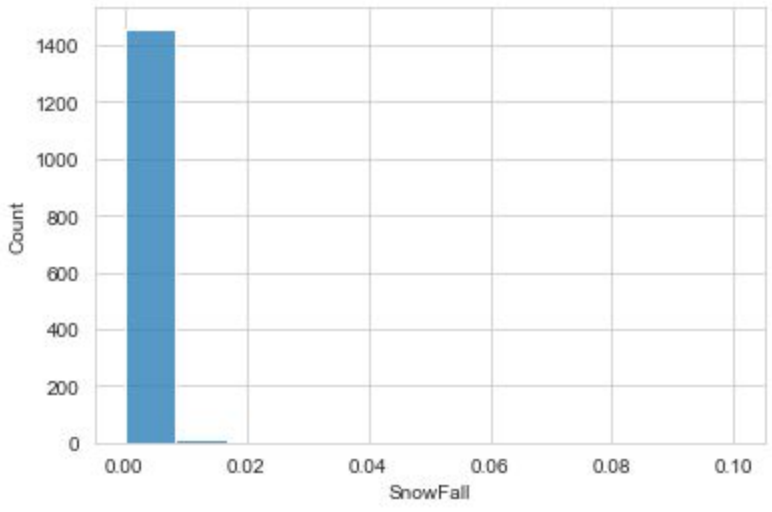

Figure 1: SnowFall Histogram

Filling missing data

Other than the three missing features, there were only 11 missing data points. The missing data was handled in two ways:

1. Fill with zero. For precipitation, I assumed a missing value was equal to zero, consistent with snowfall reporting. I found no indication that extensive rain would have caused equipment malfunction, or any other reason to believe that missing data could have been because of too much rain, rather than none at all.
2. Forward fill. For other features with missing data, there were no concurrent omissions, which is an indication that at no point was a weather station down or out of service. For individual features, for example Tavg, missing data is random and does not appear sequential. I use forward fill because adjacent days are a better approximation than the median or mean. The plot below is Tavg for one weather station over a short period of time where one missing value is. It's clear that the median, the orange line, does approximate the missing segment well.

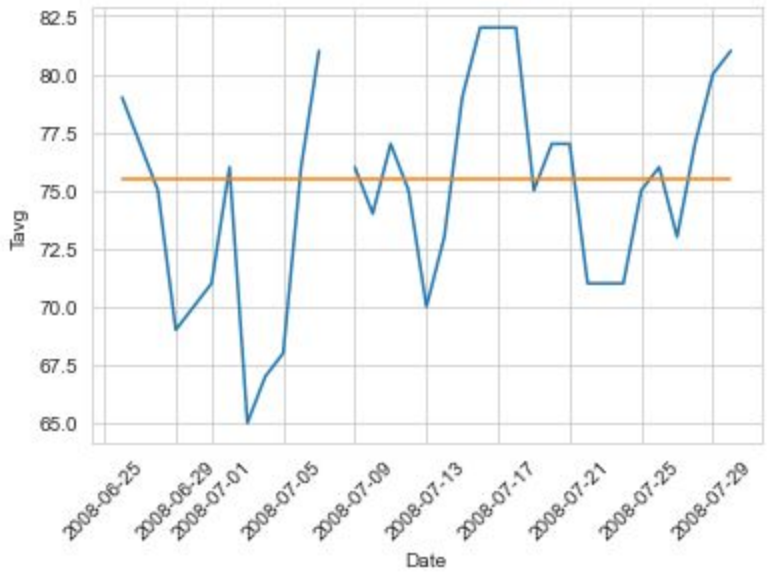

Figure 2: Tavg over a short period of time

Handling two weather stations

There are two weather observations for each day, one for each weather station. I separate each station and re-merge on data. This yields one observation for each day, but a

considerable number of features that are collinear. The weather at each station, although statistically different, is not practically very different.

The chart below shows the difference of the two weather stations as a percentage of the overall mean for the feature. Most are below a 0.2% difference, with only one feature above 0.4% difference. Practically, this difference is not significant, and for that reason, I average the two weather stations into one.

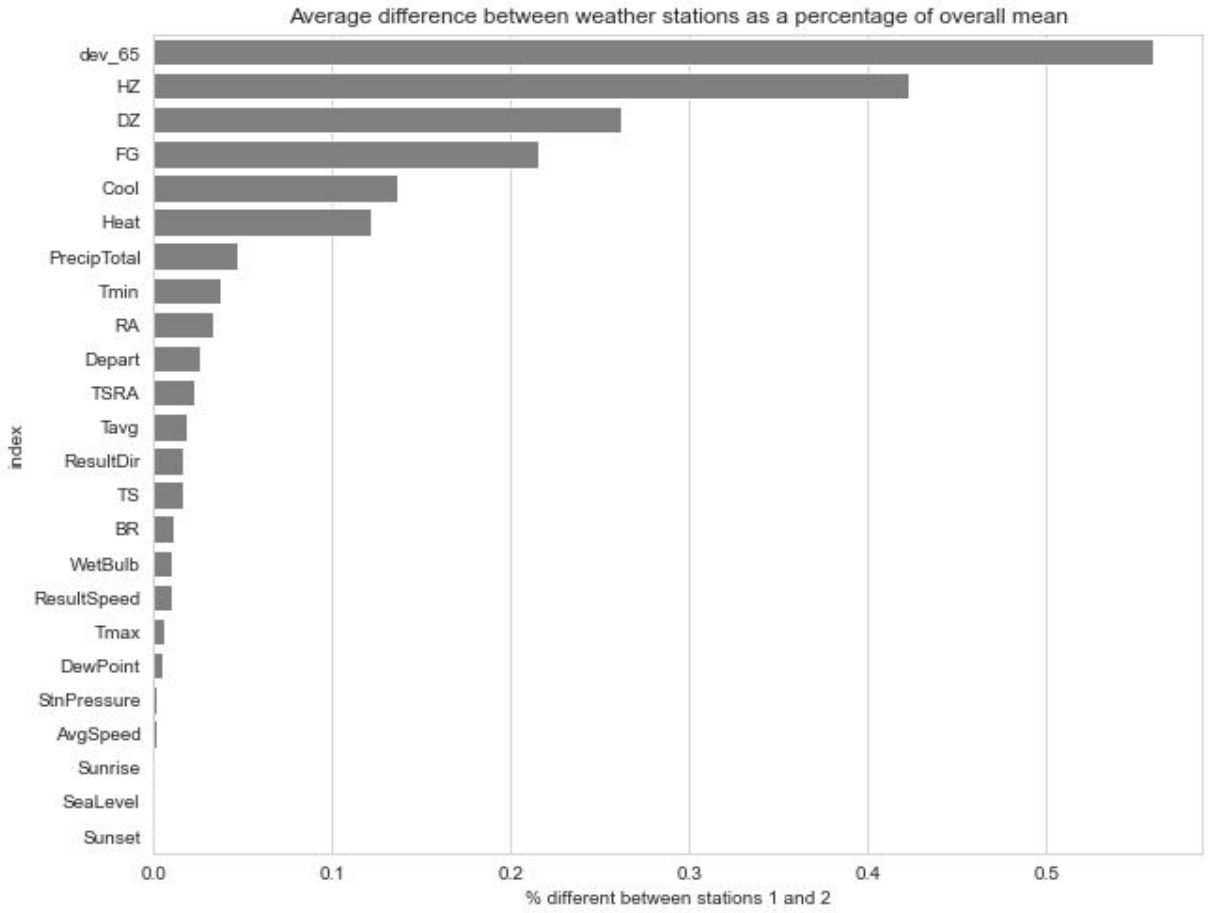

Figure 3: Average difference between the weather stations as a percentage of overall mean

Spray Data

The spray data set was missing no data and required no cleaning. It provided the date, time, and location of each spray. The map below plots all spray locations, which goes back several years. Spraying occurs primarily near large water basins.

Mosquito Trap Data

The City of Chicago Department of Public Health monitors 84 mosquito traps weekly. Below on the left is a map of the trap locations, and on the right locations of spraying. The mosquito data, like the spray data, was clean and simple. It provides

- The date the trap was checked
- Trap type
- Trap location
- Number of mosquitos found
- WNV present

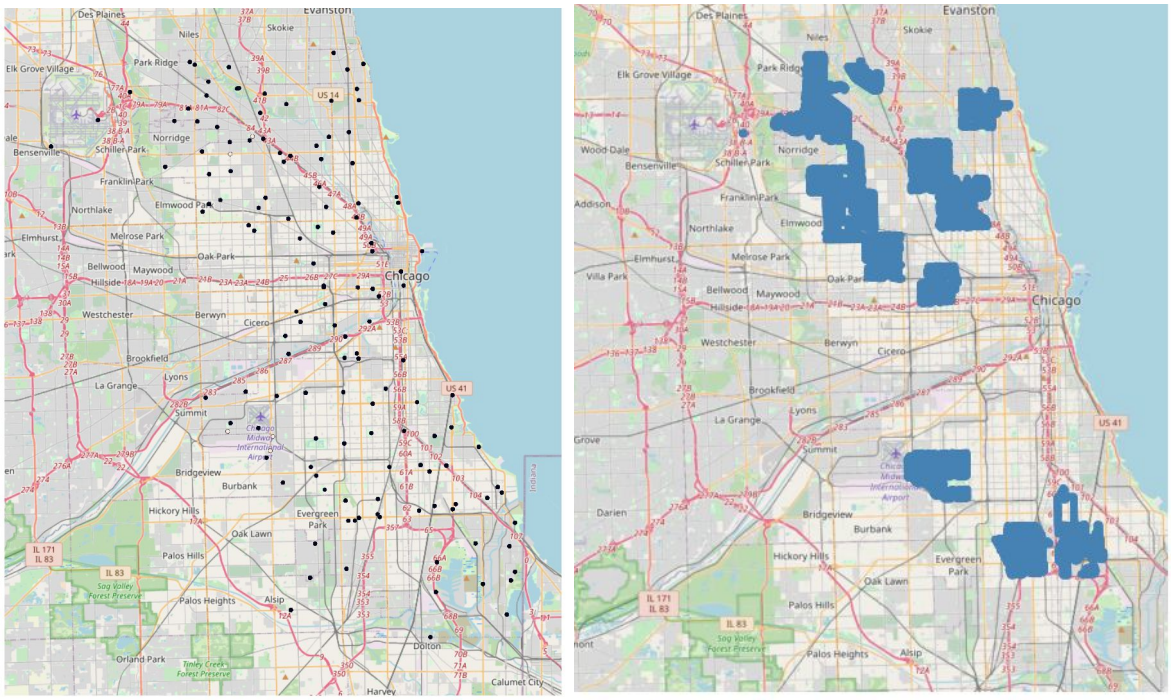

Figure 4: left: trap locations, right: spray locations

In order to join the spray data to the larger mosquito monitoring data set, I used the trap locations and spray locations to see if there had been an insecticide spray within one mile of the trap on any particular day, month, or year. One mile is used, because that is the farthest observed distance a mosquito of the species that carries WNV has been recorded traveling in its lifespan. Later in the analysis, I find that an exponentially weighted moving average of the 'day' metric was the most valuable for prediction.

Uniting the data sets

In order to unite the mosquito/spray data set, which contains the target variable WNV present, with the weather data, I merge the weather data onto the mosquito/spray data set. For each trap observation there is then information about spray occurrences and weather based on date and location.

EXPLORATORY DATA ANALYSIS AND FEATURE ENGINEERING

Mosquito Species

It is well documented that only two highly related mosquito species, Pipiens and Restuans, can carry and transmit the WNV. Unfortunately, in the distribution of trapped mosquitoes, they are the vast majority of mosquitos found in the Chicago area. In other words, only Pipiens or Restuans can carry the WNV, but the vast majority do not, and since few mosquitos trapped are not of these species, no locations or situations can be discounted because of the species observed. Any consistent observation of different species is likely due to sampling error, and drawing conclusions or making predictions based on that is tenuous at best. For the analysis, I create a new feature that is a binary indicator of mosquitos observed, where 1 is positive for Pipiens/Restuans, and 0 is other.

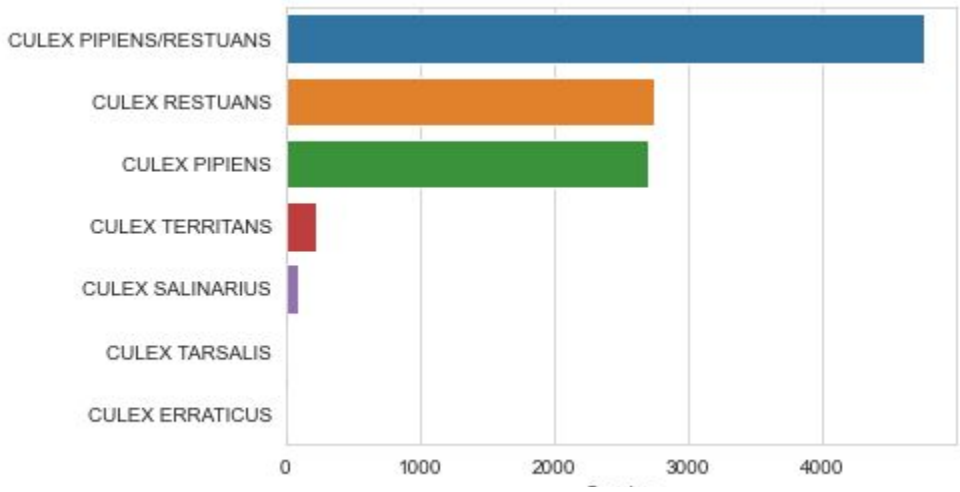

Figure 5: Mosquito frequency distribution

Lagged Features

The effects of many weather phenomena are not limited to the day they occur. Rain on one day, for example, can create stagnant water for the following few days, or possibly even more. The effects of high heat, sustained winds, and others can also similarly have effects that go beyond their immediate happening. And finally, insecticide spray has such an effect as well

For this reason, I lag all relevant weather and spray features using two strategies and at a variety of time intervals.

1. Uniform moving average. All observations for n days prior to the current day and including the current day are uniformly weighted and averaged. n ranges from [2,8) with step 1, [8,18) with step 2, and [18,60) with step 4.
2. Exponentially weighted moving average (ewm). All observations with a certain span are weighted exponentially, so more recent observations have more influence in the current state. Span ranges from [2,8) with step 1, [8,18) with step 2, and [18,60) with step 4.

All lagged features, over 1000 in total, are incorporated into the existing DataFrame.

Paired Feature Interactions and Feature Distributions

Even without the lagged features, there are 31 features to plot pairwise, which is an unreasonable amount of information to digest. I subset the data with hypothesis driven questions to create more manageable plots and relationships.

Pair plots with larger image sizes for legibility, as well as with groupings incorporated can be found here, in the plot folder of my github repository.

Pair Plots for temperature are shown below. As expected there are strong linear relationships between temperature-related variables. In order to guard against multicollinearity, most of these features will need to be removed or combined, which I accomplish later in the process with information value assessment and variance inflation factor calculations.

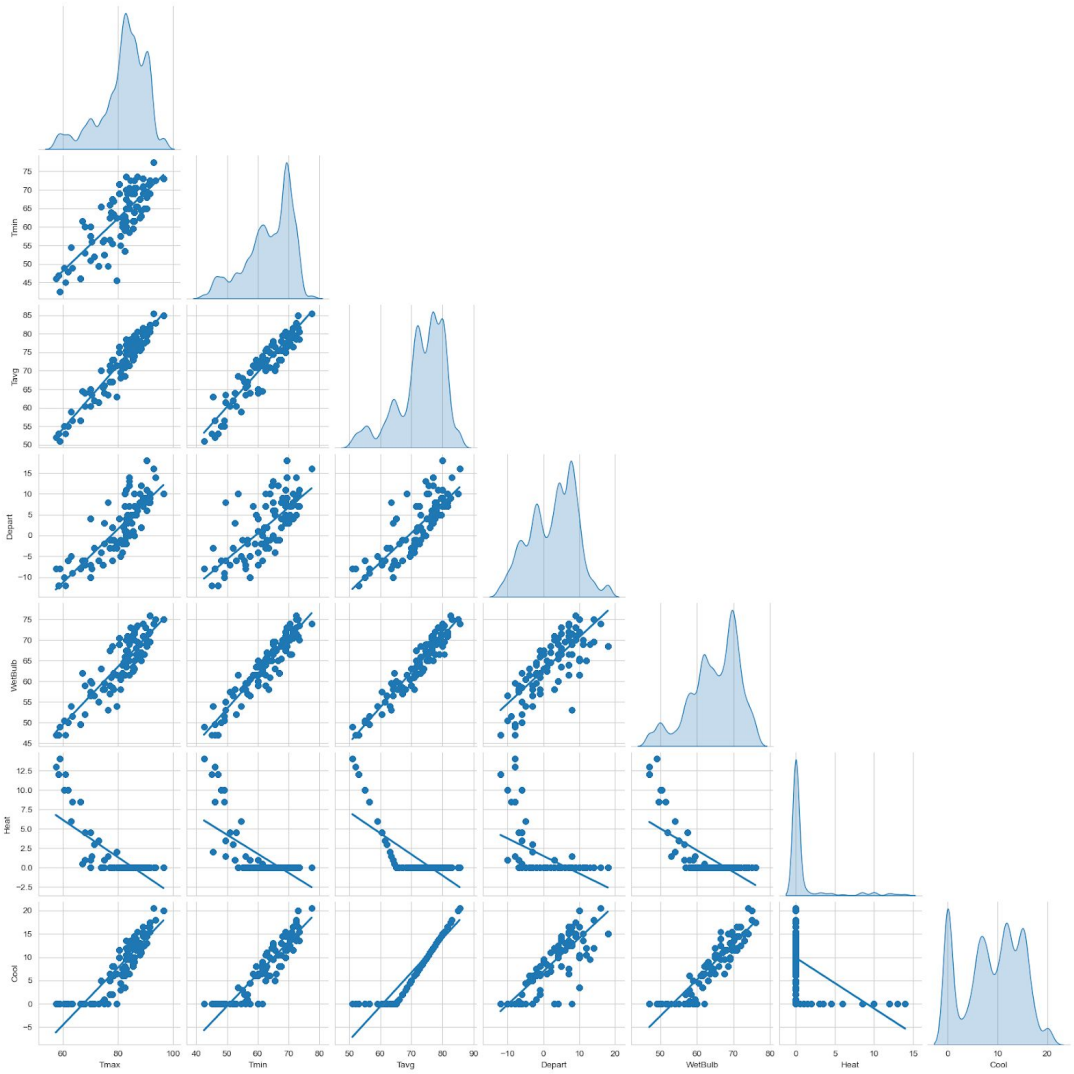

Figure 6: Temperature feature pairplots, regressions, and distributions

The plot below shows relationships between several other features in the weather set. Again, there are highly collinear relationships which are addressed later in the process. There are also some variables, such as wind speed and precipitation, and pressure measures, which are less related, a good sign for feature engineering.

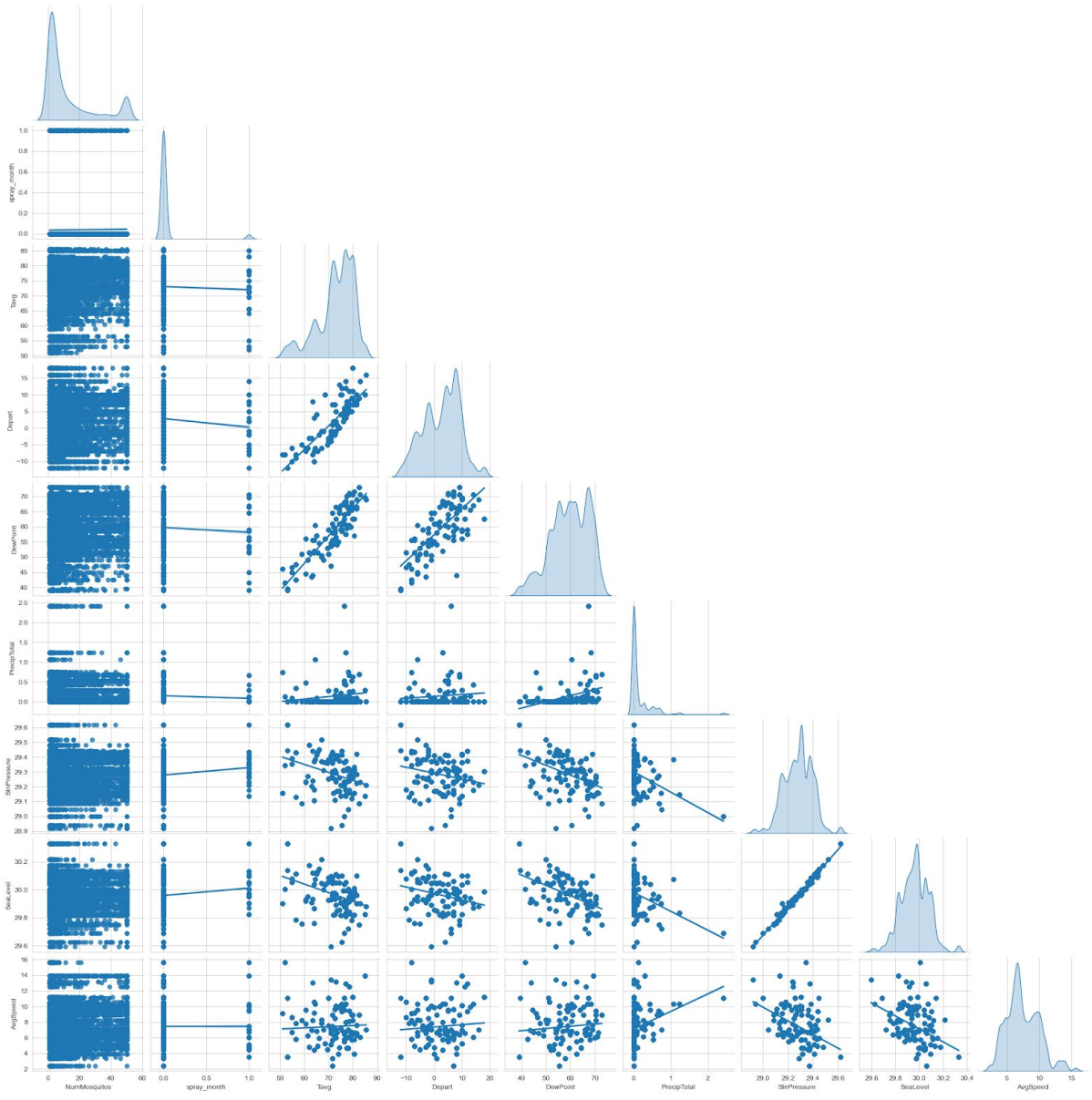

Figure 7: Various feature pairplots, regression, and distributions

Lastly, the pairplot of final features, which can again be seen in greater details in the plot folder of my github repository, reveals little to no clean feature interaction. Features appear to be independent and correct for predictive modeling. None of the final features have strong linear relationships to WnvPresent, the target variable, which is a good sign as well. To high of a correlation could indicate sampling error and hinder predictive modeling.

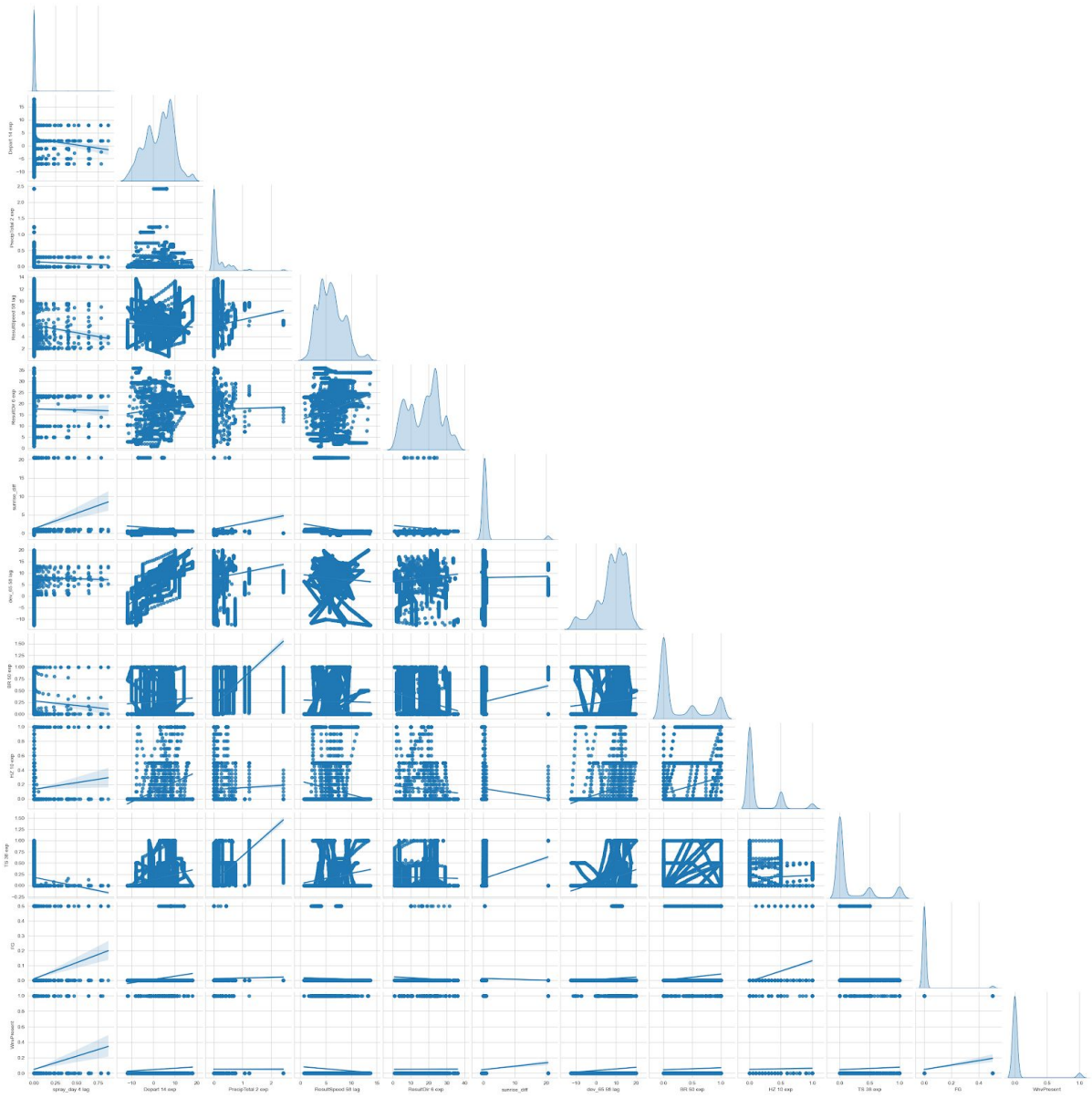

Figure 8: Final feature pairplots, regression, and distributions

Information Value Analysis

In total, 1143 features were engineered for the purpose of this analysis. Many of those features are time-lagged variations on one another. I first reduce the number of features through an information value (IV) analysis. Broadly, IV is a measure of the predictive power of an independent feature and the target feature. An IV of less than 0.01 or greater than 0.8 indicates incredibly poor predictive power or suspiciously high predictive value, respectively, and those features are dropped.

After this initial IV sorting, 474 features are left, many of which are near repeats, for example average temperature lagged for 47 days and lagged for 46 days. For each variable, the lag or iteration with the highest IV is kept and the others are dropped.

At the end of the analysis, 17 features remain. Sunrise and sunset times are suspiciously good predictors, and they are directly related, measuring the same phenomena. Variance inflation factors reduce the feature set even more.

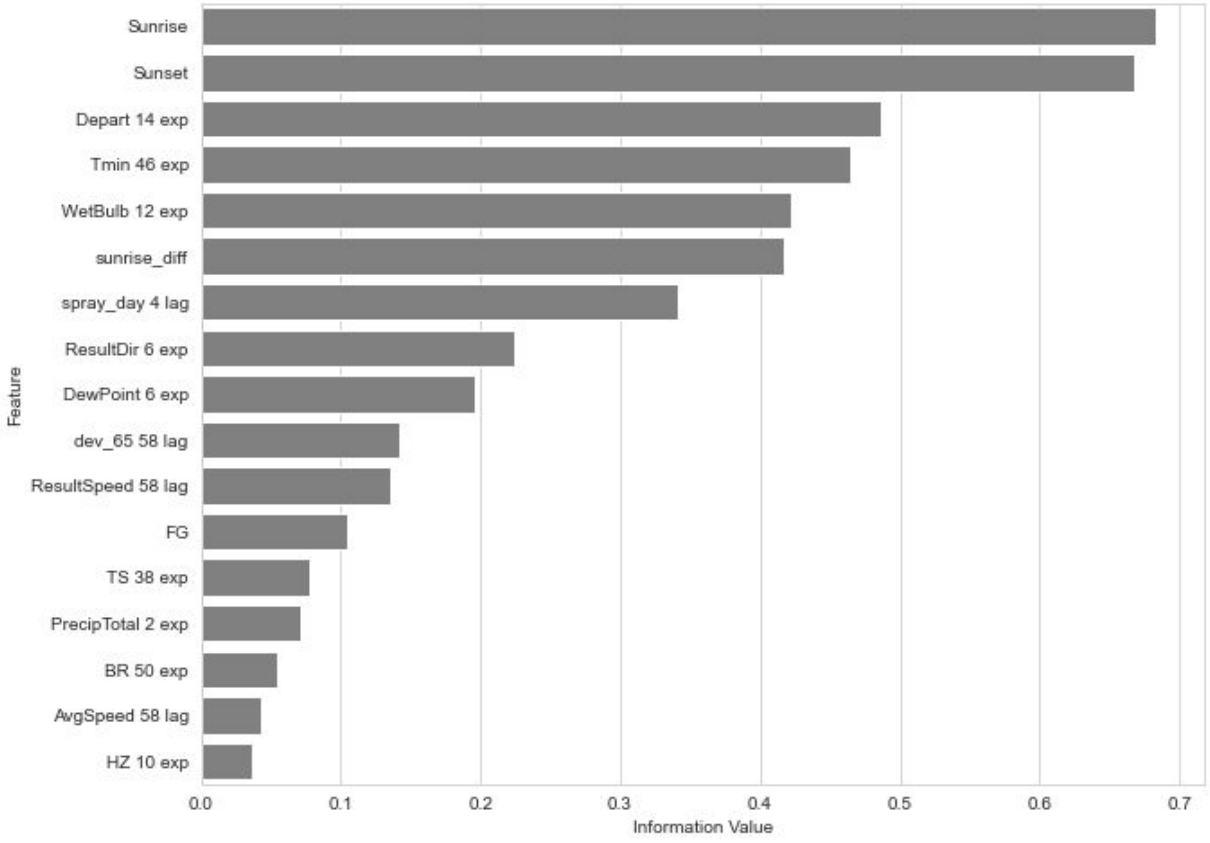

Figure 9: Final feature pairplots, regression, and distributions

Variance Inflation Factor

Variance inflation factor (VIF) is a measure of multicollinearity between features. It's a more robust measure than covariance alone for removing highly-related features, because it quantifies variance relationships between one and multiple variables rather than one on one relationships, which can bury more complex relationships between features. The procedure for dropping features based on high VIF is to 1. calculate the VIF for all features, 2. drop the feature with the highest VIF, and repeat the process until the maximum VIF in the feature space is under a set threshold. Thresholds are commonly in the range of 5-10, and for this analysis the threshold is 6.

Below are the final features and their VIFs after 7 iterations and 6 variables dropped.

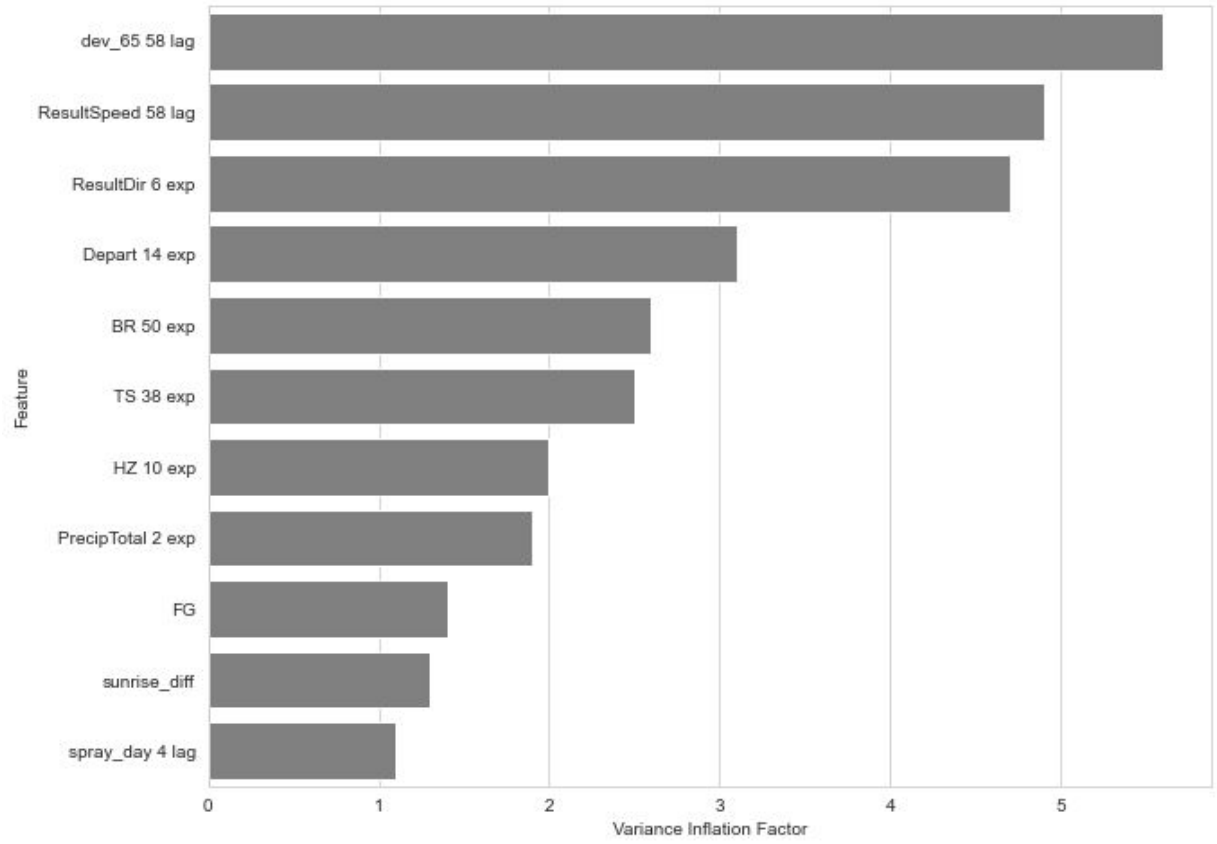

Figure 10: Variance inflation factors for final features

MACHINE LEARNING ANALYSIS

Train/Test Split and Cross Validation

Prior to any model training or analysis, the data set is split into training and test sets. The test set guards against overfitting a classifier to the particular data set, and is used to assess the models ability to generalize. The test set is 20% of the total data, randomly selected, while the training set is the remaining 80%.

For all model training that follows, five folds of cross validation are used. Cross validation folds further segment the training set into n subsets. The classifier is trained on four of the subsets and tested on the remaining set, and then the process is repeated so that all five subsets act as the test set for a classifier trained on the other four. The average score across all folds is used to assess to overall performance of the classifier, rather than any individual fit.

Selecting a Classifier

With a robust set of features prepared for machine learning (ML) analysis, different classifiers are now explored and assessed. The scoring function for the analysis and assessment is the area under the receiver operating curve (AUC), which is robust against class-imbalanced data sets, which the WNV set certainly is. Other points of consideration are interpretability and transparency. Since the primary stakeholders in this model building are not necessarily technical professionals, the model needs to be readily explainable, and feature importance needs to be so as well.

I explore four classifiers — random forest classification, k-nearest neighbor, C-support vector classification, and logistic regression — with a broad random search of hyperparameters to give a general sense of performance. The random search cross validation is not intended to be exhaustive, but to instead narrow down the hyper-parameter range for further tuning, and to offer a general assessment of the classifier's performance. Only the random forest and k-nearest neighbor classifiers, the best performers by a large margin, are discussed below. More detailed results for all classifiers can be found here.

A Random Forest Classifier with 1050 estimators and a max tree depth of 9 is the final selection. The final AUC is around 0.80.

Random Forest

The Random Forest (RF) classifier is an ensemble method which combines a large number of naive decision makers to make one final classification. Probabilities for class membership are

calculated based on the number of estimators that decide a positive class divided by the total number of classifiers.

Search Parameter	**Value**
Number of estimators	[50,2000)
Max tree depth	[2,20)
Iterations	10
Cross validation folds	5
Total fits	50
Best AUC training sets	0.80
Test AUC score	0.82

As shown in the figure below, the optimal parameter space is clear for the max tree depth, at around 8-11. The optimal number of estimators, however, is not clear, with equally good scores from 245 to almost 1750 estimators.

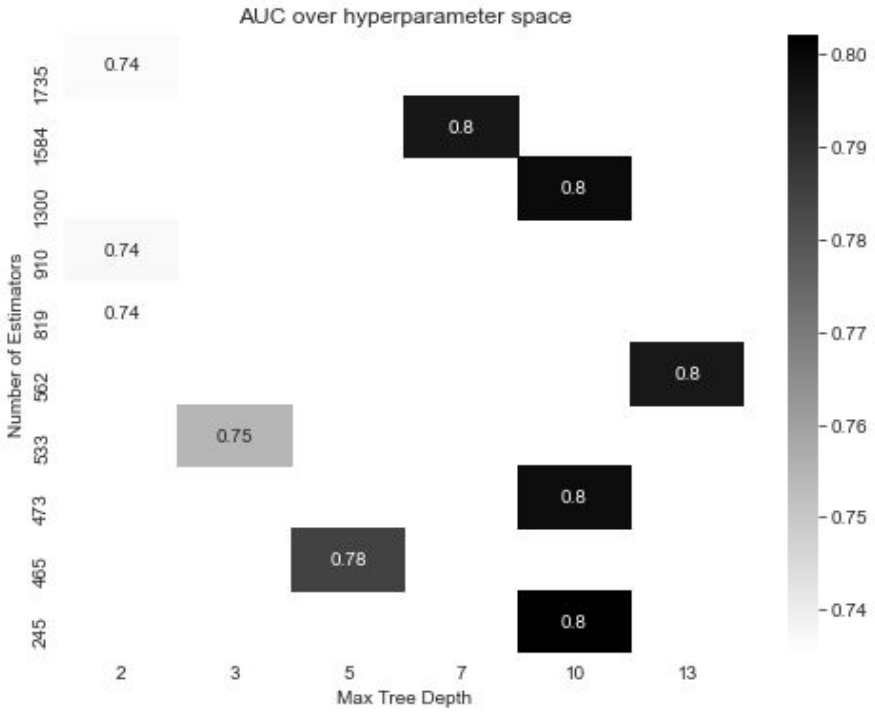

Figure 11: AUC scores for different parameter combinations

The receiver operating characteristic curve (ROC) is shown below for the best estimator of the above fits

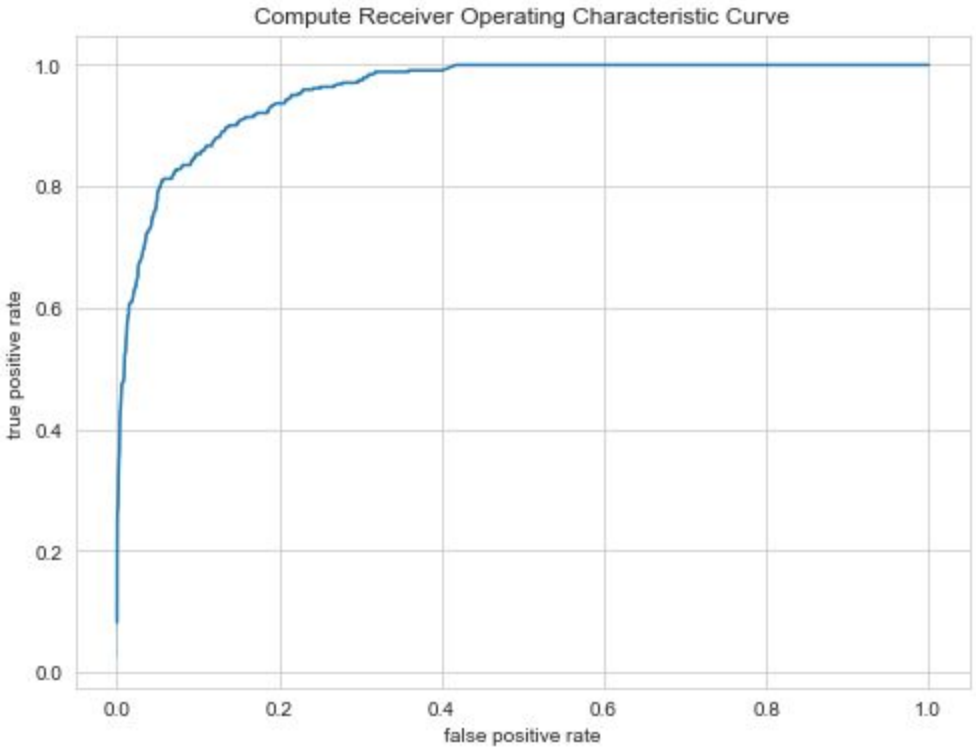

Figure 12: ROC curve for the RF classifier

K-nearest neighbors

K-nearest neighbor (KNN) classifiers make class membership decisions based on proximity in the feature space. Due to random initialization of class boundaries, there is the potential for newly trained models to make different class decisions based on its initial state.

The parameter to tune is the number of neighbors to take into account in the feature space to decide class membership. KNN is fast to train in comparison to RF, so considerably more interactions were run

Search Parameter	**Value**
Number of neighbors	[2,200]
Iterations	200

Cross validation folds	5
Total fits	1000
Best AUC training sets	0.81
Test AUC score	0.81

The KNN classifier performs almost exactly as well as the RF classifier. The optimal number of neighbors was found to be 72. The plot below shows that there are large returns in AUC score for increasing the number of neighbors at first, and then diminishing returns afterwards.

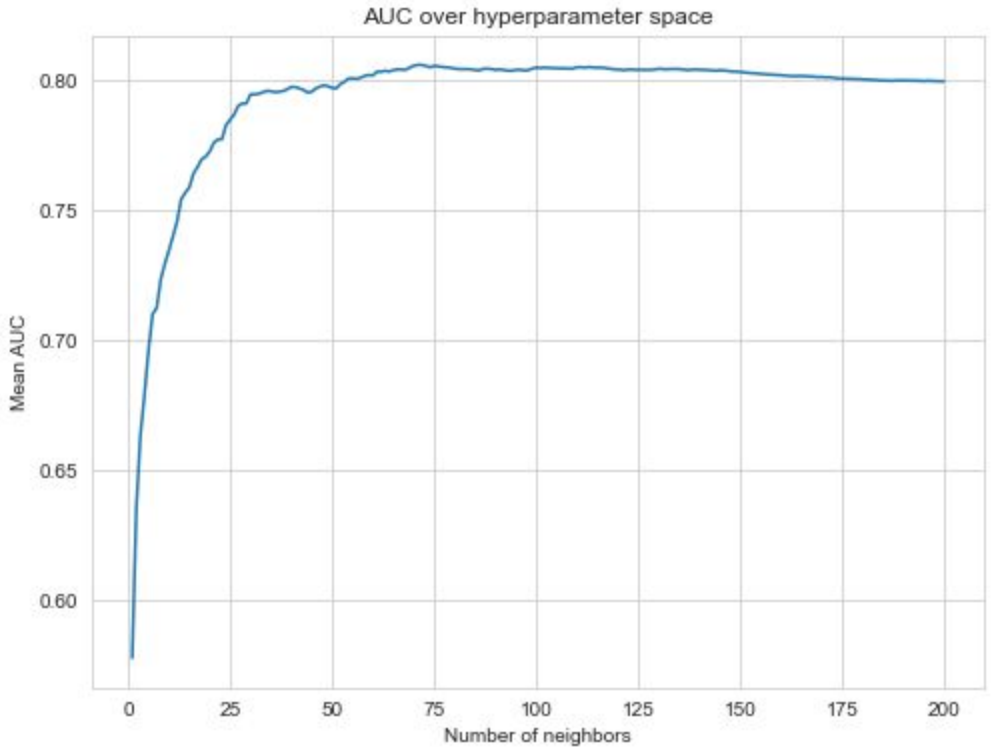

Figure 13: AUC scores at different numbers of neighbors

The ROC curve with 72 neighbors is shown below. Although the area under the curve is almost exactly equal to that of the ROC curve for the RF classifier, we see in the plot that a true positive rate of 1 is only reached when the false positive rate is 1 as well. The true positive rate starts much higher than in the RF curve, but it gains slowly regardless of the threshold. The RF curve, by contrast, reaches a peak quickly.

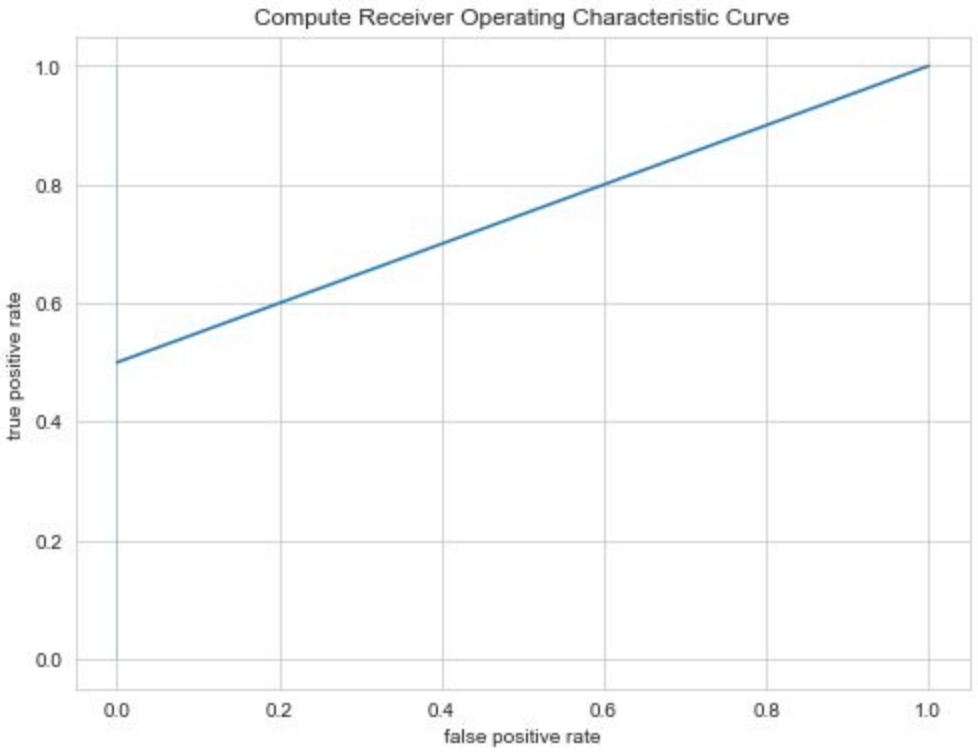

Figure 14: ROC curve for the KNN classifier

Classifier Choice

Two classifiers perform well enough to be considered for further exploration and use in the predictive modeling, the RF and KNN classifiers. Based on the AUC scores aline, performance is nearly equal across the two. Yet, the ROC curves tell a different story. Although the RF classifier starts with a low true positive rate, its peak is reached quickly and at a lower false positive rate than the KNN classifier. Below, in Figure 15, the curves are overlaid. Clearly, the RF classifier is the better performer of the two.

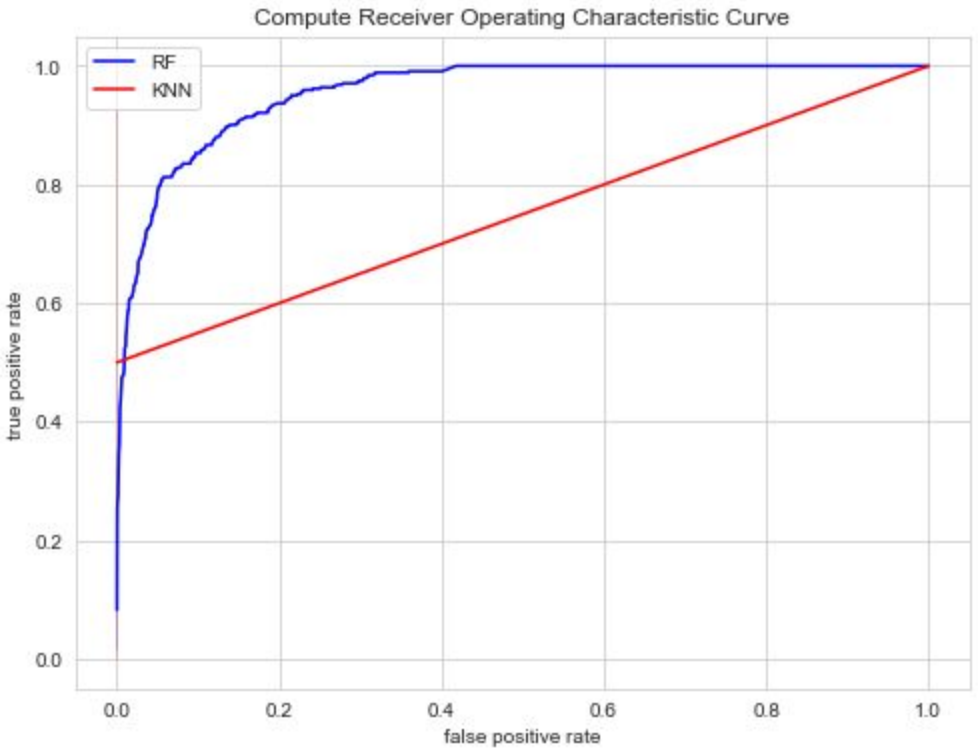

Figure 15: ROC curves for RF and KNN

Random Forest Classifier Training and Performance

The final RF classifier is trained with a grid search over a narrowed parameter space based on the results of the initial random search.

Search Parameter	**Value**
Number of estimators	[750,2000) step of 50
Max tree depth	[6,11)
Cross validation folds	5
Total fits	625
Best AUC training sets	0.80
Test AUC score	0.82
Optimal number of estimators	1050

Optimal max tree depth	9

The figure below shows the grid search and resultant AUC scores. Once again we see high scores distributed across the number of estimators space, but clearly a max depth of 9 or 10 is the optimal choice. The max depth is limited to 10 in order to protect against overfitting.

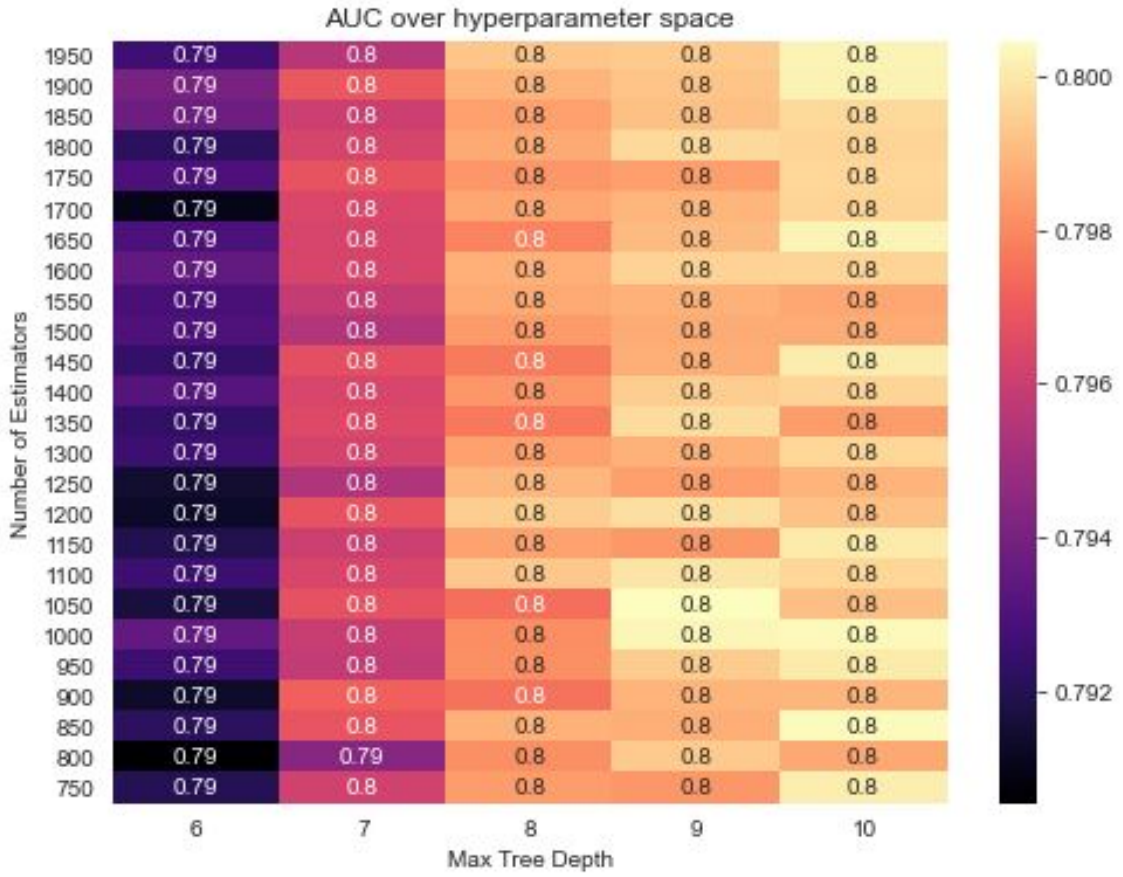

Figure 16: AUC scores for different parameter combinations

The ROC and the Precision-Recall curves are shown below for the best estimator, 1050 estimators, and max depth of 9.

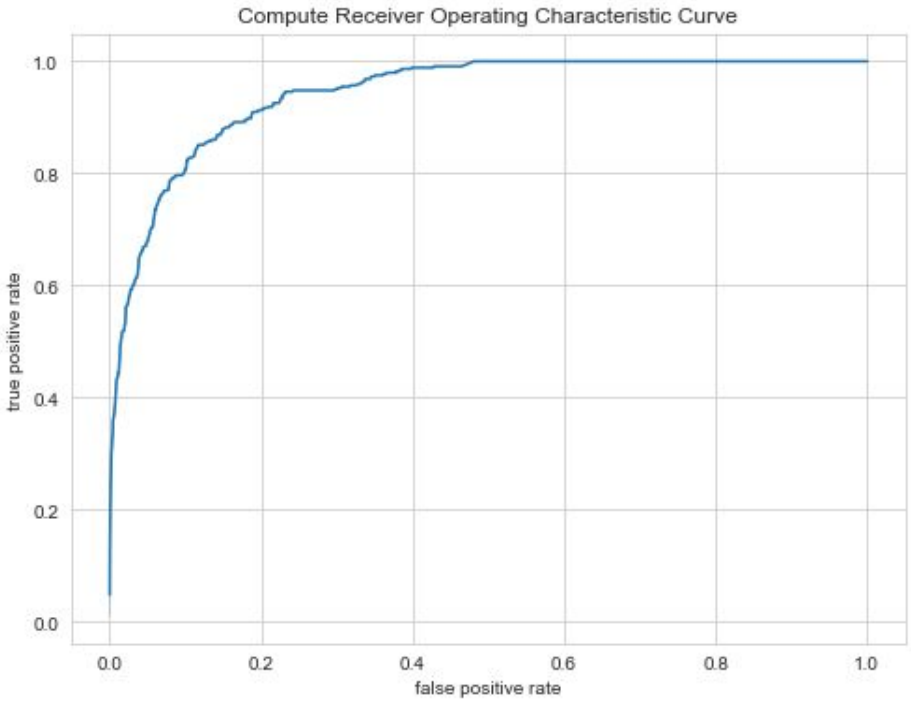

Figure 17: ROC curve for the RF classifier

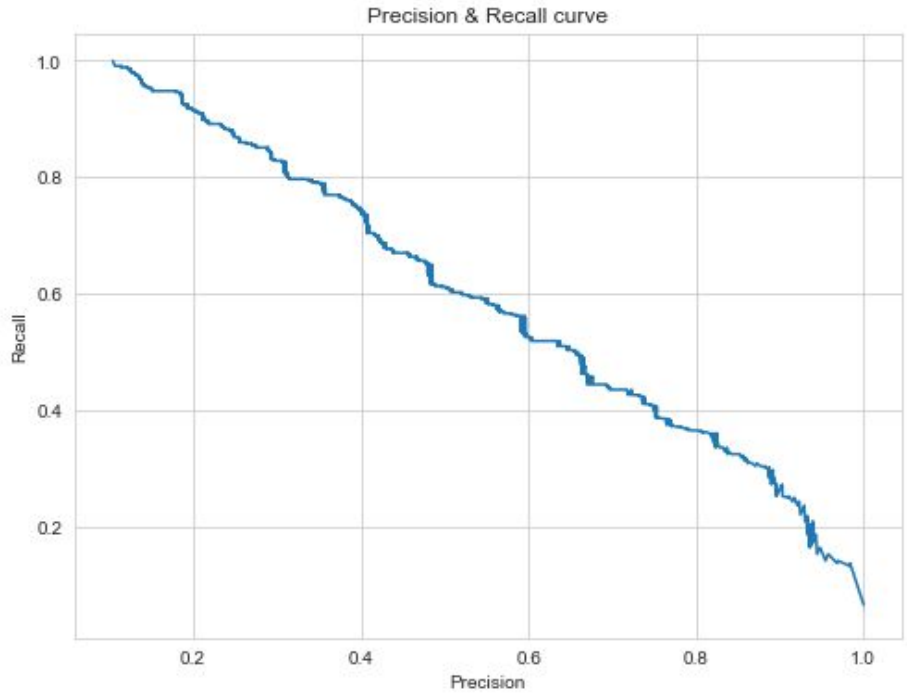

Figure 18: Precision and Recall curve for the RF classifier

Feature Importances and SHAP Analysis

SHapley Additive exPlanations (SHAP) is a method introduced by Lundberg and Lee in 2016 to explain individual predictions. It's based on Shapley Values, a mathematical concept in game theory which quantifies the magnitude and direction of individual contributions to a game's outcome. Dr. Dataman, a contributor to *Towards Data Science,* a publication on medium.com, uses the analogy of several individuals driving a stake into the ground. Shapley values quantify how many inches each individual drove the stake into the ground.

SHAP's Tree Explainer allows for the explanations of both individual decisions as well as aggregate decision making. Figure 19, below, shows average SHAP value magnitudes for each feature. Precipitation, and then two temperature features have, overall, the greatest contribution to individual decisions of class membership made by the RF classifier.

SHAP dependance plots show how contributions vary for each feature based on its value.

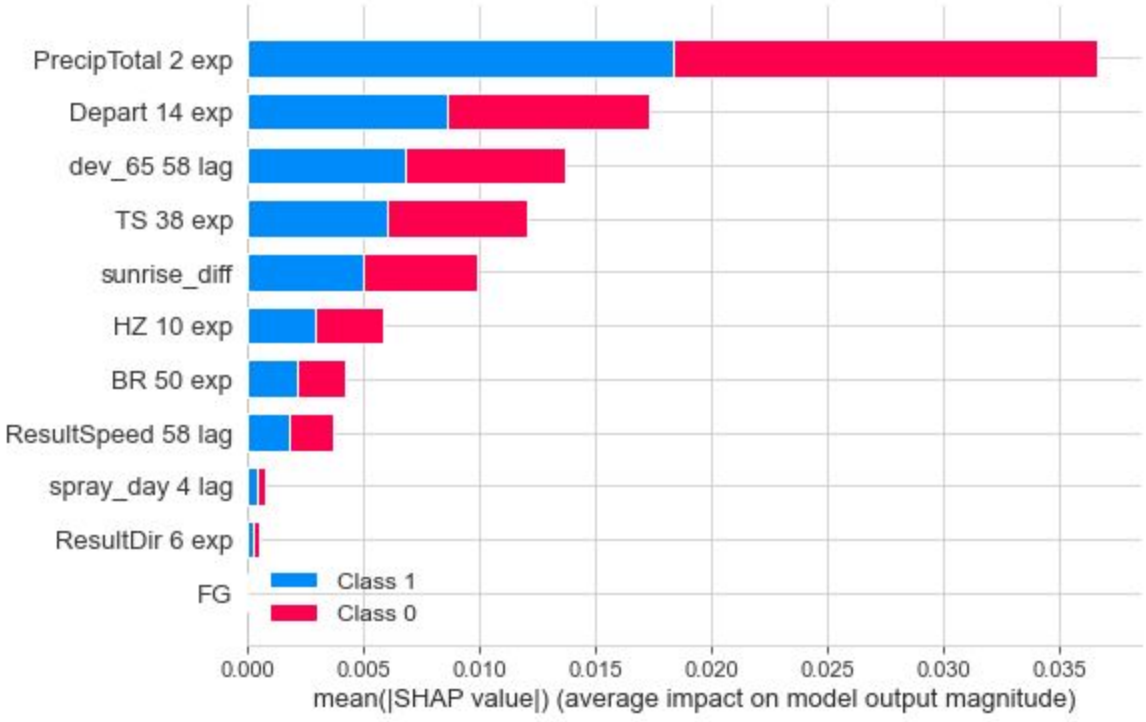

Figure 19: Average SHAP value magnitudes for each feature

Precipitation

PrecipTotal

PrecipTotal 2 exp is a 2-span exponentially weighted moving average of the daily precipitation. Here, there is a negative trend. When there is more precipitation in a two day span, we expect to see less WNV. This is a surprising result, as it's well known that stagnant water is necessary for mosquito populations. The short timescale likely hides this relationship.

In an article published in 2004 by the *International Journal of Biometeorology,* Arthur T. DeGaetano finds that individual heavy rainfall events tend to reduce trap counts, while high monthly precipitation is associated with more trapped mosquitos overall. The overall effect of high precipitation events is not well understood, even in more recent literature, yet the association to mosquito trap numbers is consistently negative.

Consistency with other research findings confirms that the 2-day precipitation feature is an important indicator of WNV presence, if only because mosquitos are less likely to fly and get trapped or bite people.

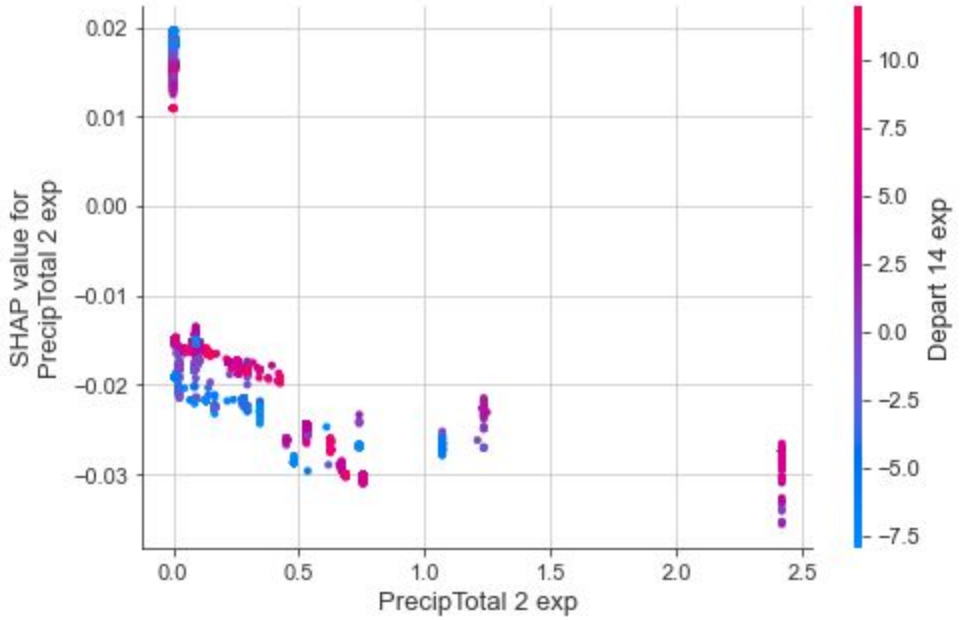

Figure 20: SHAP dependance plot for PrecipTotal 2 exp

TS 38 exp

TS 38 exp is a 38-span exponentially weighted average of thunderstorm occurrences. As opposed to precipitation, we see in the dependence plot below that more thunderstorms are

clearly associated with more WNV sightings. At a longer time-span, this confirms the relationship of stagnant water to mosquito populations. And this finding is consistent with DeGaetano's finding that monthly precipitation is positively associated with mosquito trapping.

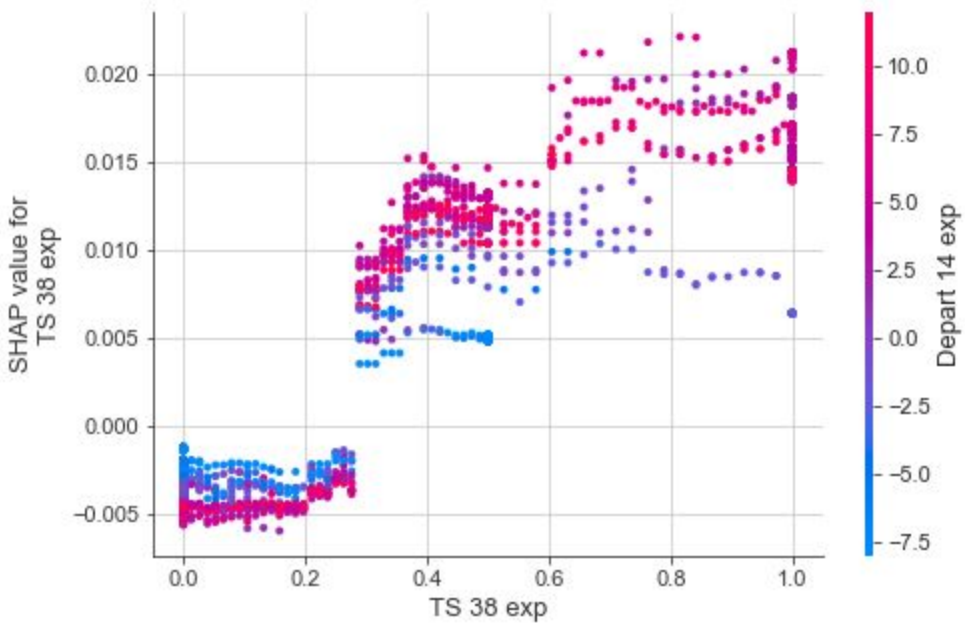

Figure 21: SHAP dependance plot for TS 38 exp

Temperature: Departure and Deviance

On a given day, departure is the difference between that day's average temperature and the historical temperature average for that day. Day-to-day, year-to-year, the base temperature shifts. Long measures of departure are useful in tracking climatological shifts like climate change.

Deviance is similar to departure, except the base temperature is uniformly 65F. It's a measure commonly used in the energy industry to predict heating and cooling loads on the grid. Typically, if a day's average temperature is below 65F, the metric is labeled Heat and calculated as 65F - Tavg, otherwise 0. If Tavg is greater than 65, the metric is labeled Cool and calculated as Tavg - 65F, otherwise 0. I combine Heat and Cool into one metric that tracks the deviance from 65F on any given day.

These temperature metrics, despite seeming similar, do measure different phenomena. Departure is a climatological measurement, relating to and quantifying natural variation in temperature,, while deviance is a social construct, measuring perceptual 'hotness' and

‘coolness’ as we define it in our cities and urban landscapes. Additionally, the features are lagged with different strategies and timespans.

Depart 14 exp is a 14-span exponentially weighted moving average of daily departure measures. Dev_65 58 lag is a 58 day trailing uniform average of deviance.

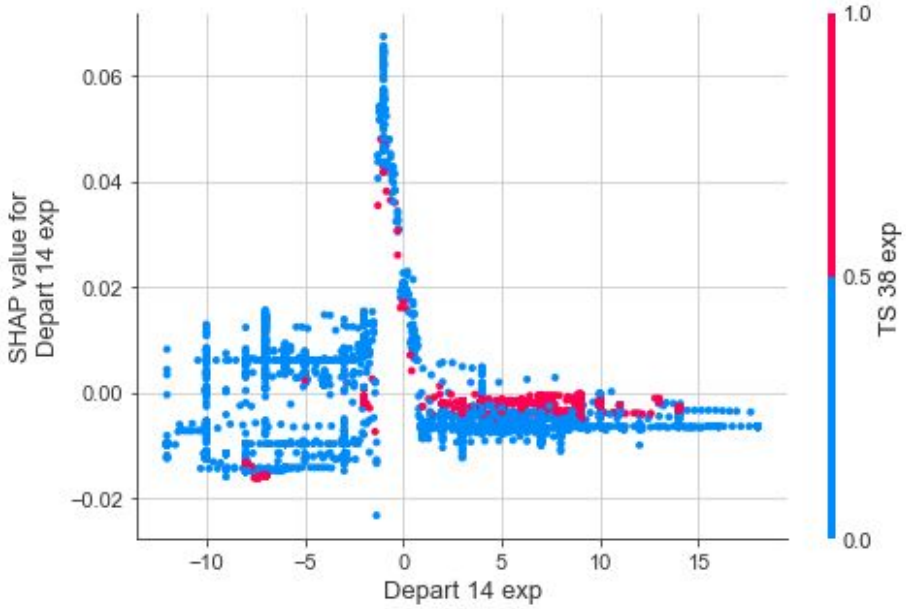

Figure 22: SHAP dependance plot for Depart 14 exp

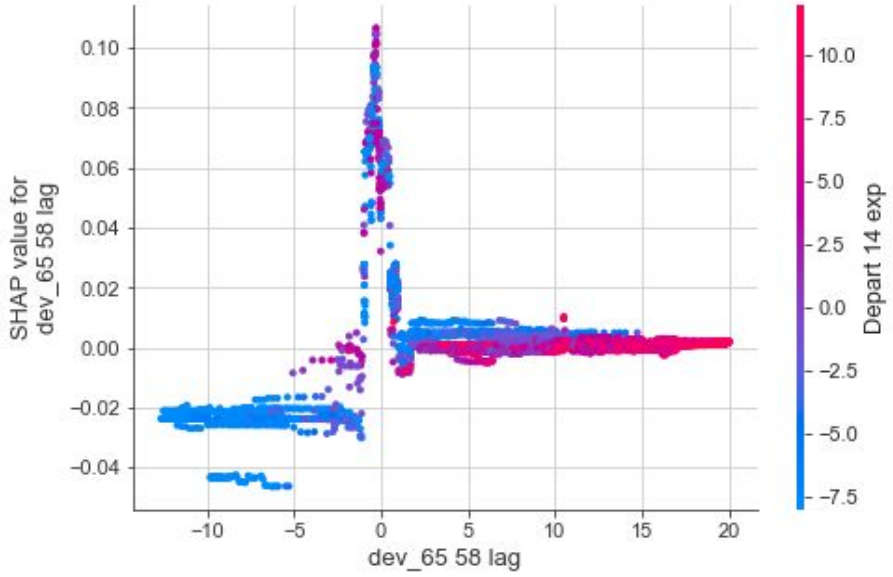

Figure 23: SHAP dependance plot for dev_65 58 lag

As we see in figures 22 and 23, both features strongly positively contribute to WNV present when they are near 0. This is an indication that average temperatures around 65F are the most conducive to WNV sightings and trappings.

For depart 14 exp, lower temperatures over two weeks with exponential decay do not clearly lead to a positive or negative SHAP contribution. Higher than average two week temperatures are associated with a slight negative SHAP contribution. DeGaetano notes that above average temperatures contribute to mosquito mortality.

Average deviance from 65F averaged over two months offers similar insights. Temperatures more than a few degrees colder than 65F for long periods of time uniformly negatively contribute to the SHAP value, and by extension are negatively associated with WNV presence. Average temperatures over two months more than a few degrees warmer than 65 appear to have a slight positive SHAP value, but have 0 association within the margin of error. This indicates that sustained higher-than-average temperatures have little association to WNV.

Insecticide Spraying

Insecticide spraying is the primary method for controlling mosquito populations. The CDPH treats 40,000 each year to hinder mosquito larvae growth, and downstream, reduce the number of mosquitoes in the air.
Spray_day 4 lag is a 4 day uniformly weighted moving average of spray occurrences within one mile of a mosquito trap. The distance is one mile because that is the furthest recorded distance mosquitoes of the correct species can travel in their lifetime, so presumably the effect of the spray can extend up to one mile away from the actual spray location.

The dependence plot, below, is non-monotonic. The moving average ranges from 0 to 1, where 1 indicates four sprays within four days. A great deal of spray activity 0.5 - 1 has little consistency with positive or negative SHAP values. Fewer sprays in four days between approximately 0.15 and 0.4, have positive SHAP values of low to moderate magnitude. Deviation, by comparison, peaks over 0.1, over ten times the indicative magnitude. And finally, low spray values, < 0.15 appear to have low SHAP values on average, and a negative association with WNV.

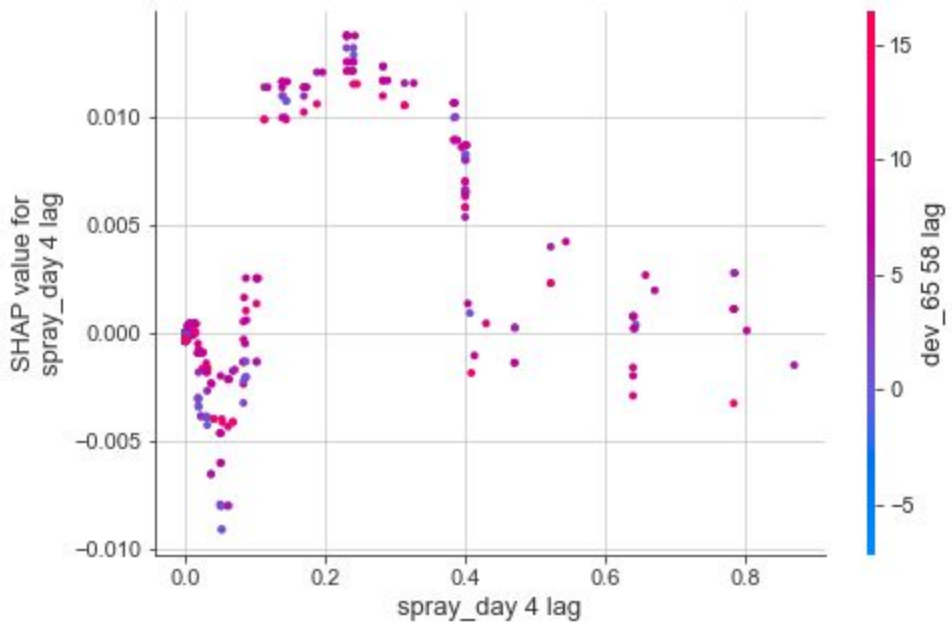

Figure 24: SHAP dependance plot for spray_day 4 lag

This relationship follows expectations about how spraying is related to WNV. The middle spray values, which have positive SHAP values, could result from any number of situations, but there is one I would like to highlight and discuss. Mosquito sprays are done in response to large mosquito populations, not preventatively. A moving average with window 4 will capture either the initial stages or the tail end stages of the city's response. The spray feature is a proxy measure of mosquito populations as a whole. In the middle of a spraying campaign, at the high end of the spray_day 4 lag range, mosquito populations likely experience a sharp decline, fewer are trapped, and we expect to see less WNV. Similarly, at the low range for the feature it's likely an area whose mosquito population is not large or concerning enough to attract the attention of public health departments, and subsequently insecticide spraying.

This complex relationship between mosquito populations, public health official's decisions, insecticide spraying, and WNV presence means that the overall value of this metric for decision making and association with WNV is low. In figure 19, the summary plot we see that its mean magnitude SHAP value is incredibly low compared to other factors, and largely it is acceptable to ignore within the context of this model.

Seasonality

There is only one feature which is a measure of seasonality, sunrise_diff, which is the difference in sunrise time between one day and the day prior. A negative value indicates that

days are growing longer, the sunrise is occurring earlier in the morning, and the summer solstice has not occurred. A positive value indicates the opposite.

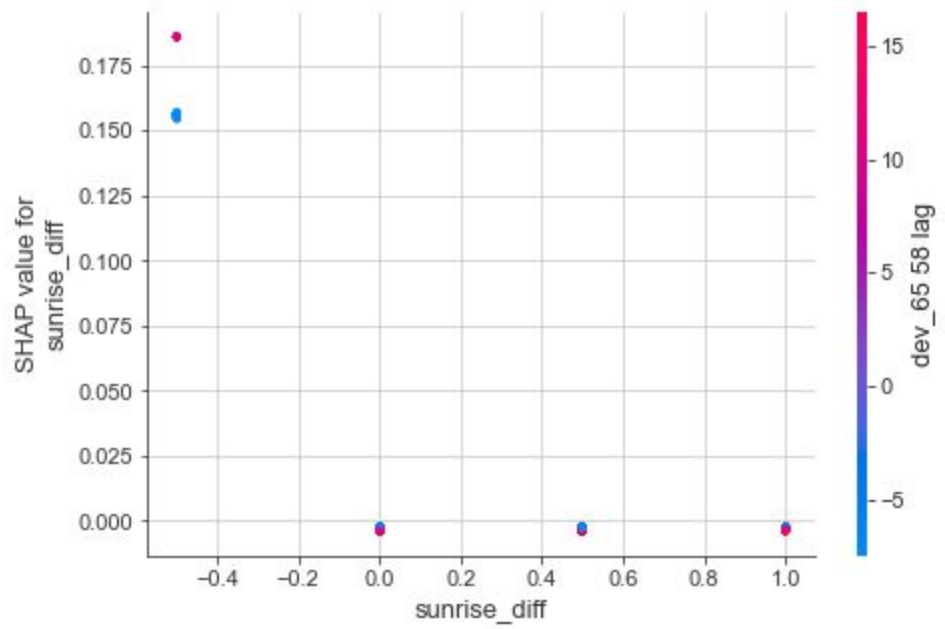

Figure 25: SHAP dependance plot for sunrise_diff

The dependence plot above indicates that WNV presence is highly associated with the early days of summer, when the days are growing longer, and there is relatively little SHAP contribution otherwise.

The causality behind this association is unknown, but it's likely intertwined with human intervention in the mosquito ecosystem. It could be truly seasonal, that more mosquitos and WNV sightings occur before the peak of summer. Or, this relationship could again represent the reactivity of the mosquito control system in place at the moment. Only once mosquito populations have truly taken hold, presumably later on in the summer with more consistently warm temperatures, does the public health infrastructure in Chicago react with insecticide spray and other treatments.

Regardless of the causal mechanics behind this relationship, it's powerful for prediction of WNV if the status quo of mosquito control is maintained. If public health officials change mosquito control techniques or strategy, the relationship seen here could quite possibly shift. And if the relationship does not shift, that might act as evidence for true seasonal dependence.

Mist

BR 50 exp is a 50-span exponentially weighted moving average of daily mist occurrences. Mist forms when warm air over water suddenly encounters cooler temperature on land. The moisture in the air condenses into small droplets and hangs around the air. It's a proxy indicator of humidity, pressure, and temperature.

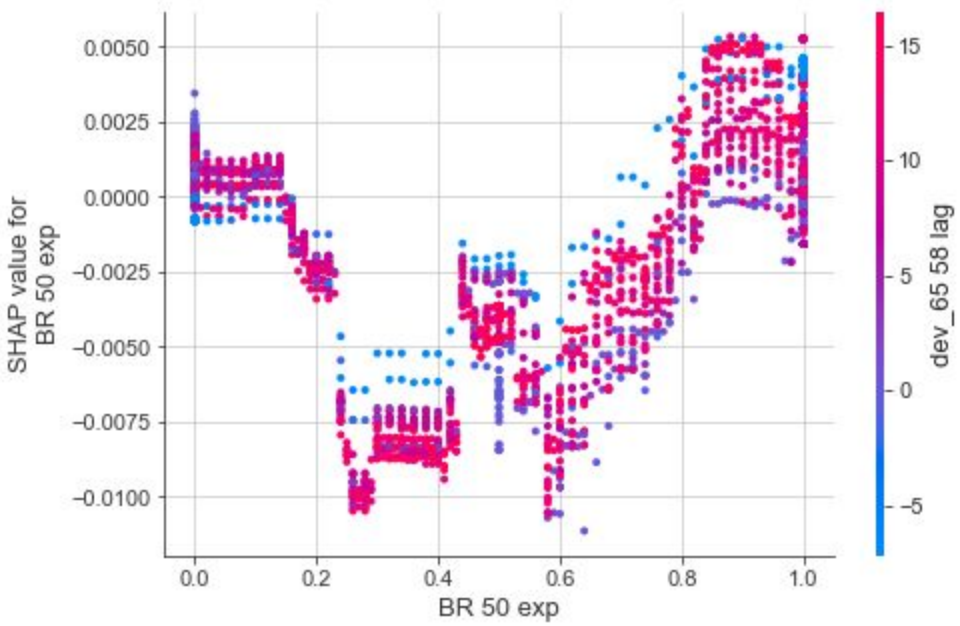

Figure 26: SHAP dependance plot for BR 50 exp

Lake Michigan, which Chicago is situated on, is an incredibly large thermal mass and maintains a more consistent temperature than land does. Wind is consistently, on average, coming off of the lake, and not once in the 10,000 observations does the prevailing wind come from an inland direction. High values of BR 50 exp indicate cool and damp nights and early mornings with wind coming from lake Michigan, while low values indicate relatively warm and dry nights and mornings.

High trailing averages of mist occurrences (>0.8), meaning long periods of cool, humid nights have strong positive relationships to WNV. Interestingly, low trailing averages of mist occurrences (<0.15) are also slightly positively associated with WNV. Yet, the middle ground (0.15-0.8) is negatively associated with WNV. This relationship is surprising, and there are no clear reasons that we expect to see this. More work needs to be done to identify the unique climatological conditions that affect mosquito populations.

Haze

Haze, as opposed to mist, is caused by the suspension of dry particles, often pollutants, in the air. HZ 10 exp is a 10-span exponentially weighted moving average of haze occurrences.

Interestingly, as with mist occurrences, the relationship between HZ 10 exp and SHAP values is non-monotonic. The shape of the dependence curve below is roughly parabolic and peaks around 0.5, and again a bit lower around 0.9. Generally, however, baring extremely high values, there is a positive association between HZ 10 exp and WNV.

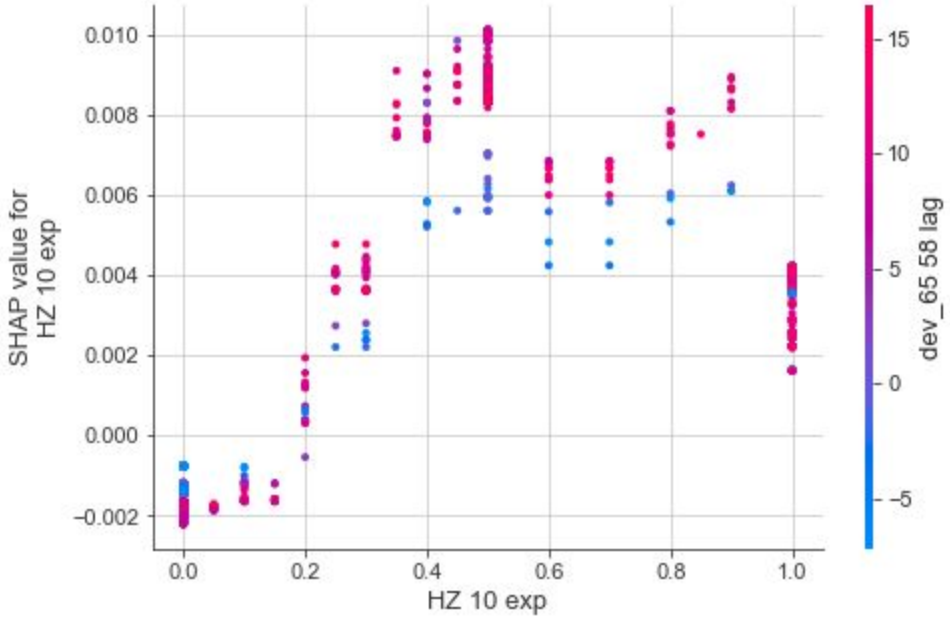

Figure 27: SHAP dependance plot for HZ 10 exp

It's possible that very high levels of haze caused by wildfires, intense pollution, or other sources, keep mosquito activity down and reduce the risk for WNV virus. In the absence of more detailed data on exactly what pollutants are in the air, it's difficult to come to any real conclusions about the nature of this relationship.

CONCLUSIONS

Bringing WNV cases down to zero per year is a complex problem at the intersection of many human and natural factors. This model's predictions operate primarily on natural factors, yet mosquito populations at large are highly influenced by the actions of humans, like insecticide

spraying, and human actions are likely confounding variables in many of the observations in the data set.

With the features engineered and explored above, a RF classifier with a max depth of 9 and 1050 estimators predicts the presence of WNV carrying mosquitoes relatively well. With an AUC of 0.8 on the training set and 0.82 on the test set, the model shows strong enough performance and generalizability to be considered for practical implementation and use.

Feature analysis with the SHAP library largely confirmed existing knowledge about the interactions between mosquito populations and climatological factors. Yet, there are interesting relationships which deserve more attention and insight from the research community. The association between WNV and haze, seasonality, and mist are not well documented or discussed, yet they are prevalent factors in predicting the WNV according to this analysis and modeling.

If the system of mosquito control changes, the underlying relationships on which the model is built may experience change as well. In the absence of more metrics and features that quantify human interaction with the mosquito ecosystem, it's challenging to advocate for the full integration of model-based decision making.

Instead, a more robust approach is to use the model as an additional tool in the arsenal of mosquito control. As its predictive power is validated in the field, and we better understand the relationships between the metrics that are tracked and the presence of WNV, the tool can only improve. A significant step in the integration would be more detailed and diverse metrics about human behavior, at which point model-based decision making might become a reality for Chicago's Department of Public Health.