SMARRT
Consulting Group

Analytics
Solutions
Public Safety

Final Report

2019 Chicago West Nile Virus Action Plan

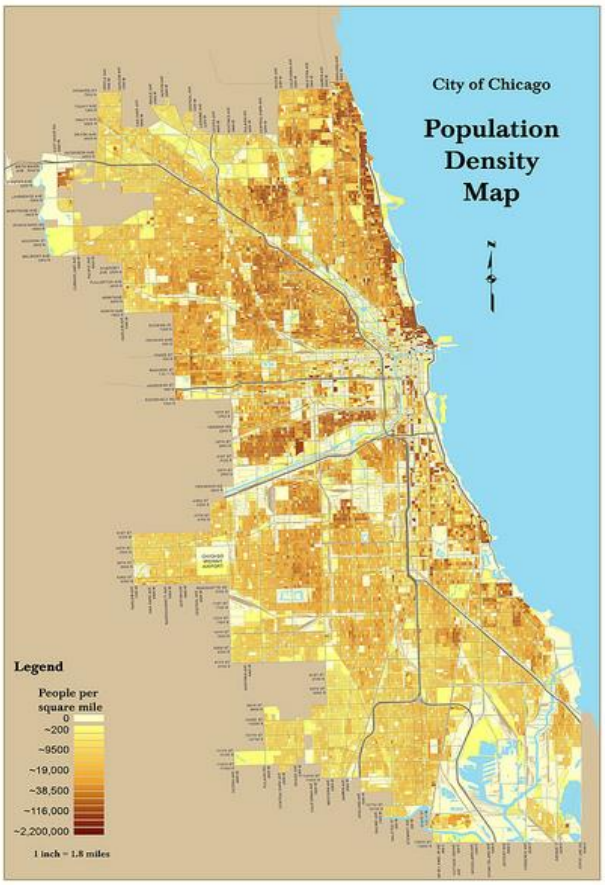
Team:
Andrew Cooper
Rachel Dudle
Stephen Hage
Mike Kapelinski
Ted Inciong
Rahul Sangole

Table of Contents

Executive Summary

Overview

In September 2001, West Nile virus was first identified in Illinois when laboratory tests confirmed its presence in two dead crows discovered in the Chicago area. This comes only two years after West Nile virus first emerged in the United States in New York in the fall of 1999. By the end of 2002, Illinois had counted more human cases (884) and deaths (64) than any other state in the United States[1].



This is where SMARRT consulting group can help to prevent this costly pandemic from resurfacing and preserve public safety. SMARRT Analytics has focused exclusively on consultation in matters of public health to assess the risk of disease outbreaks in cities across the world to provide analytical expertise ultimately providing recommendations for intervention and prescriptive prevention.

The project outlined below proposes means to assess and provide analytical insights to prevent a West Nile outbreak in the Chicago area. The goal is to assess Chicago for West Nile prevalence. We will also establish the most influential factors which may contribute to an outbreak. We can then predict the potential of an outbreak and the best mitigation and prevention program for a city of 2.7 million residents.

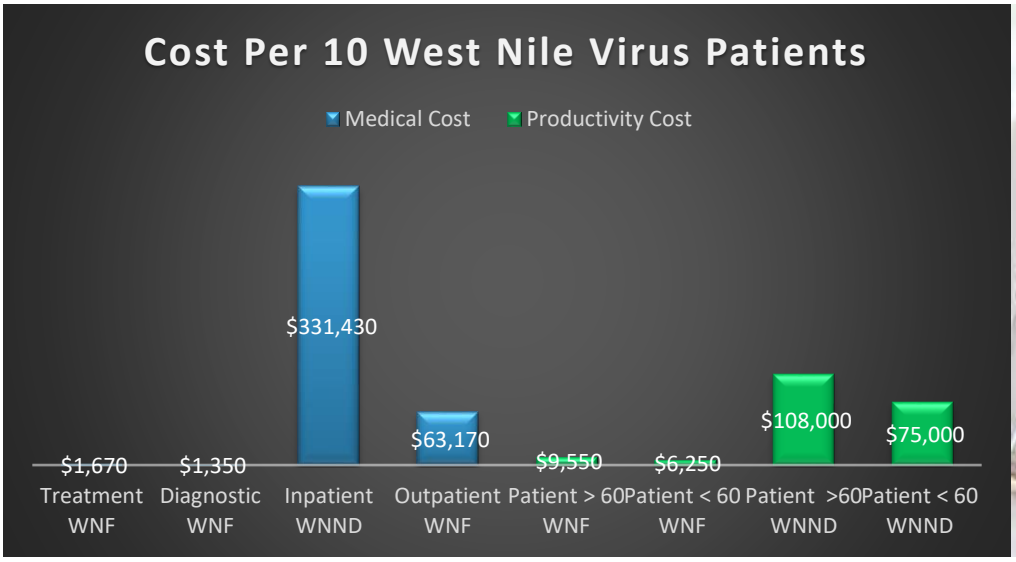
[1] http://www.dph.illinois.gov/topics-services/diseases-and-conditions/west-nile-virus

Business Case

In 2002, West Nile virus was discovered in Chicago for the first time with over 225 cases reported. In response, the Chicago Department of Public Health (CDPH) has implemented a city-wide surveillance and mosquito control measure program. Based upon a 2010 study produced by the CDC, a 2005 outbreak of West Nile virus cost Sacramento County, California $2.98 million. West Nile Virus (WNV) can have two different effects on a human host, West Nile Fever (WNF) and it's much more severe and costly West Nile neuroinvasive disease (WNND). A cost benefit analysis was also performed during this study, which indicated that preventative measures such as spraying would only need to prevent 15 cases of WNND to make the control measure cost effective.[3] Today the population of Chicago is 2.7 million which is 1.8 times the size of Sacramento County, California and the outbreak we experienced in 2002 having 225 infected compared to the 163 cases in California would result in an overall cost of over $4 million in medical and productivity costs alone.

The CDPH recently commissioned a study which revealed the city of Chicago can significantly reduce costs associated with treating West Nile virus in hospitals and clinics through simple preventative measures. These include but are not limited to community level control programs such as targeted mosquito spraying of high-risk area and removal of debris associated with mosquito breeding (i.e. areas with stagnant water). In addition, personal protective measures such as use of mosquito repellant and wearing long sleeves have shown to decrease exposure to infected mosquitoes.

Therefore, the CDPH opened a requisition to modernize their mosquito controls system. The CDPH would like a system to identify top and emerging high-risk areas for the West Nile virus as well as a dashboard which allows users to monitor the West Nile virus in real-time. This modernized system will allow the CDPH to prevent the spread of West Nile virus.



[2]Healthy Chicago Data Brief - West Nile Virus. (n.d.). Retrieved from https://www.chicago.gov/city/en/depts/cdph.html
[3]Economic Cost Analysis of West Nile Virus Outbreak, Sacramento County, California, USA, 2005 - Volume 16, Number 3-March 2010 - Emerging Infectious Diseases journal - CDC. (2010, December 14). Retrieved from https://wwwnc.cdc.gov/eid/article/16/3/09-0667_article

Goals

SMARRT Consulting Group will help Chicago decrease the amount of West Nile virus infections in a cost-effective way. In order to accomplish this goal, guidance will be given to the city where preventative measures (like spraying) would be most effective, and where the most vulnerable populations live. We recognize that Chicago Department of Public Health already undertakes mosquito abatement via screening, targeting areas based on mosquito trap test results. These abatement efforts curtail mosquito population growth and reduce transmission of West Nile virus but cannot eliminate it. SMARRT Consulting Group's advanced models can improve upon existing methods for targeting mosquito spraying by identifying times and places where risk remains high and identification of risk areas early can prevent mosquito problems from increasing. In addition to using advanced predictive modeling techniques, SMARRT will use ancillary data sources that will better identify areas where neighborhood characteristics contribute to mosquito growth.

In addition to refining predictive models to identify areas to target with spraying, SMARRT will identify high risk regions where there is increased risk for human transmission and neuroinvasive disease. We will identify areas with a high concentration of vulnerable people using demographic data, school and senior center locations and other data sources.

This effort would also alert consumers as to the risk of West Nile virus in their location, so the population can take preventative measures as well. Traditional public service announcements and flyers educate the public about the risk of standing water and ways in which the public can reduce risk of mosquito-transmitted infections, but these efforts are rarely targeted to the neighborhoods and times when education will have the largest impact. We can change this by making timely risk data available to the public and communicated in easy to understand terms. Ultimately, this will lead to a safer and healthier Chicago, at limited taxpayer expense.

Objectives

Goal Type	Business Objective	SMARRT Deliverable	Success Criteria
Deliver Predictive Models	City of Chicago can perform strategic and targeted intervention activities by identifying areas which have high risk of mosquitos carrying West Nile Virus	Deliver predictive models using mosquito activity data, weather data and data from the city of Chicago like zoning, demographic, income characteristics etc. Stretch Goal: Evaluation of factors which affect spraying effectiveness.	Models evaluated on common classification and regression metrics using cross-validation and hold out datasets.
Deliver Real Time Actionable Insights	City of Chicago and its residents will have access to up to date predictions of West Nile Virus threat levels among other valuable insights	Deliver a dashboard with real time updated threat levels, predictions of outbreaks, identification of high-risk regions like hospitals, playgrounds, senior living facilities etc. Weekly or monthly reports by zip-code or neighborhood can be generated.	Voice of customer feedback score on usability and value of dashboard or mobile application.

The Approach

Data Sources

To develop the predictive models and insightful dashboards, the team will scrape data from various sources from the internet. As we have shown in the table below, data needed to address this problem are varied in size, complexity, variety, quality and availability. After a preliminary investigation into these data, the team has assigned qualitative scores to help determine feasibility for model building.

	Size	Complexity	Variety	Quality	Availability
	Unknown	Unknown	Unknown	Unknown	Unknown
	<100MB	No preprocessing required for consumption	Numerical	…	Some risk in availablity
	100-500MB	Some preprocessing required for consumption	Numerical + Categorical	Quality, not vetted	Partially Available
	500-1GB	Substantial preprocessing & preparation required	Temporal + Numerical + Categorical	…	Available, not vetted
	1GB+	Substantial & complex data munging required	Spacio-temporal + Numerical + Categorical	High Quality & vetted	Available & vetted

Source	Dataset	Description	Size	Complexity	Variety	Quality	Availability
Chicago Department of Public Health	Mosquito trap and West Nile Virus, 2007-2018	27000 records over 11 years with location, mosquito species, and presence of west nile virus					
National Oceanic and Atmospheric Administration	Daily weather data, 2007-2018	Daily precipitation and high/low/average temperature readings for Chicago and surrounding areas					
Webscraping or Satellite Imagery Analysis	Geospatial Water Body Information	Locations and metadata of water bodies and marshlands for Chicago and surrounding areas					
Unknown	Aviary data	Bird population and death rates by location for Chicago and surrounding areas					
United States Census Bureau's American Community Survey	Sociodemographic data, 2007-2018	Poverty rates, socioeconomic status, education status, unemployment status over 11 years					
Cook County Data Portal	Geospatial Hospital & School Locations	Locations of hospitals, schools, and senior assisted living facilities to characterize areas of highly vulnerable populations					
Unknown	Financial Data	Financial impact of West Nile Virus: estimated per-infection treatment costs, spraying and prevention costs, impact on businesses and local economy					

1. The primary dataset to be used for predictive modeling, the Mosquito trap and West Nile Virus test data, 2007-2018, were obtained from Chicago Department of Public Health (CDPH) via the Chicago Data Portal. These data provide mosquito trap locations (latitude & longitude), species-specific mosquito counts and West Nile virus test results. This dataset is small: roughly 27,000 observations for 12 variables.
2. Daily weather data were obtained from the National Oceanic and Atmospheric Administration (NOAA), which is part of the United States Department of Commerce. Data were extracted from Daily Summary

data and Local Climatological Data for two main Chicago airport weather stations, plus a subset of Daily Summary data were extracted for 303 regional weather stations in or near Cook County, Illinois.

3. Geospatial predictors such as locations of bodies of water & marshland were considered. Web scraping or satellite imagery analysis are two options to identify areas with a higher concentration of bodies of water and standing water. *Further feasibility study is required to determine if obtaining such information is possible, and it is outside the scope of this project, but could prove useful to CDPH in the future.*
4. Although aviary data for specific bird populations would be very useful, these data are not available in any known dataset.
5. To assess potential human impact of West Nile virus outbreaks in mosquitos, we will investigate sociodemographic data from the United States Census Bureau's American Community Survey (ACS) obtained from ftp2.census.gov. We extracted measures of poverty, low socioeconomic status, and vulnerability to West Nile virus and neuroinvasive disease. We used 5-year ACS summary data at the Census block group, Census tract and zip code levels. Many CDPH mosquito traps were already geocoded with latitude & longitude, and we geocoded the remaining traps and performed spatial joins to identify the Census block group, tract, zip code, and Chicago neighborhood community area in which each trap is found.
6. We obtained locations of hospitals, schools, senior assisted living facilities and other areas with high concentrations of vulnerable populations. Hospital and school locations were obtained from the Cook County Data Portal. These data locations (latitude & longitude) were spatially joined to obtain Census block group, tract and Chicago neighborhood community area. All spatial joins were performed using Census TIGER/Line files (i.e. GIS shapefiles).
7. To assess the financial impact of preventing West Nile virus infections, we used published cost analyses that have assessed the burden of medical care for West Nile virus patients. These data will be used to estimate financial impact of West Nile virus given the predicted mosquito count, likelihood of West Nile virus being present, and demographics of the human population nearby. We will also use published analyses of the cost of mosquito abatement programs for West Nile virus.

Description of Data

West Nile virus trap results from CDPH were comprised of 27,196 rows of data for ~194 unique trap locations in Chicago over a twelve-year period (2007-18). These data were structured with a separate row for each batch of 50 mosquitos tested, with separate rows of data for each mosquito species. Only observed species were reported. Unique trap names and locations were extracted from these data, and traps that were not already tagged with latitude/longitude pairs by CDPH were geocoded. An indicator variable was constructed for "satellite" traps which are typically set up in close proximity to another trap where CDPH wanted additional surveillance for mosquitos. When trap names were reused for completely different locations and times, new names were assigned to differentiate between them. The final set consisted of 194 unique locations situated in 146 Census block groups, 138 Census tracts, 47 zip codes, and 63 (out of 77) Chicago community areas. Spatial joins also provided zoning data for each trap location (e.g. residential, commercial, etc.). The trap result data were aggregated to a single row for each trap and date for which results were available, with separate columns for each species-specific result (count of mosquitos and whether any of them tested positive for WNV) and an additional overall summary (count of mosquitos across all species and whether any were positive). The resulting dataset consisted of 13,631 rows of trap results spanning 2007-18. Data could be further reshaped to produce weekly or monthly results for each trap, but this was determined to be unnecessary for most modeling approaches discussed further in this report.

NOAA weather data were obtained after the partial federal shutdown ended on January 25, 2019. In order to construct lag terms for all possible variables, complete data from 2006-18 were obtained for 303 weather stations in or near Cook County, Illinois. These daily data consisted of 54 variables. However, due to large missing data, initial analyses were planned using three temperature variables (daily min, max and average) and one precipitation variables (total daily precipitation) from two main Chicago airport weather stations (Midway and O'Hare). By using data from two stations located in different parts of the city of Chicago (far northwest and mid-southwest), we anticipate that advanced modeling techniques will result in proximally appropriate station data being weighted more heavily for prediction purposes. These data were reshaped, resulting in a dataset with 4747 daily observations 2006-18 of eight variables (four for each station). Data were >98% complete, and the missing data were imputed via Multivariate Imputation by Chained Equations using the MICE package in R. Moving averages were computed across multiple time periods (e.g. 2-week, 4-week and 90-day averages and minimum daily values) and lag terms created from these complete daily data, then they were appended to the observed WNV trap result data by matching on date. *Stretch goals for weather data include using data from more weather stations and more variables.* A distance matrix was computed using Haversine distance from each weather station to each trap, and the nearest stations for each trap were then determined after limiting to stations with >= 95% complete data for a given variable. This allows each trap to obtain more proximate weather data to be used while keeping missing data down to a level at which imputation methods are relatively effective. This is an appropriate technique for precipitation; daily temperature data, however, is more prone to issues when there are no weather stations within 8-10 miles of a given WNV trap (for example, if the nearest weather station is adjacent to Lake Michigan, temperatures there will be markedly cooler than at the trap location which is inland). There are also additional weather fields available from an additional NOAA data source of monthly/daily/hourly Local Climatological Data. These include variables for dew point, Heating Degree Days (HDDs) and Cooling Degree Days (CDDs). Incorporating those data points into models will be a stretch goal.

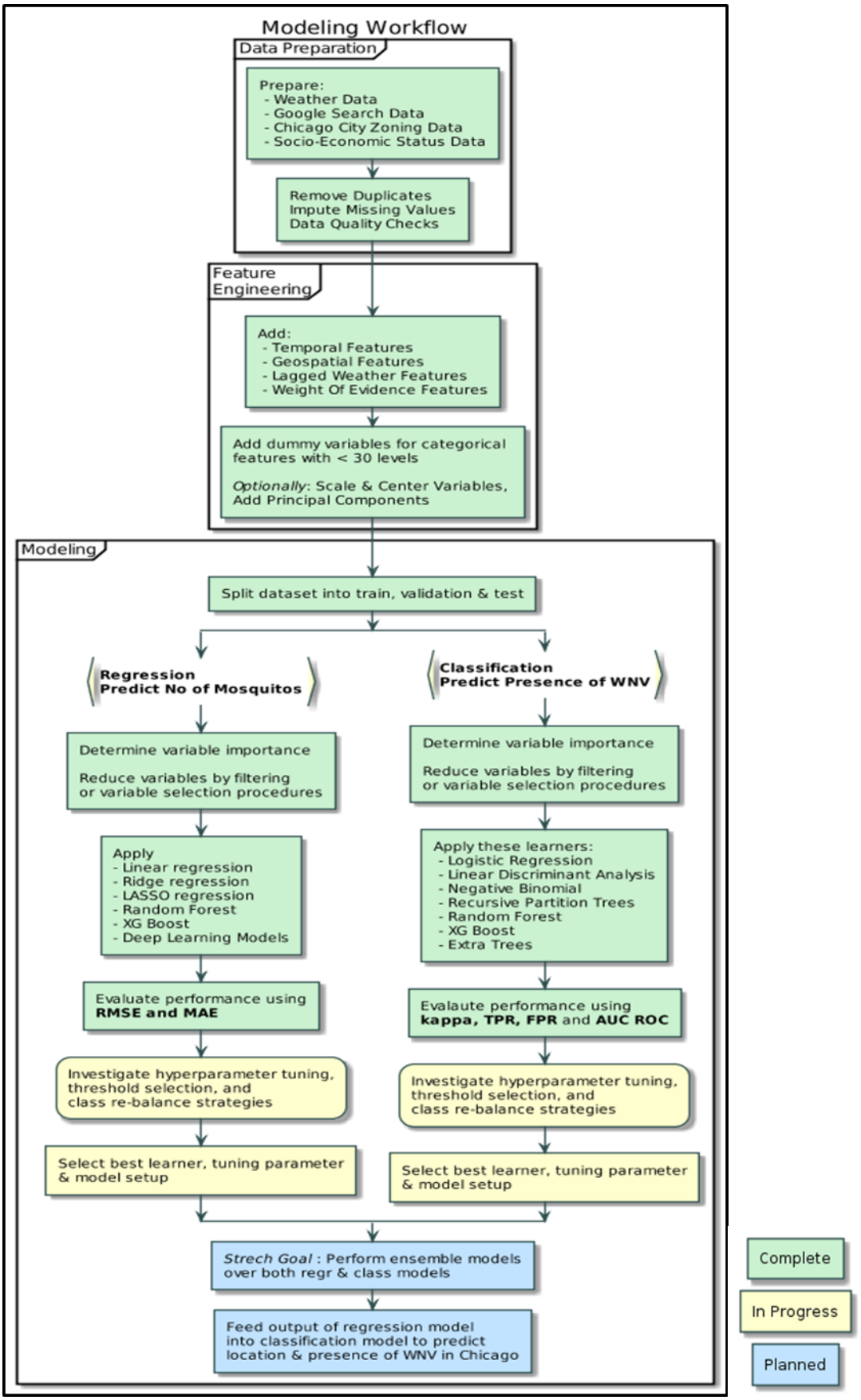
Census American Community Survey (ACS) data were extracted from 5-year summary files for each year from 2005-09 to 2013-17. Specific summary tables were extracted for topics including median household income, education for adults over age 25 in the household, income relative to poverty level, total population, and population by age, sex, race and ethnicity. Estimates for different subcategories were tabulated and summary variables calculated at the block group and Census tract level. The three socioeconomic status (SES) indicators and demographic characteristics were extracted from 2013-17 data at the two geographic levels and appended to daily WNV trap data that had already been geocoded and tagged with geospatial IDs. Although these data do not vary over time, they provide geospatially-specific measures of the population near each trap and could be useful to the extent that neighborhood SES characteristics are correlated with conditions that affect mosquito growth such as abandoned properties and standing water. In addition, we developed year-specific estimates of these measures where available, as they represent slowly changing dimensions that could provide useful time- and location-specific proxies for neighborhood characteristics that influence mosquito population growth.

To further address neighborhood conditions correlated with mosquito population growth in published literature (i.e. abandoned properties and standing water), we obtained lists of vacant properties (n=2352) and building violations (n>1 million) from the Chicago data portal. We geocoded each of these data sources and performed spatial joins to obtain Census block group, tract, zip code and Chicago community area for each row. We then aggregated data to compute 180-day summaries (total number of vacancies or property violations in the previous 180-day period) for each day in the 12-year WNV trap result period and each Census block group, tract, zip code and community area. In the end, this provides only a few additional variables that

were joined to the trap result data on a combination of date and geographic identifier. While the data were somewhat sparse, particularly for vacancies and especially for more granular geographic summaries (e.g. block group), they may have value for modeling purposes because they vary both temporally and geospatially. This has the potential to add value to our models. Other data that are location-specific but fixed in time are less likely to be useful since advanced models can already derive much of the information from trap location information alone. Re-aggregating over multiple time periods and searching a grid of temporal and geospatial aggregations to find the combination that adds the most value to a model using feature selection and dimension reduction methods was a stretch goal that was deemed out of scope at the current time but could be implemented in future revisions.

Hospital, school and senior center locations were obtained from the Chicago and Cook County data portals. These data were geocoded when that was not already done, and spatial joins were performed to obtain Census block group, tract, zip code tabulation area and Chicago community area. These data have value for visualization purposes and presentation but were not used in predictive models. The same is true of cost data.

Model Process Flow

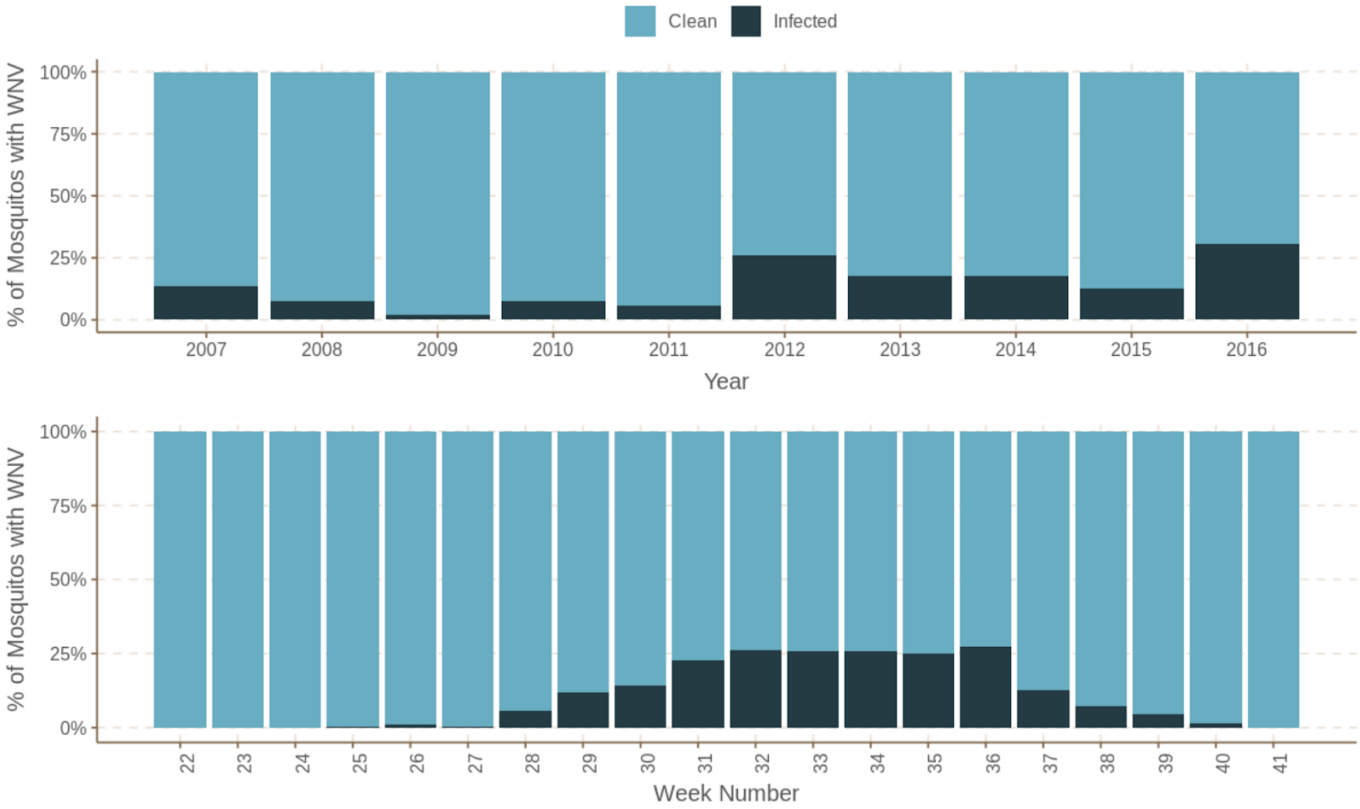


Analysis of Data

Exploratory Data Analysis

At this stage, SMARRT Consulting Group has identified actual data sets and is ensuring the data is of high quality and reflective of the intended purpose of the models. An initial look at the data set reveals the current year, 2016, has the highest number of WNV cases followed by 2012 (Figure 1). Furthermore, the WNV season typically lasts fifteen weeks and begins around week twenty-five (June), peaking between week 30 through 33 (July - August), and ending around week 40 (October). While the week in the year may provide some predictive power, year does not appear to provide any predictive strength.

Figure 1



As we see in *Figure 2* below, the data set contains categorical location variables which identify the trap's name, community/neighborhood, and zip code for which the trap is set. One sees trap T090B, T900, and T143 account for a high ratio of positive WNV cases while traps T917, T909, and T040 account for no cases for WNV. Trap name will be investigated as a feature variable during the model development stage.

Zip code and community describe the trap's location in a similar fashion. One sees zip codes 60018, 60631, and 60639 have the highest cases of WNV while zip code 60616 has zero cases of WNV. Similarly, several communities are identified as having zero cases of WNV. Both zip code and community will be investigated as feature variables for predicting WNV though there is overlay between the two variables.

Figure 2

Figure 3 provides trend analysis for WNV by zip codes and year through a bubble chart. One sees in 2007, 2012, and 2013 WNV impacted these zip codes in a similar fashion, i.e. cases of WNV were not isolated to a few zip codes. Furthermore, 2009 and 2011 (excluding 60018) saw consistently lower cases of WNV. These trends may suggest correlations to external factors (e.g. weather) or cyclical life-cycles which may be associated with a specific mosquito type (e.g. annual cicada vs. 13- and 17-year cicada). Given the difficulty with predicting the weather and mosquito class, these attributes will be excluded from the model development process.

Figure 3

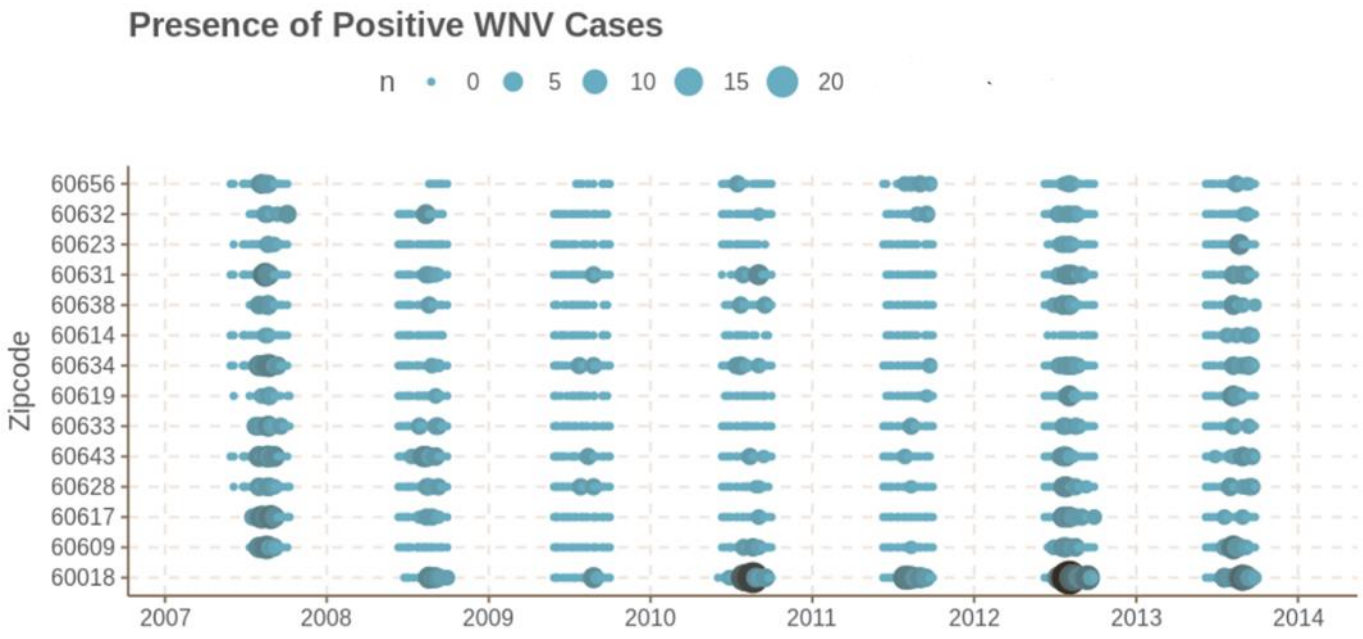


Figure 4 shows Google searches pertaining to mosquito bites, WNV, and WNV symptoms. Interestingly, WNV symptoms searches start peaking after searches for mosquito bites and WNV. This would suggest people are

being bit by mosquitos, have concerns about WNV, and are finally concerned about contracting WNV. The results of the Google search index act as a benchmark for the Chicago WNV data set. When compared to *Figure 1 and Figure 3* both data sets are directionally aligned having 2009 as a low period and 2012 is a high period for WNV cases.

Figure 4

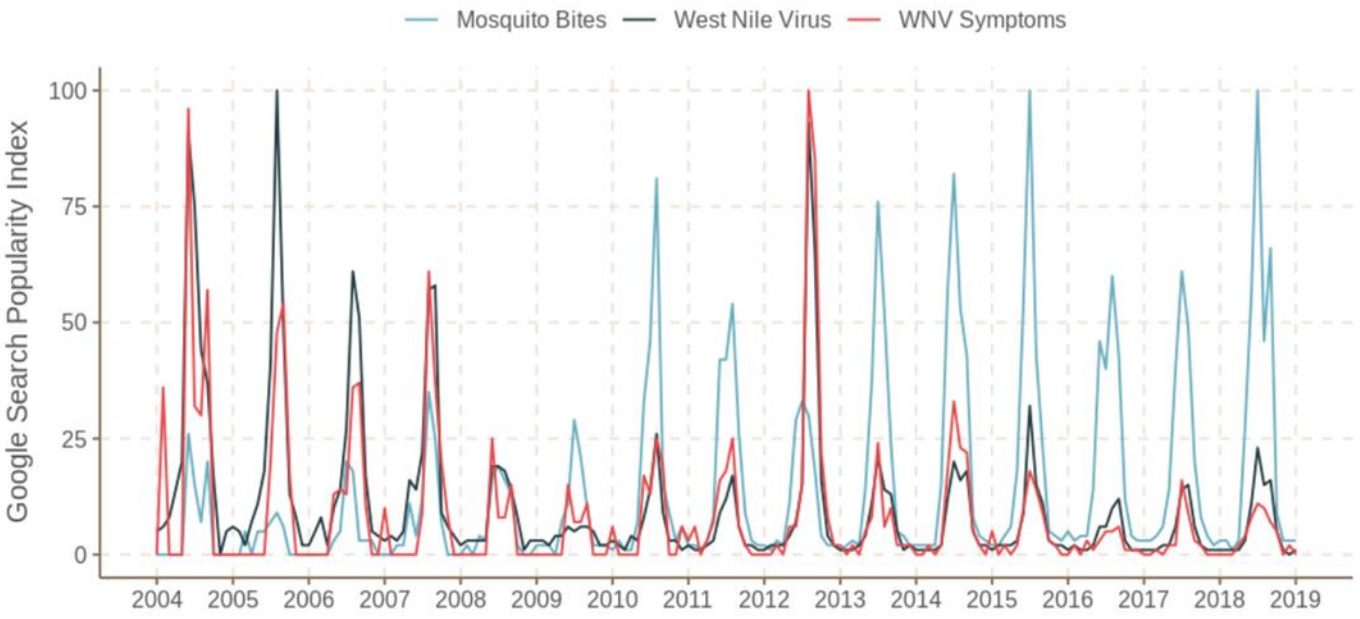


Figure 5 shows annual counts of mosquitos by species for positive and negative WNV test results. This indicates that some mosquito species are much more prevalent in Chicago than others. Culex Pipiens and Culex Restuans in particular account for the majority of mosquitos found in the traps. In later years, there are more records differentiated between those two species, but there are test result data through all years where the two are combined. There is an uptick in Culex Territans in later years. While there is value in continuing to monitor species-specific rates with descriptive analyses, these results suggest that differentiating between species may not be important for public health planning purposes, and models may therefore focus on total count of mosquitos and any presence of WNV in a given location & time.

Figure 5

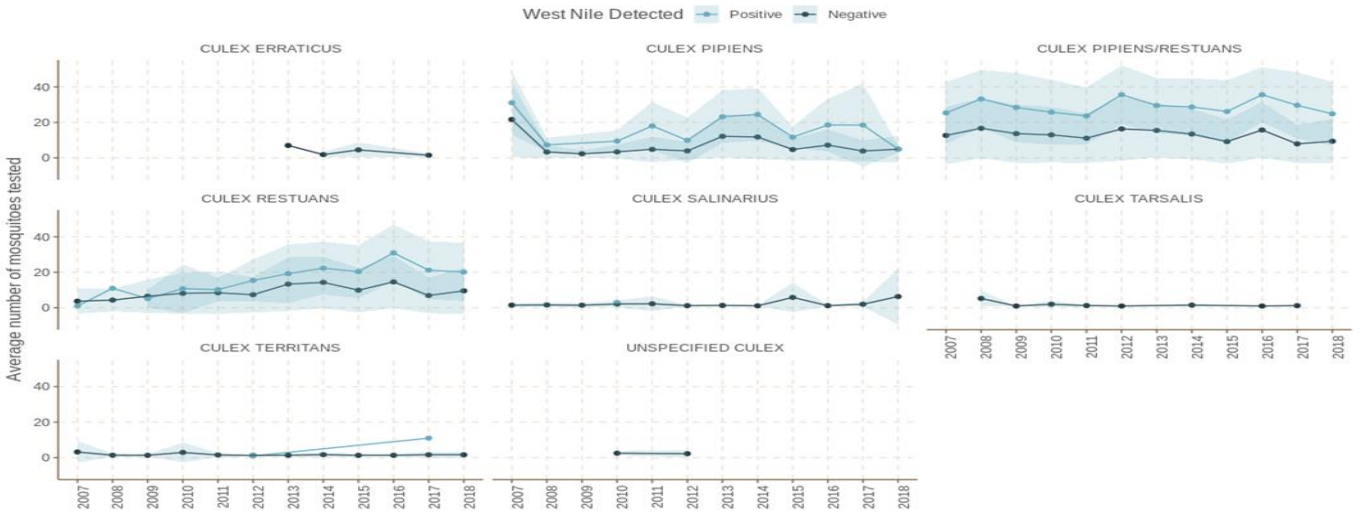
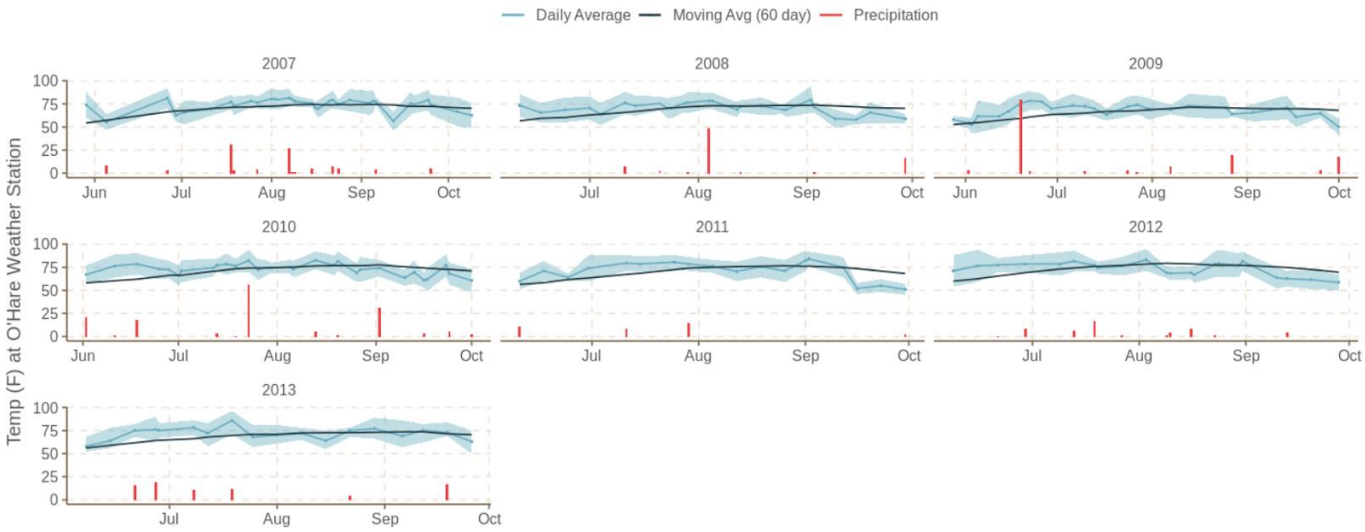


Figure 6 presents time series plots of NOAA weather data for the O'Hare Airport weather station. Daily average temperature and 60-day moving average show that temperatures tend to peak in late summer but with a fair amount of daily variability. Red bars indicate daily precipitation which is much more randomly dispersed. Literature review indicated that daily temperature is correlated with mosquito activity. In addition, mosquitos need rainfall to leave moisture or standing water where they can breed, but then mosquito activity tends to peak in subsequent weeks and is associated with drought periods.

Figure 6



Transformation of Data/Feature Engineering

Based off of the EDA, the team as put together new features prior to the modeling activities. These are the new features developed.

Temporal Features

As we saw in the EDA, the mosquito and weather datasets show seasonality. Thus, we added temporal features like Month, Day, Quarter, Week Number, Day of the Year, Day of the Week, and Day Name which would be useful for regression and classification models. We removed Year from the equation, since we did not see any inter-year trends, only intra-year trends.

Weather Features

The original dataset contains daily precipitation, minimum, average and maximum temperatures for O'Hare and Midway airport weather stations. We know - through literature review- that mosquito populations have a strong correlation to recent-past temperatures & precipitation values. Weather data also tend to have a large degree of day-to-day variability. This variability doesn't offer much in terms of predictability - sometimes, it's better to look at overall trends.

To account for these two ideas, we created many styles of lag variables on "smoothened" weather data. For example:

- 7 Day Moving Averages for Min, Max & Average Temperatures & Precipitation
- 30 Day Moving Averages for Min, Max & Average Temperatures & Precipitation

- 60 Day Moving Averages for Min, Max & Average Temperatures & Precipitation
- 60 Day Total Precipitation Values
- 60 Day Moving Averages, with a lag of 1, 2, 3, … 8 weeks for Min, Max & Average Temperatures & Precipitation

Geospatial Features

Apart from the original geospatial features in the dataset like latitude, longitude, zip code, block ID, zone information, one feature we created is the distance of each mosquito trap to the weather station for which the weather data is used. We use *haversine distance*, which determines the distance between two points on a sphere given their longitudes and latitudes.

Towards the tail end of the project, if we need some additional predictive power, the team will investigate other geospatial features like:

- Haversine distance for each trap to the nearest weather station, and usage of that weather station data
- Haversine distance to the closest localized WNV outbreak epicenter in the recent past

Dummy Variables

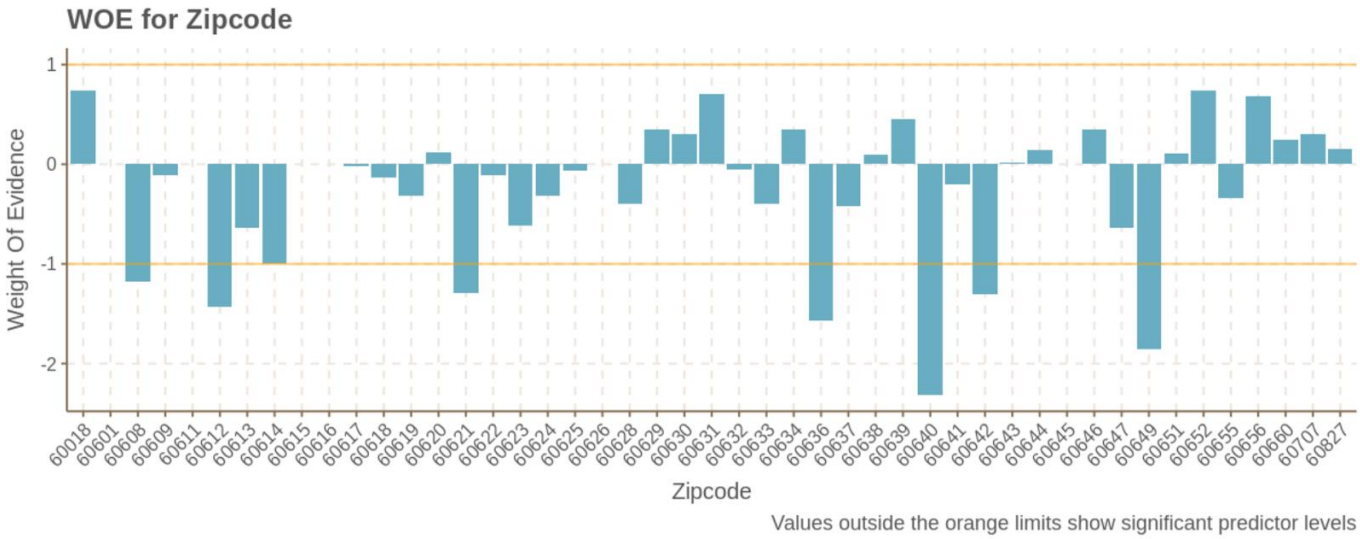
For those variables with categorical data, like *Zone Class, Month Name* etc., we added dummy variables to the data set. A dummy variable is a numeric variable that indicates the presence or absence of some level of a categorical variable. There will be n-1 levels of categories as 0 represents the base level.

Weight of Evidence Features

We have a few categorical variables with extremely large number of levels like *Trap Name (179 levels), Zip Code (47 levels) or Community (63 levels).* The poses a challenge when using certain types of models which require conversion of categorical variables to dummy variables, like traditional regression models or machine learning models like XG Boost. This is because conversion to dummy variables results in creation of a large number of sparse columns. For example, one column *Trap Name* would become 178 (or 179) columns. Thus, there would be an explosion of the size of the model - increasing computation time - while sparsity can also affect model stability. Furthermore, some implementations of models like Random Forest in R cannot accept variables with greater than 53 levels.

To overcome this, SMARRT Consulting used a concept called *Weight of Evidence (WOE)*. For each categorical predictor variable, the WOE values for each level is a number indicating how strongly there is evidence to support the hypothesis that the level is *predictive* of the response variable. For example, *Figure 7* shows the WOE for *Zip Code*. Numbers towards 1 indicate zip codes more likely to have WNV, while those less than -1 strongly indicate that WNV will not be present.

Figure 7



Each categorical variable is replaced with its corresponding WOE variable. Thus, there is no increase in the number of variables. Also, the WOE variables are numeric in nature, so no sparsity is introduced in the data either. This approach has proven quite beneficial to the team in the classification models.

SMARRT Consulting used *Information Value (IV)* to estimate variable importance for classification models. IV is a weighted sum of the WOE for each categorical variable. A higher IV indicates more predictive nature of a variable, as shown in *Figure 8.*

Figure 8


Variable Importance Using Info Value
mos_tot_num_mosquitos
mos_pipiens_restuans_num_mosquitos
t_mo
t_qtr
t_wk
mos_pipiens_num_mosquitos
t_eval_day
t_yr
t_eval_wk
googtrend_sym_wnv
wea_ohare_tavg2_ma60_lag7
wea_ohare_tmin
googtrend_mosq_bites
trap_trap_name
loc_census_block_group_id
loc_census_tract_id
loc_zipcode
loc_lng
loc_community
t_day_of_wk_name
t_day_of_wk
zone_class
ses_lt_pov_pct_tract2017
ses_lt_hs_pct_blk_grp2017
googtrend_westnile
loc_lat
ses_median_hh_inc_tract2017
wea_ohare_tavg2_ma60_lag14
ses_lt_pov_pct_blk_grp2017
ses_median_hh_inc_blk_grp2017
nbrhud_comm_180d_violation_cnt
zone_type
mos_species
ses_lt_hs_pct_tract2017
loc_lat_lng_src
nbrhud_blk_grp_180d_violation_cnt
nbrhud_zipcode_180d_violation_cnt
t_day_of_yr
trap_trap_type
mos_salinarius_num_mosquitos
mos_unspecified_num_mosquitos
mos_erraticus_num_mosquitos
mos_tarsalis_num_mosquitos
trap_satellite_ind
t_day
mos_territans_num_mosquitos
nbrhud_comm_180d_vacancies_cnt
wea_ohare_tavg2_ma7
nbrhud_blk_grp_180d_vacancies_cnt
nbrhud_zipcode_180d_vacancies_cnt
wea_ohare_prcp
googtrend_deadbirds
wea_ohare_prcp_sum7
mos_restuans_num_mosquitos
wea_ohare_tavg2_ma60_lag21
wea_ohare_tmax
wea_ohare_prcp_sum60_lag7
Predictor
-0.25
0.00
0.25
0.50
0.75
1.00
Cross Validation Adjusted Information Value
IV > Cutoff Threshold of 0.03 are stronger predictors

Final Data Review

The final dataset of WNV trap results consisted of 13,631 observations and 135 variables. The full data dictionary is presented in the Appendix.

Data were partitioned into training, validation and test sets by splitting on years. Seven years of data (2007 - 2013) were used for training with three years (2014-2016) set aside for validation & parameter hyper tuning. Two years of data (2017-2018) were held out to be used as a test set.

Partition	Years	Observations	Mean # of Mosquitos	Percentage of trap results testing positive for WNV
Training	2007-13	8,222	26.3	9.6
Validation	2014-16	3,430	26.6	15.0
Test	2017-18	1,979	16.3	13.7

Regression Models

Since West Nile virus (as well as other diseases) are transmitted through mosquitoes, it is critical to make predictions about the presence and density of mosquitoes in each area of the city. There are 194 mosquito traps in Chicago, which serve as a gauge for the number of mosquitoes in the area.

We initially considered conventional time series approaches such as ARIMA and ETS. However, through our EDA, we determined that while there is a strong seasonal component to the regression problem, there is no apparent linear relationship between year and mosquito counts. Therefore, we know that there is a limit to what we can accomplish with time series methods unless we use external regressors. For true forecasting, we would need to predict future values without most of those external regressors. In essence, this would require developing models for weather forecasting since intra-year temperature and precipitation patterns are among the strongest available predictors of mosquito counts. True weather forecasting is outside the scope of this project since it's a notoriously difficult task and there are already government and private sector efforts specializing in this. We feel our efforts are better focused on predicting mosquito counts given known external regressors, whether that is observed weather data in a hold-out sample or forecast data from another source. Beyond the weather problem, conventional time series approaches are very good at making near-term forecasts using observed data lagged over short intervals (e.g. 1 week or 1 month). Demonstrating the value of near-term forecasts is a reasonable stretch goal but not the main focus of our predictive models. While conventional time series methods respect time (training on only past data at each given time point), there are many other regression methods that are well-equipped to capture seasonality, trends and information gleaned from predictors such as weather data, so our regression methods will focus on them.

The following regression models were identified for predicting mosquito counts:

1. Linear Regression (e.g. Ordinary Least Squares)
2. Ridge Regression
3. LASSO Regression
4. Random Forest

Feature selection and dimension reduction are important steps in successful modeling. SMARRT Consulting collected data for 135 different variables, which necessitates either uncovering which are most predictive of the count of each species of mosquito, or the use of models that are robust to so many variables.

Linear Regression models are not our primary models because they cannot adequately capture geospatial latitude/longitude data without introducing splines. We can nonetheless use linear regression to make predictions using some variables, and stepwise regression is one viable method for feature selection.

A model that is actually robust to many variables is Ridge Regression, which shrinks the weight of unimportant predictors towards zero. It captures some of the trend but does not do a good job of predicting the number of any species of mosquito. Interestingly, the variables given the most weight are date variables, followed by moving averages of temperature and precipitation. This implies that, for the prediction of mosquito density, there is a sequential or time-series component. LASSO Regression is a similar regularization technique that, unlike Ridge, actually shrinks some coefficients to zero, thereby functioning as a feature selection tool. LASSO and Ridge are part of a family of Elastic Nets functions. While Ridge Regression can produce good predictive models, LASSO regression can further be used to select which features should be used in other modeling techniques.

In order to uncover other patterns, a deep neural net was developed. Since neural nets are sometimes called "universal function approximators", a multi-layered neural net can also help determine variable importance, and the interplay between the different variables. For some species, in particular the *Pipiens Restuans*, it does an adequate job predicting the density of mosquitos. It tends to, if anything, predict spikes in density early, which will need to be adjusted in final modeling.

In theory, random forests are robust to many variables, though if they are not properly tuned, this can result in slow models and difficulty with interpretation, even when examining metrics of variable importance. Regression trees are inherently prone to overfitting. Random forests are ensembles of trees that effectively reduce overfitting. Random forests can be fit to data that have already been processed with feature selection and dimension reduction techniques (e.g. LASSO and Principal Components Analysis).

Additional development is underway for sequential models for cohorts of traps. It is evident that O'Hare, for example, has a higher propensity to have mosquitoes that are WNV-positive. There is some evidence that different models for each trap, or for traps that have strong similarities, will outperform models that predict all traps. Moreover, some models (such as Recurrent Neural Networks) capture sequences well, which are expected to improve performance. Finally, SMARRT is developing automated model selection/ensembling in order to programmatically improve predictions for each Chicago neighborhood.

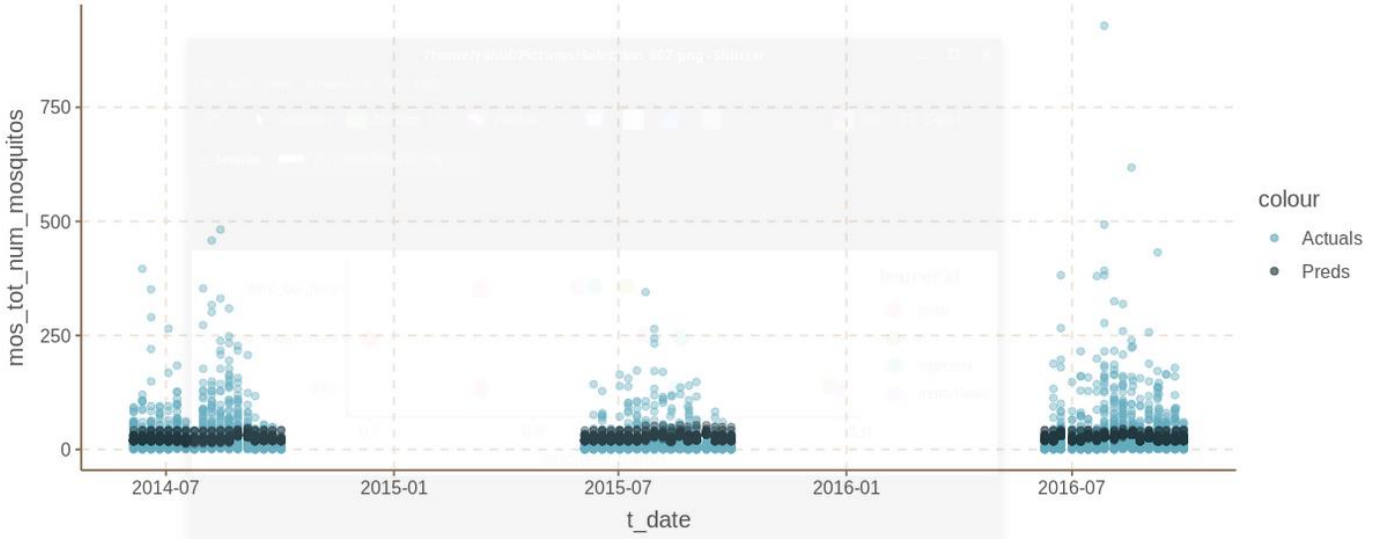
Regression model results are incomplete at this date as several models are still being revised, so performance metrics are not yet presented. A table shell shows how final results will be presented.

Modeling the number of mosquitos proved to be a difficult task, with no single model consistently capturing the mosquito trends or the influence of weather. The Ridge Regression model performed admirably in some situations but had too small a range of predictions. Actual mosquito counts in the validation set varied from 1 to 929, where Ridge only predicted up to 49. Similarly, Random Forest and an LSTM Neural Net made predictions within a tight range. Only a Multi-Layered Neural Network captured the range of mosquitos, though it had only the second best MSE of all the models.

The reason those model types were selected was because they were not only robust to a large number of variables, but also because they learn in different ways. While no individual model performed exceptionally well, combined they improved dramatically. As a result of a basic Nelder-Mead model weighting optimization, the RMSE fell from 48 (Ridge, the best individual model by this metric) to 44.

The biggest gains, though, were obvious in evaluating when a spike in mosquito counts would occur. The Neural Net model was good at predicting spikes but overemphasized them. Other models were good at predicting local averages and minor trends. By combining the models, not only was the overall accuracy improved, but the predicted spikes in mosquitos coincided with actual spikes. As a result, recommendations for when and where to spray will be very effective.

Model Type	MSE
Ridge Regression	48
Neural Net	130
Random Forest	48
LSTM	52
Assembled	44



Classification Models

Classification models are being developed to use the outputs of the regression models (predicted mosquito count), along with other metadata, to predict WNV presence. The determination of WNV is thus a binary classification problem.

SMARRT Consulting is using a variety of classification tools to predict WNV. Some of the models being investigated are:

1. Elastinet
2. Lasso Logistic Regression
3. C5.0 Trees
4. Recursive Partition Trees
5. Random Forest
6. XG boost
7. Extra Trees

Each modeling approach offers strengths in different aspects. While approaches 1 and 2 offer a high degree of explainability - why the model is predicting the results, which variables contribute to what extent - allowing for subject matter experts to critique the model, often, for complex data, these models fall short in predictive power. Models 5 through 7, have the opposite characteristics - they are very strong predictive models, but offer limited explainability. These types of models are called *Ensemble Models*, meaning they are a combination of 100s of smaller and weaker tree-based models. Though each tree in the ensemble is weak - i.e. on its own it has limited predictability, when combined together, the ensemble eliminates the weaknesses and enhances predictive power. Since the objective for SMARRT Consulting is to predict the number of WNV cases, our key goal is predictive power, even at the cost of explainability.

SMARRT Consulting used weather, geospatial, socioeconomic status indicator, neighborhood, trap information and the total mosquito count as predictor variables for the classification modeling. We did not use the mosquito species as a predictor variable since it was perceived to be quite difficult to forecast. We discuss in the “Combined Model” section why and how we had to modify this approach.

A typical means for evaluating binary classification problems is to evaluate the performance by calculating the *accuracy*. This is the percentage of correctly predicted labels over all predictions. However, this metric works well on balanced datasets - i.e. on datasets where both classes have roughly equal number of observations. This metric’s validity is compromised when used on an imbalanced dataset like the one we have, where ~ 10% of the observations have WNV. This metric can be misleading by yielding a model with high accuracy when predicting the class of unimportance but performs poorly on class of interest that is less represented in the data. Thus, we will use metrics such as Kappa, False Positive Rate, False Negative Rate, Area Under the Curve (AUC) of a Receiver Operating Curve and a custom cost metric for WNV infections. Some of these terms are explained further in the appendix under classification model metrics, while the cost metric calculation is explained a bit later.

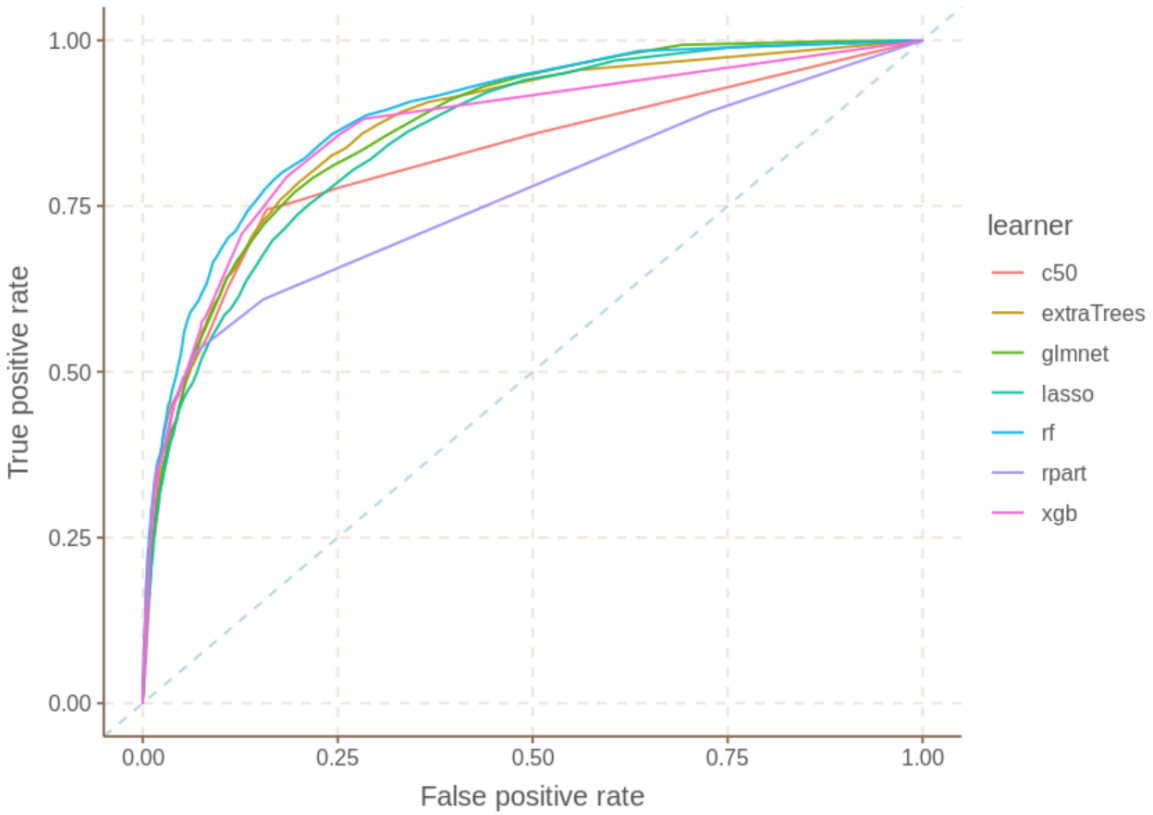
Baseline models were constructed to give a standard to compare future models as well as to provide insight into variable importance. Random forest was used without any tuning to see what variables showed the greatest promise of predictability. Stepwise variable selection was used for the logistic regression model to aid in variable reduction.

The Classification Model Comparison Table below shows the performance of the models for a 3-fold Cross Validation resampling approach.

Classification Model Comparison					
Techniques Used	**Validation Set Performance Metrics**				
Model	**Misclassification Error**	**True Positive**	**False Positive**	**False Negative**	**AUC**
C5.0 Trees	0.166	0.730	0.153	0.270	0.816
Elastinet	0.132	0.621	0.101	0.379	0.872
Lasso Logistic Regression	0.168	0.662	0.146	0.338	0.858
Recursive Partition Trees	0.113	0.530	0.068	0.470	0.756
Random Forest	0.139	0.718	0.120	0.282	0.890
XG Boost	0.888	1.000	1.000	0.000	0.869
Extra Trees	0.149	0.680	0.128	0.320	0.871

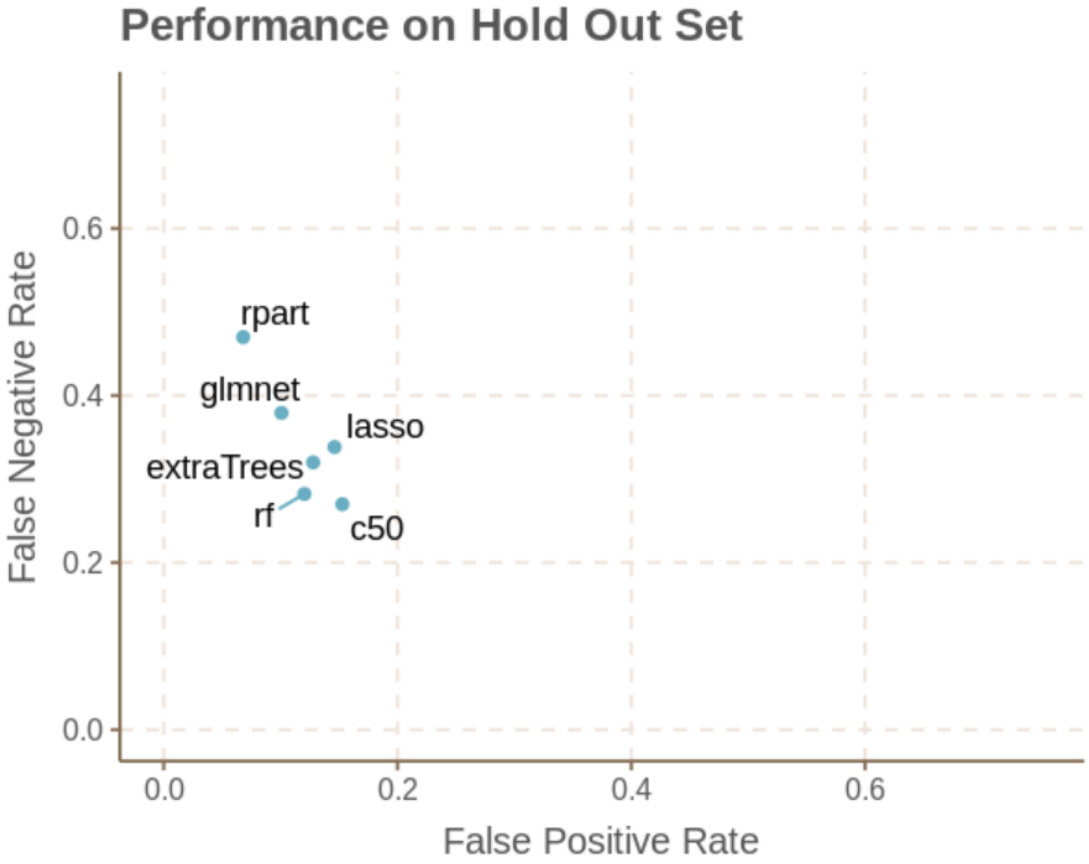
The ROC curve in *Figure 9* shows the performance of all the models. In the ROC plot, the models with the curves closes to the top left corner are strongest performers. Apart from the *rpart* model, the rest of the three models are quite close contenders, yet the top performing model is the randomForest model.

Figure 9



Though AUC is a quick way to compare model performance, a closer look at two metrics - False Positive Rate and False Negative Rate gives us an interesting insight into the relative model performances. *Figure 10* shows this relationship for all the classification models built.

Figure 10



SMARRT Consulting has performed two approaches to tune and improve the classification models:

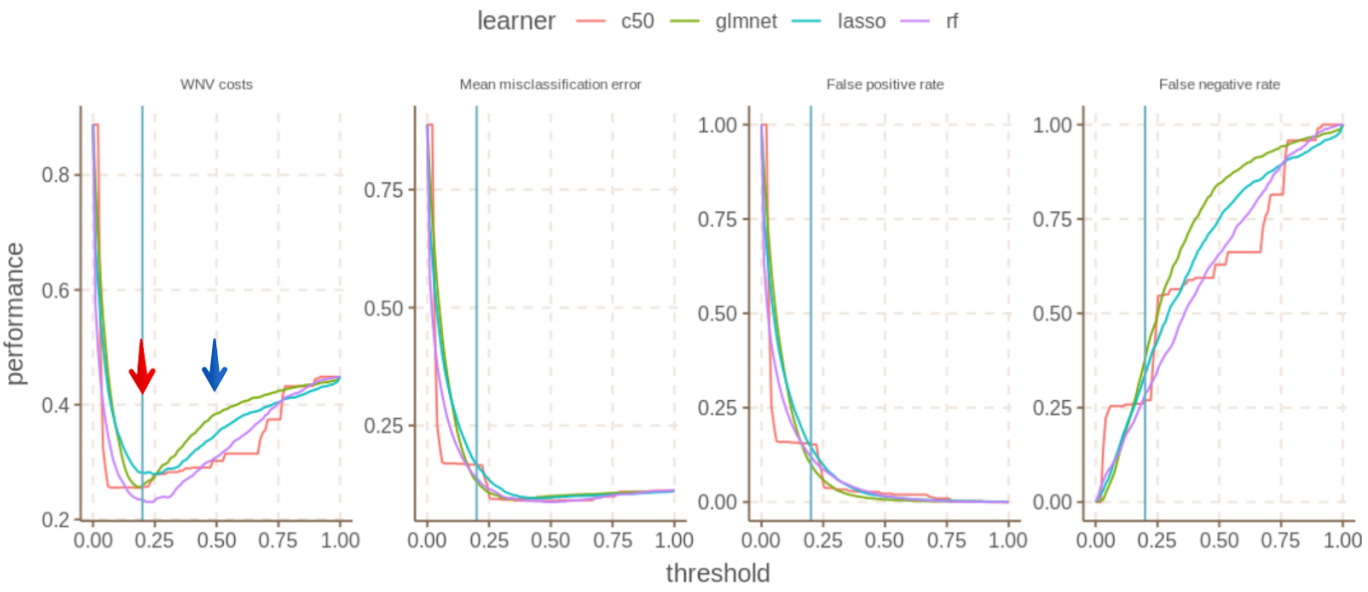
1. **Hyperparameter Tuning**

For the top performing models, we carried out tuning of the hyperparameters. For example, a randomForest approach needs selection of model parameters like the number of variables to randomly select per iteration, or the total number of trees to build in the forest. For elastinet models, we need to select the appropriate balance between the two penalty functions (ridge and lasso). The search for such parameters is automated using grid search combined with a stratified 3-fold cross validation approach.

2. **Incorporation of Class-Dependent Misclassification Costs**

This is discussed in the next section.

Figure 11



The performance of the randomForest classification, after hyperparameter tuning, and cost-based thresholding is shown in *figure 12.* The model gets 78% of the infections correctly identified, on average, for the test portions of the 3-fold cross validation resampling approach. We can see that the large cost on false-negatives means that we only 3% of *Clean* predictions are Infected in truth. [For the default threshold of 0.5, this number was 7%].

Figure 12

		Prediction	
		Clean	Infected
Truth	Clean	9062	1229
	Infected	355	945

		Prediction	
		Clean	Infected
Truth	Clean	78%	11%
	Infected	3%	8%

Classification Modeling Business Impact

An important impact which SMARRT Consulting will consider while tuning these models is relative impact of the errors on the City of Chicago. Each type of error - FPR, FNR - also called Type 1 and Type 2 Errors, have a different cost impacts to the city.

False Positives mean that the model predicted the presence of WNV at a certain location & date, however, in reality, no WNV was found. The impact to the city would be that, for these locations, the city would end up conducting activities like eliminating stagnant water sources or spraying to eliminate mosquitoes. There is public perception of a human health risk due to exposure to chemicals used for mosquito abatement although

environmental reviews have concluded that the human risk is low and offsets the health benefits (by reducing mosquito-transmitted disease). Nonetheless, the public would be encouraged to stay indoors and avoid direct exposure to the spraying sites which can cause consternation. What is the cost to the city in this scenario?

False Negatives means that the model did not predict the presence of WNV, however WNV was present. This could cause an unseen breakout if sufficient mosquitoes with WNV bite and spread infection. The cost to the city would be a medical burden of addressing these infections and taking emergency precautionary steps to avoid further WNV spread. What is the cost to the city in this scenario?

An additional consideration is that all models are developed under the presumption that there are ongoing, unobserved mosquito abatement efforts in the form of mosquito spraying, public education, and removal of stagnant water. Any areas identified as having elevated risk are considered such *even after* these abatement efforts are underway, and the predictions assume that future abatement efforts are similar to past efforts. Any change to spraying strategies, for example, will impact model performance. This must be considered when making decisions about the tradeoff between False Positives and False Negatives.

SMARRT Consulting had to estimate the relative costs of these errors to the city. We found some evidence that the cost of roughly $1.9 m per year in damages for WNV infections (At an average of $21,000 per person per infection). On the other hand, annual spraying costs are around $1.1 m per year, which reduces the occurrence by WNV by 65%. Estimation of the relative costs is a complex project in itself. To be on the conservative side, our team has proceeded with an assumption of a 4:1 cost, as shown on the side. These are *relative costs*, used to tune the posterior probability thresholds for the classification models.

		Prediction	
		Infected	Clean
Truth	Infected	0	4
	Clean	1	0

The default threshold cutoff for posterior probabilities - shown by the blue arrow in *Figure 11* - are replaced by an improved threshold of 0.2 shown by the red arrow. As we can see, a custom cost function build to minimize the cost of incorrect predictions as weighted by the 4:1 cost matrix is shown on the very left hand side graph. We can see that using an improved threshold of 0.2, the randomForest model has the lowest cost to the city.

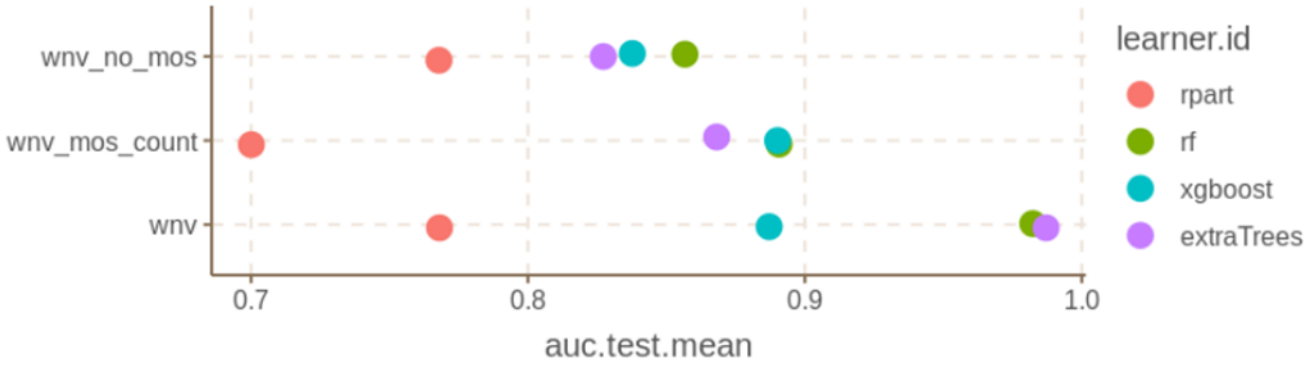
Combined Models

The original intent of the modeling approach - as shown in the modeling flow - was to predict the total number of mosquitoes, per trap, for each date (using regression modeling), and feed this as in input into the classification model. The classification model uses this variable as one of its key predictors. As we did showcase in the regression section above, the regression modeling was quite challenging using the current predictor set. The lower quality predictions on the number of mosquitoes would have an adverse effect on the classification model.

To overcome this challenge, the approach we have adopted *in the interim* is to eliminate the mosquito count from the classification model, to keep the quality of the classification model outputs under check. We do expect new classification models without this predictor to be lesser in performance.

SMARRT Consulting has evaluated the performance of both types of models and documented them here. *Figure 13* shows the relative performance of 4 modeling approaches, using 3 types of input datasets. *DS1: wnv_no_mos* uses predictors without any mosquito dependent predictors, *DS2: wnv_mos_count* adds the mosquito count predictor, while *DS3: wnv* adds even more mosquito predictor variables, viz species-specific counts. Switching from *DS2* (models described above) to *DS1* drops the AUC by ~3.5% to 0.85. The team believes this is an acceptable drop in model performance while avoiding the risk of uncertain mosquito count predictions.

Figure 13



On the other hand, if we were to add predictions at a species level, we can prove that we could improve test set AUC by ~11% to 0.97.

Conclusions

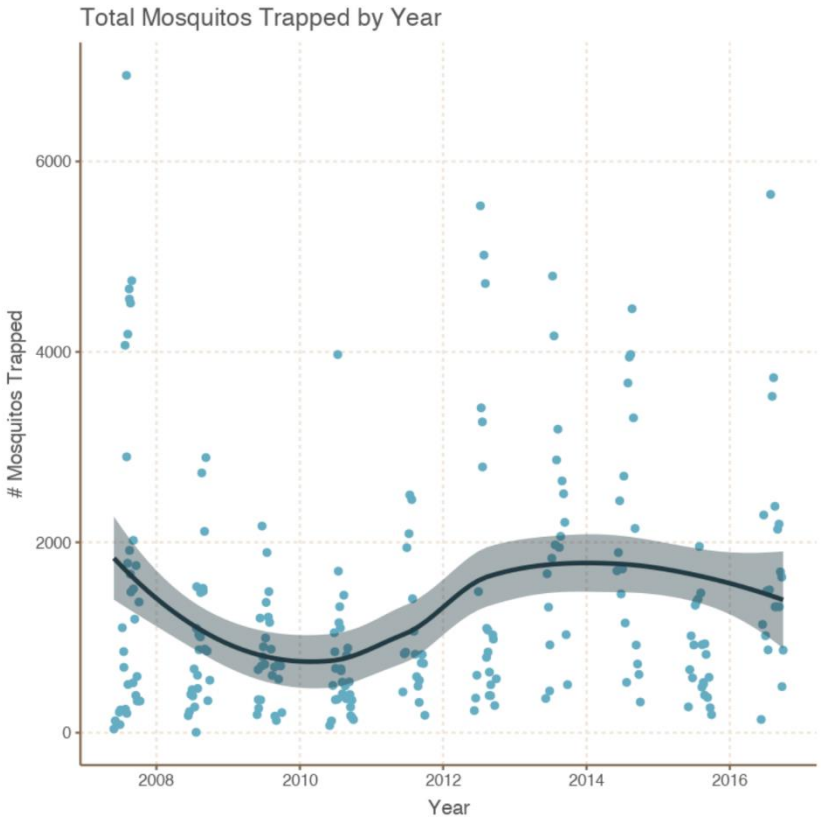
Modeling

SMARRT Consulting has demonstrated that we can effectively identify, gather and combine disparate data sources to make effective predictive models for identifying temporally and geospatially specific mosquito and WNV risk. However, there are many difficulties encountered with predicting mosquito levels. There is a strong seasonal component to mosquito population growth (Figure 14), but it is moreover tied to specific environmental conditions including sustained daily temperatures (heat), presence of stagnant water during breeding periods (precipitation and physical characteristics of neighborhoods), and drought conditions that increase mosquito activity and contact with birds and humans later (low precipitation).

SMARRT Consulting demonstrated that model performance improves when considering as much of these data as possible and using proxies for characteristics that are not easily measured directly (e.g. measures of neighborhood poverty and buildings with vacancies and violations). Despite that, predicting mosquito counts is very difficult. One of several key challenges is that ongoing mosquito spraying efforts actively target neighborhoods with active WNV mosquito populations, so the signal for mosquito presence in the data is slim. The natural course of unchecked mosquito population growth is rarely observed. Since high risk neighborhoods are already being targeted, it is difficult to extract additional information that will be of value for targeted spraying.

In addition, weather data are highly important, so models that use known weather data (observed historical data) will outperform models that rely on weather forecasts. Making predictions for an entire mosquito season or calendar year is problematic since forecasts are not typically reliable far into the future. Making these predictions could provide high level information to Chicago Department of Public Health when planning for anticipated resource levels (e.g. volume of mosquito spraying that will need to be conducted during the course of the mosquito season and budgeted appropriately). Making short term predictions will result in more accurate and useful models. Depending on how far in advance one forecasts, conventional time series methods such as ARIMA and ETS may actually provide business value.

Figure 14



Dashboard Visualization

Maps are a powerful tool that can be used to effectively visualize geospatial data. SMARRT Consulting group will deliver several maps to support the findings from our analysis. The maps will be incorporated into different Tableau dashboards that will be interactive. This will enable the CDPH to do some self-service analytics within the environment we create.

The main type of map we plan to deliver are heat maps. Heat maps are a graphical representation of data that transform the quantitative data into color. Typically, these colors are on a spectrum that is ranging from red to yellow to green, with the different color gradients representing a different numeric value. The main benefit to using heat maps is the ability to quickly ingest a large amount of data. Looking at a list of Chicago neighborhoods or zip codes and the number of West Nile virus infections in each area would be a lot of data for the naked eye to comprehend. However, looking at that same data on a map that is divided into Chicago's neighborhoods or zip codes and color coded according to the number of West Nile virus infections, is a completely different story. On the map, it will be very easy to immediately identify pockets where the number

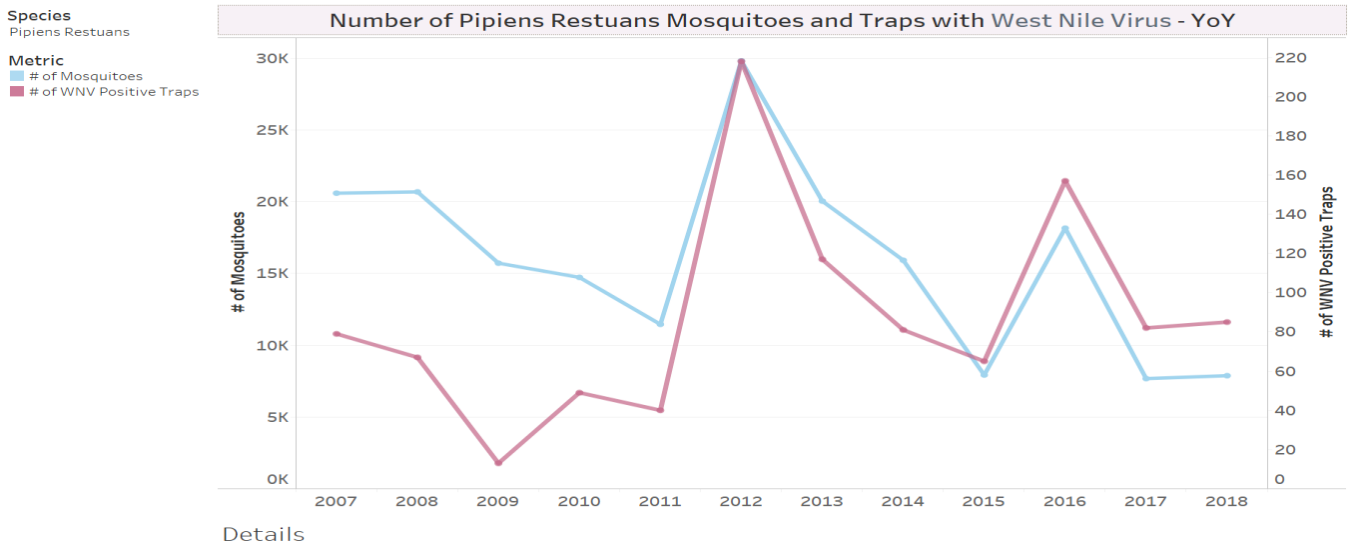
of West Nile virus infections are very high (in red) or very low (in green). Combining the color dimension with the geospatial will allow very simple ease of use for the CDPH.

Other aspects we will incorporate into our dashboards are markers to indicate high risk areas. We are defining high risk areas as those where a large outbreak of West Nile virus infections would have the worst impact due to a more sensitive population group. This would include areas with a lot of schools or daycare centers or areas with retirement homes.

One way this can be done is using the interactive hover feature that is built into Tableau. Following along the same example as earlier, if you were to hover your mouse over a certain neighborhood or zip code that had a high number of West Nile virus infections and was colored red, a small text box would appear and give you additional information on that particular area. This information could include the population in that area, whether a spraying was done, the number of schools and daycares, the number of retirement homes, etc. There are several options for what types of supplemental data can be provided in these text boxes.

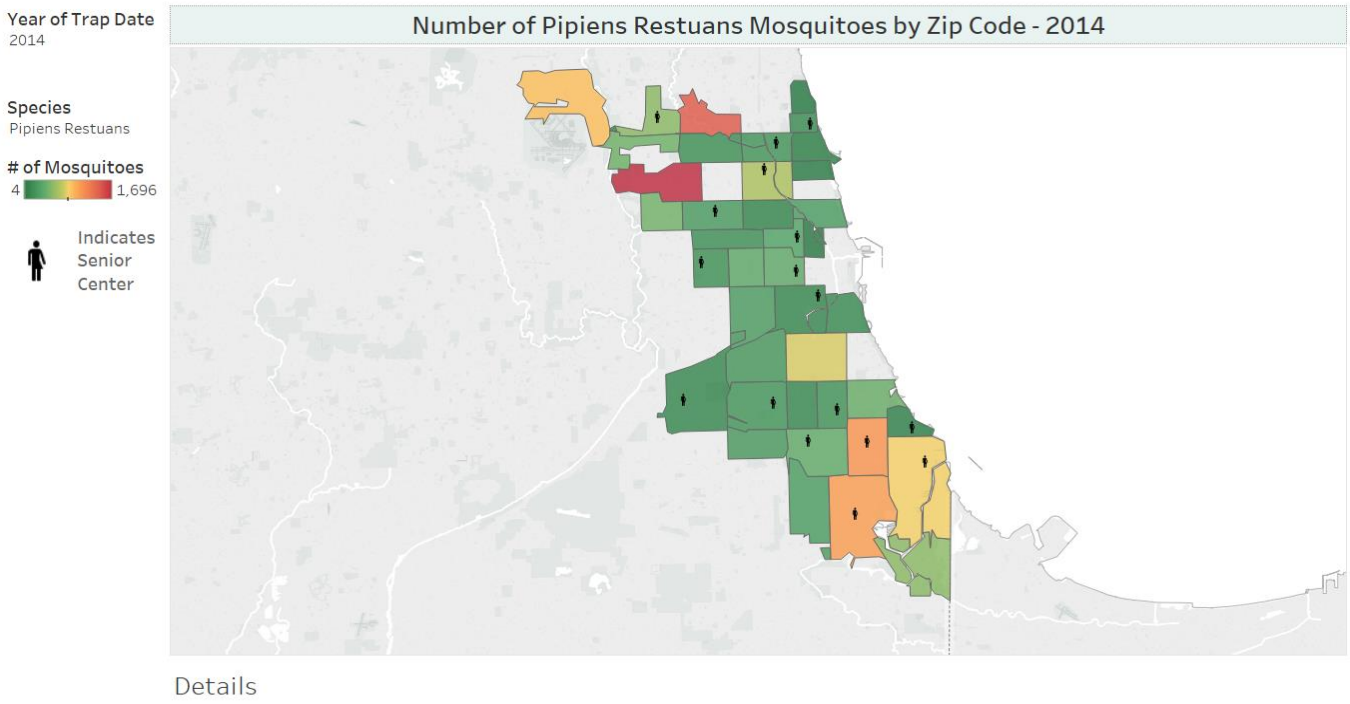
Another more visual option for indicating these high-risk areas would be to overlay small pictorials in the areas. For instance, a small clipart picture of a school in school zones.

So far, we have been able to successfully complete a few our initial deliverables. We currently have three separate dashboards, each with a different objective. The first is just a simple look to show over the 12-year time frame (2007-2018) the total number of mosquitoes vs. the number of positive WNV traps by species. The species can be selected (control is in the top left corner) and the dashboard will update. The idea with this dashboard is to help orient the user to the magnitude of the data that they will be viewing on the heat maps. See below example for the Pipiens Restuans mosquito species.



The next two dashboards are where we will focus the bulk of our time and attention. The first dashboard shows the number of mosquitoes aggregated at a zip code level in a heat map. This map is also able to filter

specifically on a certain mosquito species or to look at all species together. You also have the functionality to change what year you are looking at. We have been able to incorporate the senior center locations onto this map, notated by a person pictorial. See below example for year 2014 for the total number of mosquitoes in mosquito species Pipiens Restuans.



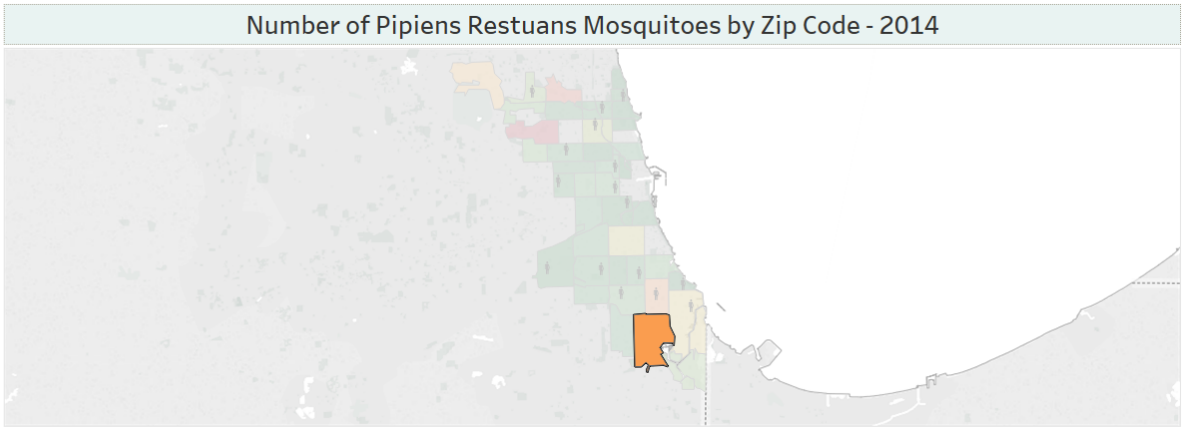
An additional feature we have on the visualization is the “Details” section on the bottom. This section will expand when you click on a zip code or highlight multiple zip codes to give the user additional information for that area in particular. These details include the zip code, community/neighborhood, trap year, trap quarter, trap type, trap name, species, and number of mosquitoes. See below example when you click on zip code 60628.

Year of Trap Date
2014

Species
Pipiens Restuans

# of Mosquitoes
4 1,696

Indicates Senior Center



Details

Zipcode	Community	Trap Year	Trap Quarter	Trap Type	Trap Name	Species	# of Mosquitoes
60628	WEST PULLMAN	2014	2	GRAVID	T135	Pipiens Restuans	372
			3	GRAVID	T135	Pipiens Restuans	114
			4	GRAVID	T135	Pipiens Restuans	13
	PULLMAN	2014	2	GRAVID	T102	Pipiens Restuans	11
			3	GRAVID	T102	Pipiens Restuans	285
			4	GRAVID	T102	Pipiens Restuans	1
	RIVERDALE	2014	2	GRAVID	T221	Pipiens Restuans	36
			3	GRAVID	T221	Pipiens Restuans	141
	ROSELAND	2014	2	GRAVID	T095	Pipiens Restuans	18
			3	GRAVID	T095	Pipiens Restuans	63
			4	GRAVID	T095	Pipiens Restuans	3

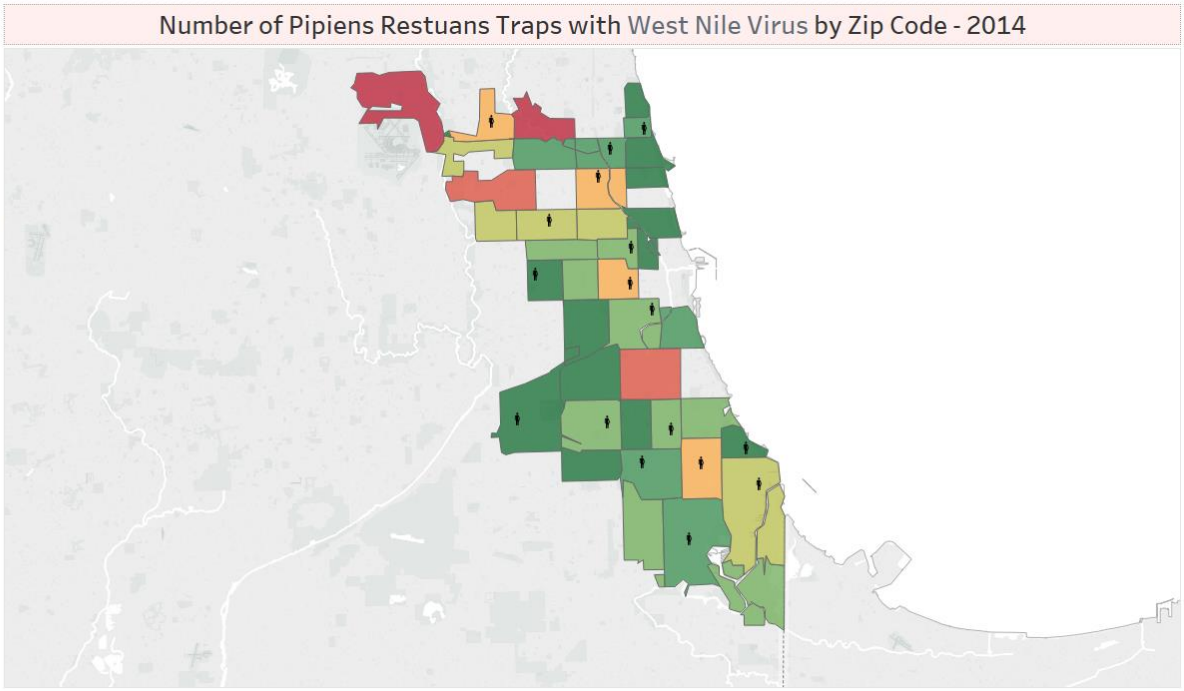
The third visualization is set up and works the same as the second visualization except that instead of showing the number of mosquitoes, it is showing the number of positive WNV traps. You will have the same functionality to filter on year and mosquito species and will again see the pictorials indicating senior centers. See below example for year 2014 for the total number of positive WNV traps for mosquito species Pipiens Restauns.

Year of Trap Date
2014

Species
Pipiens Restuans

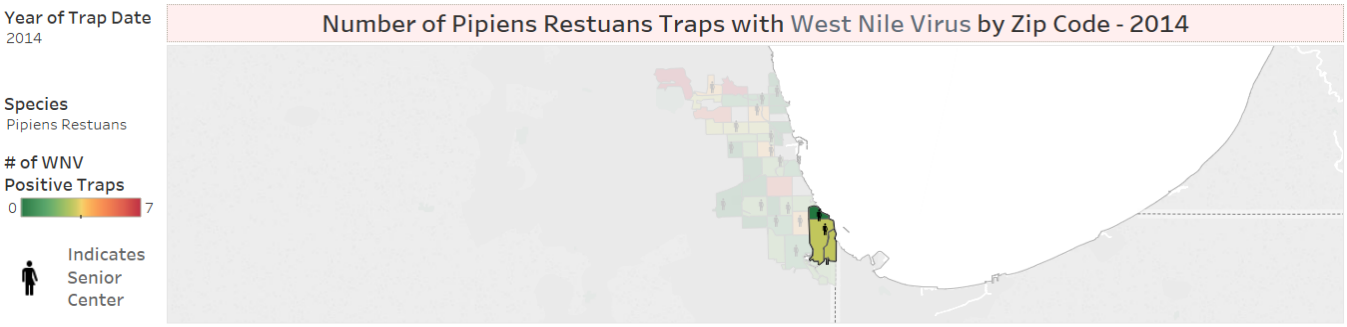
# of WNV Positive Traps
0 7

Indicates Senior Center



Details

Again, similar to the second dashboard, you can select a zip code or group of zip codes to see additional details. These details include the zip code, community/neighborhood, trap year, trap quarter, trap type, trap name, species, and number of WNV positive traps. See below example when you select zip codes 60617 and 60649.



Details

Zipcode	Community	Trap Year	Trap Quarter	Trap Type	Trap Name	Species	# of WNV Positive Traps
60617	SOUTH DEERING	2014	2	GRAVID	T128	Pipiens Restuans	0
					T138	Pipiens Restuans	0
			3	GRAVID	T128	Pipiens Restuans	1
					T138	Pipiens Restuans	2
			4	GRAVID	T128	Pipiens Restuans	0
					T138	Pipiens Restuans	0
	HEGEWISCH	2014	2	GRAVID	T212	Pipiens Restuans	0
			3	GRAVID	T212	Pipiens Restuans	0
			4	GRAVID	T212	Pipiens Restuans	0
	SOUTH CHICAGO	2014	2	GRAVID	T099	Pipiens Restuans	0
			3	GRAVID	T099	Pipiens Restuans	0
			4	GRAVID	T099	Pipiens Restuans	0
60649	SOUTH SHORE	2014	2	GRAVID	T079	Pipiens Restuans	0
			3	GRAVID	T079	Pipiens Restuans	0
			4	GRAVID	T079	Pipiens Restuans	0

The next steps in our dashboarding/visualization work will be to drill down further into the data to try to incorporate the longitude and latitude data instead of aggregating at the zip code level. We are also working to include the indicators on the maps for schools and hospitals. Ideally, we will be able to combine the dashboards into a more streamlined solution for our users. We are hoping that our final product can also include our forecasted projections for future years from our modeling team.

SMARRT Consulting group is dedicated to working closely with the CDPH to deliver the solution that would be most beneficial.

Mobile Application

The primary objective of SMARRT Consulting Group is to decrease the amount of West Nile virus infections in a cost-effective way. According to the Center for Disease Control and Prevention (CDC) “The most effective way to prevent infection from West Nile virus is to prevent mosquito bites. Mosquitoes bite during the day and night. Use insect repellent, wear long-sleeved shirts and pants, treat clothing and gear, and take steps to control mosquitoes indoors and outdoors.” By providing accurate and real-time forecasts on high risk areas for testing positive for WNV, the public can take preventative measure against infection from WNV. SMARRT Consulting Group will deliver the WNV forecasts to the public through a dashboard on a public website as well as through a free mobile application.

WNV forecasts will be determined through two methodologies. A series of models will determine the likelihood of an area to test positive for WNV. The severity will be identified through a red, amber, or green monitoring system (RAG). The RAG status will be determined by calculating the proximity of forecasted WNV to a school, senior center, or hospital. These establishments are used since children, the elderly, or people with weakened immune systems are more likely to contract WNV.

The data engineering team processed several WNV files and produced a WNV master file. This is the primary file used in the applications. Additional input files include a Chicago shapefile and several files which include locations of Chicago hospitals, schools, and senior centers. The modelling data will be integrated into the master WNV file for ease of use.

The mobile applications are divided into three sections: Cluster, Heat-map, and Graph tabs. The clusters tab shows a dynamic map of Chicago and provides longitude and latitude coordinates of historical and forecasted locations for positive cases of WNV. The heat-map of Chicago overlays the severity for WNV; red areas indicate a high severity and green indicating low severity for WNV. The graph tab shows key WNV metrics such as total cases identified, current forecast (RAG), and areas most likely to test positive for WNV. ***(Figure Shiny Dashboard & Figure Mobile App in appendix)***.

SMARRT Consulting has demonstrated the ability to render the modelled WNV data in an "easy to understand" format and provide targeted alerts the public. The severity methodology has been revised from its original scope and only targets hospitals. Initially, three segments of the population were to be targeted and assigned a Red, Amber, or Green severity status. Schools and senior centers have been removed from the initial deliverable. The high number of schools caused performance issues with the application. Senior centers have been excluded due to insufficient data. Both will be included in future enhancements.

To provide additional context for the user, a dynamic line graph has been included on the cluster map. When the user selects a hospital, the line graph provides up to six months of mosquito history for the nearest trap from the hospital. When the user selects a trap marker the application provides up to six months of mosquito history for that particular trap. A link to the application has been provided in the appendix.

The Team

The SMARRT consulting group is a diverse team consisting of professionals from a range of educational backgrounds with a wealth of practical experience. Our team comes from a variety of industries including: R&D, academia, finance, automotive, operations, and analytics. We know that this team and all the skills that we bring to the table will be able to exceed your expectations on this project.

Andrew Cooper has a Master of Public Health in epidemiology and has 15+ years of work experience as a statistical analyst and manager of a software development team that developed web-based data entry, data management and analytic tools. In addition to contributing to analytic design, he will be involved with identifying data sources; obtaining, cleaning and reshaping data; and building predictive models.

Rachel Dudle is a new addition to the SMARRT Consulting Group. While she has only 4+ years' experience, she has established herself with strong skills in building visualizations and dashboards. Her experience ranges from the Financial sector, to manufacturing to pharmaceuticals.

Stephen Hage has a diverse career background, having worked in operations, sales, marketing and analytics. His strengths as a modeler and data storyteller will help this project mature from concept to effective product. He will be a bit involved with most aspects but will also be the sales and marketing lead for SMARRT Consulting Group.

Ted Inciong has an M.S. in Information Technology from the Illinois Institute of Technology and has 10+ years of experience in the Financial sector.

Mike Kapelinski is a Project manager that has a M.S. in Biotechnology and 9+ years working in a variety of fields across science such as clinical oncology, pharmaceutical manufacture, and R&D. He brings domain knowledge of the sciences and experience leading a variety of projects to help this team deliver goals on time and above expectations.

Rahul Sangole has 11+ years of work experience in the Automotive industry, primarily in Engineering, Quality and Analytics, leveraging both predictive modeling and six sigma to drive organizational changes using analytics.

Recommendations

A strong surveillance system is imperative in the prevention of WNV. In the absence of vaccines, WNV prevention depends on keeping infected mosquitoes from biting people. SMARRT Consulting is able to prevent infection by accurately predicting the location of WNV and providing alerts to city officials and the general public so that mitigating actions can occur.

SMARRT Consulting recommends that the models we built and deployed, the mobile application and visualizations be used to guide continued mosquito abatement activities. Identifying specific neighborhoods that are at risk, particularly those where spraying efforts are not currently underway and that are adjacent to vulnerable populations, can be very useful to prevent the spread of WNV to the human population. The models can be used to forecast mosquito and WNV risk areas for an entire mosquito season at the start of the year, and they can be updated for near-term risk as the season progresses.

Future Work

The data platform developed by SMARRT Consulting lends itself well to further analyses.

One of the potential next steps is to identify areas not covered by mosquito traps that appear likely to be high risk. Since our models can predict risk for any given latitude/longitude location and time, we can use this information to make recommendations on new trap location placement. Recommendations on trap placement (i.e. locations that could be high risk or signal WNV spreading but which don't currently have traps) will improve the quality of surveillance. Current trap locations distribute traps across the city of Chicago, with traps in 63 of 77 community areas (82%). This scheme provides coverage across most of the city. Additional traps could be placed in areas where mosquito population growth and WNV transmission occurs at key points during the season. Doing so may improve the quality of models. This may be a cost-effective way to identify high risk areas earlier and allow spraying efforts to target them quickly before WNV spreads.

Another future step is to take locations that are known to be positive for WNV and/or have an elevated mosquito population, then attempt to predict risk of it spreading to neighboring areas. This is a slightly different prediction problem than we have focused on to date since it would potentially be limited to locations proximal to known outbreaks.

Additional analytic methods could be tested on the data platform. One such approach is anomaly detection such as methods used in the banking industry. Models that could detect the beginning of an uptick or outbreak would have clear benefits for surveillance and intervention via mosquito abatement. Another approach that we recommend is the use of methods specifically developed for drawing conclusions about latent or unobserved data. Many people are familiar with recommender systems from shopping websites such as Amazon or Netflix. The approaches used in recommender systems may have some applicability to mosquito/WNV predictions where there are many unobserved neighborhood characteristics and abatement efforts that are difficult to track and tie into models.

Appendix

Codebase

The entire codebase is version controlled and available on github at
https://github.com/rsangole/capstone_project

This allows for the project result to be recreated end-to-end, from data to graphics to model results. The code is organized as follows:

- data: Holds all the raw and post processed datasets to be used for visualization, modeling and dashboarding.
    - **raw** holds the original raw datasets,
    - **processed** holds the post processed datasets which go into modeling, EDA and dashboarding.
- **munge:** Holds the scripts which converts the raw data to processed data. This allows for end to end reproducibility. Basically, **raw + munge = processed**
- **images**: Holds images from EDA and modeling activities
- **src**: Holds all the scripts for EDA, plotting, modeling and dashboarding. Does not hold scripts for converting data from **raw to processed state. These scripts reside in munge.**
- **kaggle_original_data :** Holds data and scripts pulled from the West Nile Virus Kaggle competition.
- **docs:** Holds literature research and other project related information
- **reports:** Holds markdown, jupyter notebooks, dashboards and PDF reports created throughout the project

Software & Analytical Tools

SMARRT Analytics will use the following toolkit to complete the project:

- CRAN R 3.3+
- Python 3.5
- RStudio
- Jupyter Lab
- Tableau Desktop
- Microsoft Excel
- Microsoft PowerBI
- SAS JMP Pro 13

Data Dictionary

Data Dictionary	
Cleaned Variable Names	**Data Type**
mos_tot.NumMosquitos=tot.NumMosquitos	integer **- REGRESSION RESPONSE**
mos_any.WnvPresent=any.WnvPresent	logical - **CLASSIFICATION RESPONSE**
t_date=date	date
t_yr=yr	integer
t_mo=mo	integer
t_day=day	integer
t_qtr=qtr	integer
t_wk=wk	integer
t_day.of.yr=day.of.yr	integer
t_day.of.wk=day.of.wk	integer
t_day.of.wk.name=day.of.wk.name	character (factor)
t_eval.day=eval.day	integer
t_eval.wk=eval.wk	integer
part_train=train	logical
part_validate=validate	logical
part_test=test	logical
part_partition=partition	character (factor)
trap_trap.name=trap.name	character
loc_lat=lat	numeric
loc_lng=lng	numeric
loc_lat.lng.src=lat.lng.src	character (factor)
trap_satellite.ind=satellite.ind	logical
loc_ZCTA5CE10=ZCTA5CE10	character
loc_BlkGrp.geoid=BlkGrp.geoid	character
loc_Tract.geoid=Tract.geoid	character
loc_community=community	character (factor)
zone_zone_class=zone_class	character (factor)
zone_zone_type=zone_type	integer (factor)
ses_LT_HS_pct__BlkGrp2017=LT_HS_pct__BlkGrp2017	numeric / float
ses_median_HHInc__BlkGrp2017=median_HHInc__BlkGrp2017	numeric / float
ses_LT_Pov_pct__BlkGrp2017=LT_Pov_pct__BlkGrp2017	numeric / float
ses_LT_HS_pct__Tract2017=LT_HS_pct__Tract2017	numeric / float
ses_median_HHInc__Tract2017=median_HHInc__Tract2017	numeric / float

ses_LT_Pov_pct__Tract2017=LT_Pov_pct__Tract2017	numeric / float
trap_trap_type=trap_type	character (factor)
mos_erraticus.NumMosquitos=erraticus.NumMosquitos	integer
mos_pipiens.NumMosquitos=pipiens.NumMosquitos	integer
mos_pipiens_restuans.NumMosquitos=pipiens_restuans.NumMosquitos	integer
mos_restuans.NumMosquitos=restuans.NumMosquitos	integer
mos_salinarius.NumMosquitos=salinarius.NumMosquitos	integer
mos_tarsalis.NumMosquitos=tarsalis.NumMosquitos	integer
mos_territans.NumMosquitos=territans.NumMosquitos	integer
mos_unspecified.NumMosquitos=unspecified.NumMosquitos	integer
mos_erraticus.WnvPresent=erraticus.WnvPresent	logical
mos_pipiens.WnvPresent=pipiens.WnvPresent	logical
mos_pipiens_restuans.WnvPresent=pipiens_restuans.WnvPresent	logical
mos_restuans.WnvPresent=restuans.WnvPresent	logical
mos_salinarius.WnvPresent=salinarius.WnvPresent	logical
mos_tarsalis.WnvPresent=tarsalis.WnvPresent	logical
mos_territans.WnvPresent=territans.WnvPresent	logical
mos_unspecified.WnvPresent=unspecified.WnvPresent	logical
nbrhud_comm.180d.violation.cnt=comm.180d.violation.cnt	numeric
nbrhud_BlkGrp.180d.violation.cnt=BlkGrp.180d.violation.cnt	numeric
nbrhud_zcta.180d.violation.cnt=zcta.180d.violation.cnt	numeric
nbrhud_comm.180d.vacancies.cnt=comm.180d.vacancies.cnt	numeric
nbrhud_BlkGrp.180d.vacancies.cnt=BlkGrp.180d.vacancies.cnt	numeric
nbrhud_zcta.180d.vacancies.cnt=zcta.180d.vacancies.cnt	numeric
wea_USW00014819_PRCP=USW00014819_PRCP	numeric
wea_USW00014819_tavg2=USW00014819_tavg2	numeric
wea_USW00014819_TMAX=USW00014819_TMAX	numeric
wea_USW00014819_TMIN=USW00014819_TMIN	numeric
wea_USW00094846_PRCP=USW00094846_PRCP	numeric
wea_USW00094846_tavg2=USW00094846_tavg2	numeric
wea_USW00094846_TMAX=USW00094846_TMAX	numeric
wea_USW00094846_TMIN=USW00094846_TMIN	numeric

Description of Numerical Variables

variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
googtrend_deadbirds	0	15257	15257	1.21	0.51	0	1	1	1	3	
googtrend_mosq_bites	0	15257	15257	32.01	18.59	2	18	29	42	81	
googtrend_sym_wnv	0	15257	15257	24.03	25.08	0	8	15	25	100	
googtrend_westnile	0	15257	15257	23.8	23.06	3	9	14	22	93	
mos_erraticus_num_mosquitos	0	15257	15257	0.00046	0.057	0	0	0	0	7	
mos_pipiens_num_mosquitos	0	15257	15257	7.7	75.78	0	0	1	3	2532	
mos_pipiens_restuans_num_mosquitos	0	15257	15257	20.56	50.62	0	2	6	18	914	
mos_restuans_num_mosquitos	0	15257	15257	5.68	15.32	0	0	1	5	338	
mos_salinarius_num_mosquitos	0	15257	15257	0.074	0.57	0	0	0	0	21	
mos_tarsalis_num_mosquitos	0	15257	15257	0.012	0.25	0	0	0	0	11	
mos_territans_num_mosquitos	0	15257	15257	0.24	2.01	0	0	0	0	64	
mos_tot_num_mosquitos	0	15257	15257	34.27	109.42	1	4	11	31	3002	
mos_unspecified_num_mosquitos	0	15257	15257	0.0098	0.19	0	0	0	0	7	
nbrhud_blk_grp_180d_vacancies_cnt	0	15257	15257	0.34	3.56	0	0	0	0	100	
nbrhud_blk_grp_180d_violation_cnt	0	15257	15257	24.09	28.09	0	4	15	34	245	
nbrhud_comm_180d_vacancies_cnt	0	15257	15257	10.63	45.5	0	0	0	0	531	
nbrhud_comm_180d_violation_cnt	0	15257	15257	875.63	928.04	12	180	529	1276	6072	
nbrhud_zipcode_180d_vacancies_cnt	0	15257	15257	16.83	63.39	0	0	0	0	591	
nbrhud_zipcode_180d_violation_cnt	0	15257	15257	1249.97	1046.95	4	262	1036	1967	5743	
ses_median_hh_inc_blk_grp2017	0	15257	15257	48224.45	34322.91	0	22305	42117	72386	155106	
ses_median_hh_inc_tract2017	0	15257	15257	49468.54	31794.35	0	29125	45833	68899	133636	
t_day	0	15257	15257	15.81	8.73	1	9	16	24	31	
t_day_of_wk	0	15257	15257	3.66	1.44	1	2	4	5	5	
t_day_of_yr	0	15257	15257	218.21	31.59	148	194	218	243	282	
t_eval_day	0	15257	15257	1642.54	749.33	514	956	1671	2382	2826	
t_eval_wk	0	15257	15257	234.98	106.97	74	137	239	341	404	
t_mo	0	15257	15257	7.68	1.06	5	7	8	8	10	
t_qtr	0	15257	15257	2.87	0.4	2	3	3	3	4	
t_wk	0	15257	15257	31.63	4.52	22	28	32	35	41	
t_yr	0	15257	15257	2009.9	2.06	2007	2008	2010	2012	2013	
wea_midway_tmax	0	15257	15257	81.87	8.12	58	77	83	87	97	
wea_midway_tmin	0	15257	15257	64.19	7.11	44	59	66	70	79	
wea_ohare_tmax	0	15257	15257	81.73	8.14	57	77	83	88	96	
wea_ohare_tmin	0	15257	15257	62.22	7.35	41	58	64	69	76	

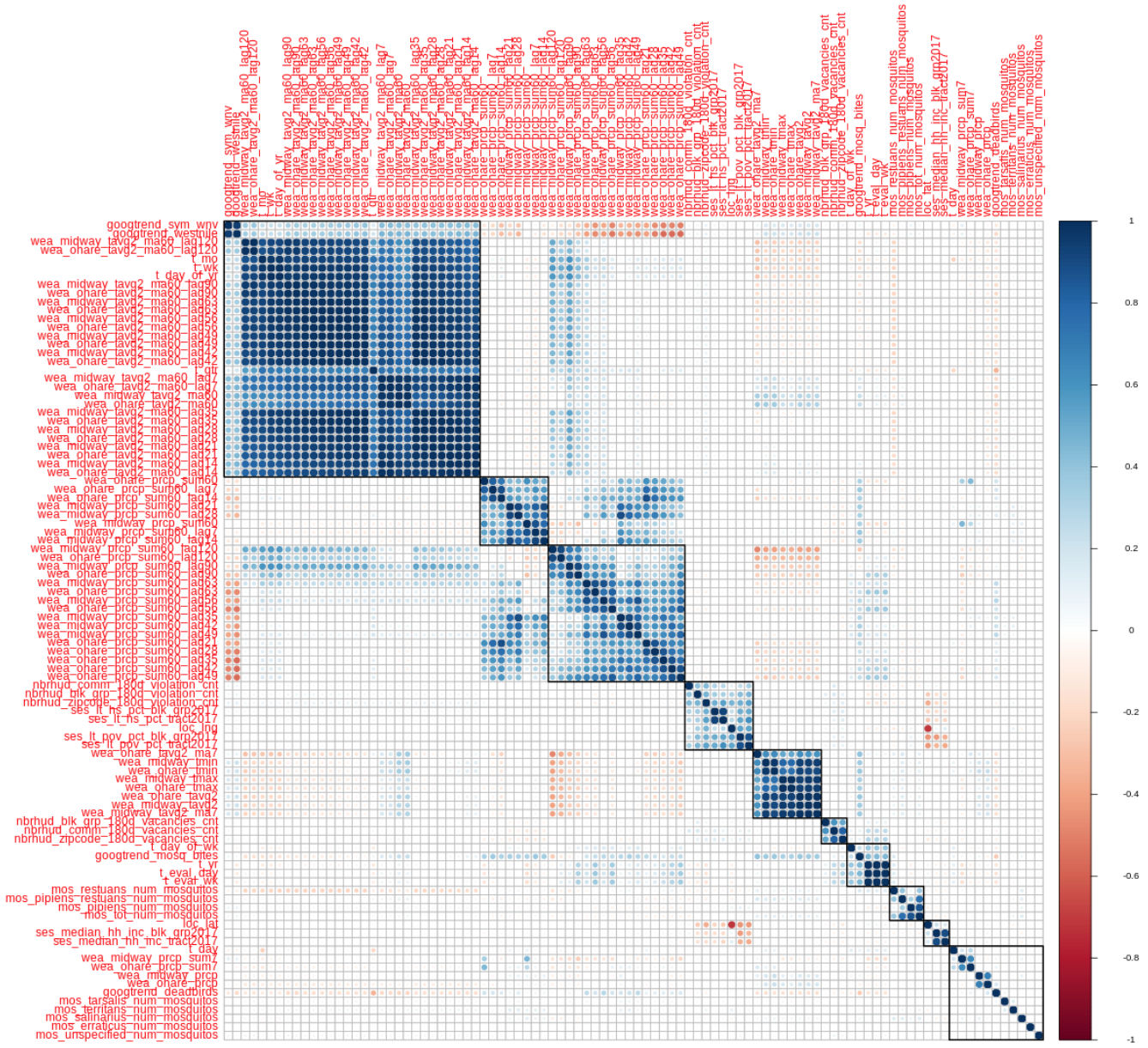
variable	missing	complete	n	mean	sd	p0	p25	p50	p75	p100	hist
loc_lat	0	15257	15257	41.84	0.12	41.64	41.74	41.87	41.96	42.02	
loc_lng	0	15257	15257	-87.7	0.093	-87.89	-87.76	-87.69	-87.64	-87.53	
ses_lt_hs_pct_blk_grp2017	0	15257	15257	14	11.63	0	4.74	11.08	20.58	53.17	
ses_lt_hs_pct_tract2017	0	15257	15257	13.85	11.11	0	4.69	12	20.81	52.12	
ses_lt_pov_pct_blk_grp2017	0	15257	15257	19	17.66	0	2.87	15.87	29.25	66.43	
ses_lt_pov_pct_tract2017	0	15257	15257	18.46	14.89	0	6.01	16.19	26.92	63.8	
wea_midway_prcp	0	15257	15257	0.12	0.32	0	0	0	0.03	3.15	
wea_midway_prcp_sum60	0	15257	15257	8.02	3.76	2.66	4.98	7.43	10.56	19.04	
wea_midway_prcp_sum60_lag120	0	15257	15257	4.79	2.1	1.42	3.17	4.45	6.02	11.73	
wea_midway_prcp_sum60_lag14	0	15257	15257	8.11	3.62	2.66	5.24	7.43	10.56	19.04	
wea_midway_prcp_sum60_lag21	0	15257	15257	7.98	3.5	2.66	5.28	7.14	10.04	19.04	
wea_midway_prcp_sum60_lag28	0	15257	15257	7.96	3.39	2.66	5.41	7.24	10.04	19.04	
wea_midway_prcp_sum60_lag35	0	15257	15257	7.86	3.23	2.72	5.41	7.24	9.79	19.04	
wea_midway_prcp_sum60_lag42	0	15257	15257	7.82	3.09	2.72	5.66	7.14	9.63	19.04	
wea_midway_prcp_sum60_lag49	0	15257	15257	7.67	3.04	2.72	5.44	7.12	8.98	19.04	
wea_midway_prcp_sum60_lag56	0	15257	15257	7.53	2.79	2.72	5.57	6.93	9.21	19.04	
wea_midway_prcp_sum60_lag63	0	15257	15257	7.28	2.67	1.97	5.41	6.71	8.98	19.04	
wea_midway_prcp_sum60_lag7	0	15257	15257	8.04	3.63	2.66	5.05	7.29	10.41	19.04	
wea_midway_prcp_sum60_lag90	0	15257	15257	6.27	2.62	1.42	4.25	6.05	7.85	12.82	
wea_midway_prcp_sum7	0	15257	15257	0.96	1.46	0	0.14	0.49	1.06	8.42	
wea_midway_tavg2	0	15257	15257	73.03	7.23	51.5	69.5	74.5	78.5	86.5	
wea_midway_tavg2_ma60	0	15257	15257	71.73	5.36	53.66	69.36	73.23	74.92	80.4	
wea_midway_tavg2_ma60_lag120	0	15257	15257	38.35	10.42	20.73	29.38	36.14	46.84	61.79	
wea_midway_tavg2_ma60_lag14	0	15257	15257	69.91	6.97	49.96	65.99	72.38	74.78	80.4	
wea_midway_tavg2_ma60_lag21	0	15257	15257	68.59	7.73	47.23	63.86	70.97	74.22	80.4	
wea_midway_tavg2_ma60_lag28	0	15257	15257	67.07	8.56	44.4	61.54	69.37	73.92	80.1	
wea_midway_tavg2_ma60_lag35	0	15257	15257	65.43	9.22	41.15	58.85	67.2	73	80.4	
wea_midway_tavg2_ma60_lag42	0	15257	15257	63.54	9.82	38.19	56.23	65.38	71.76	80.4	
wea_midway_tavg2_ma60_lag49	0	15257	15257	61.55	10.28	36.36	55.02	62.54	69.71	80.23	
wea_midway_tavg2_ma60_lag56	0	15257	15257	59.44	10.65	32.69	52.55	59.75	68.28	79.22	
wea_midway_tavg2_ma60_lag63	0	15257	15257	57.23	10.93	28.85	50.26	57.65	66.21	77.56	
wea_midway_tavg2_ma60_lag7	0	15257	15257	70.95	6.17	51.5	68.42	73	74.88	80.4	
wea_midway_tavg2_ma60_lag90	0	15257	15257	48.16	11.6	23.32	40.5	48.81	56.23	70.97	
wea_midway_tavg2_ma7	0	15257	15257	73.02	7.22	51.5	69.5	74.5	78.5	87.57	
wea_ohare_prcp	0	15257	15257	0.19	0.49	0	0	0	0.16	3.97	
wea_ohare_prcp_sum60	0	15257	15257	9.02	3.55	2.49	5.92	8.93	11.08	19.05	
wea_ohare_prcp_sum60_lag120	0	15257	15257	5.94	2.02	2.55	4.5	5.6	6.9	12.82	
wea_ohare_prcp_sum60_lag14	0	15257	15257	8.97	3.33	2.49	5.98	8.79	11.08	19.05	
wea_ohare_prcp_sum60_lag21	0	15257	15257	8.76	3.19	2.49	5.91	8.7	11.03	15.71	
wea_ohare_prcp_sum60_lag28	0	15257	15257	8.66	3.07	2.49	5.98	8.45	10.66	15.69	
wea_ohare_prcp_sum60_lag35	0	15257	15257	8.53	3.01	2.49	5.98	8.29	10.39	15.68	
wea_ohare_prcp_sum60_lag42	0	15257	15257	8.41	2.87	2.49	5.97	8.15	10.29	14.76	
wea_ohare_prcp_sum60_lag49	0	15257	15257	8.22	2.92	2.49	5.9	8.09	10.09	15.57	
wea_ohare_prcp_sum60_lag56	0	15257	15257	8.03	2.66	2.72	5.97	7.89	9.85	14.76	
wea_ohare_prcp_sum60_lag63	0	15257	15257	7.77	2.57	2.49	5.73	7.23	9.59	14.59	
wea_ohare_prcp_sum60_lag7	0	15257	15257	8.99	3.28	2.49	6.15	8.93	11.16	16.53	
wea_ohare_prcp_sum60_lag90	0	15257	15257	7.05	2.47	2.68	5.28	6.35	8.42	13.84	
wea_ohare_tavg2_ma60_lag120	0	15257	15257	37.18	10.63	18.78	27.99	35.01	46.28	60.53	
wea_ohare_tavg2_ma60_lag14	0	15257	15257	68.87	7.05	48.68	65.03	71.28	74.12	79.29	
wea_ohare_tavg2_ma60_lag21	0	15257	15257	67.55	7.81	45.85	63.18	69.88	73.62	79.29	
wea_ohare_tavg2_ma60_lag28	0	15257	15257	66.03	8.64	43.12	60.39	68.13	72.68	78.93	
wea_ohare_tavg2_ma60_lag35	0	15257	15257	64.38	9.29	39.99	57.75	66.39	71.99	79.29	
wea_ohare_tavg2_ma60_lag42	0	15257	15257	62.49	9.89	37.17	55.51	64.18	70.87	79.29	
wea_ohare_tavg2_ma60_lag49	0	15257	15257	60.49	10.34	35.27	53.77	61.58	68.64	79.11	
wea_ohare_tavg2_ma60_lag56	0	15257	15257	58.38	10.73	31.77	51.44	58.39	66.97	78.33	
wea_ohare_tavg2_ma60_lag63	0	15257	15257	56.15	11.01	27.81	48.98	56.44	65.09	76.99	
wea_ohare_tavg2_ma60_lag7	0	15257	15257	69.91	6.25	50.21	66.97	72.08	74.18	79.29	
wea_ohare_tavg2_ma60_lag90	0	15257	15257	47.06	11.74	21.53	38.88	47.63	55.7	69.88	
wea_ohare_tavg2_ma7	0	15257	15257	72.08	5.96	55.5	68.29	72.14	77	86.71	

Description of Categorical Variables

variable	missing	complete	n	n_unique	top_counts	ordered
loc_census_block_group_id	0	15257	15257	142	170: 1407, 170: 491, 170: 259, 170: 246	FALSE
loc_census_tract_id	0	15257	15257	135	170: 1407, 170: 491, 170: 465, 170: 308	FALSE
loc_community	0	15257	15257	63	OHA: 1575, SOU: 682, HEG: 667, AUS: 663	FALSE
loc_lat_lng_src	0	15257	15257	2	CDP: 13231, Goo: 2026, NA: 0	FALSE
loc_zipcode	0	15257	15257	47	600: 1407, 606: 809, 606: 797, 606: 791	FALSE
mos_species	0	15257	15257	8	pip: 6935, res: 4488, pip: 3148, ter: 457	FALSE
t_day_of_wk_name	0	15257	15257	5	Thu: 6190, Wed: 3676, Sun: 1987, Mon: 1944	FALSE
trap_trap_name	0	15257	15257	179	T00: 246, T11: 234, T16: 234, T13: 233	FALSE
trap_trap_type	0	15257	15257	3	GRA: 14632, CDC: 624, OVI: 1, SEN: 0	FALSE
zone_class	0	15257	15257	37	RS-: 3064, POS: 2614, RS-: 1813, PD : 1407	FALSE
zone_type	0	15257	15257	7	4: 6354, 12: 2815, 5: 2426, 3: 1492	FALSE

variable	missing	complete	n	mean	count
mos_any_wnv_present	0	15257	15257	0.12	FAL: 13476, TRU: 1781, NA: 0
mos_erraticus_wnv_present	0	15257	15257	0	FAL: 15257, NA: 0
mos_pipiens_restuans_wnv_present	0	15257	15257	0.083	FAL: 13993, TRU: 1264, NA: 0
mos_pipiens_wnv_present	0	15257	15257	0.034	FAL: 14736, TRU: 521, NA: 0
mos_restuans_wnv_present	0	15257	15257	0.023	FAL: 14910, TRU: 347, NA: 0
mos_salinarius_wnv_present	0	15257	15257	0.00079	FAL: 15245, TRU: 12, NA: 0
mos_tarsalis_wnv_present	0	15257	15257	0	FAL: 15257, NA: 0
mos_territans_wnv_present	0	15257	15257	0.00013	FAL: 15255, TRU: 2, NA: 0
mos_unspecified_wnv_present	0	15257	15257	0	FAL: 15257, NA: 0
trap_satellite_ind	0	15257	15257	0.023	FAL: 14912, TRU: 345, NA: 0

Correlation Plot



Regression Model Metrics

MSE= Mean Squared Error

RMSE= Root Mean Squared Error

Classification Model Metrics

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative

Accuracy: Total number of correct predictions divided by total number of observations.

Precision: Given all the predicted labels for a given class, how many instances are correctly predicted?

Precision = TP / (TP + FP)

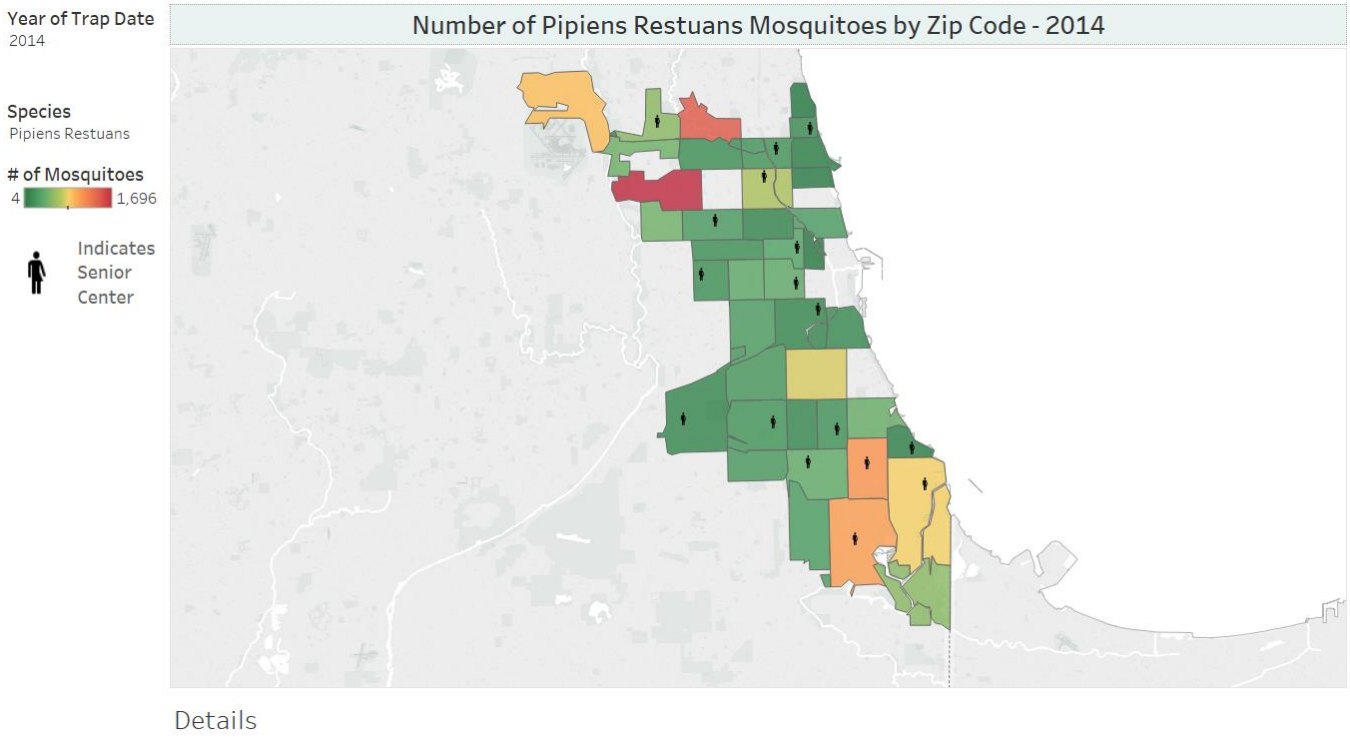
Recall: For all instances that should have a label, how many of these are correctly captured?

Recall = TP / (TP + FN)

Kappa (Cohen's Kappa) is similar to classification accuracy, except that it is normalized at the baseline of random chance on your dataset. It is a more useful measure to use on problems that have an imbalance in the classes.

AUC is the area under the ROC curve. The AUC represents a models' ability to discriminate between positive and negative classes. An area of 1.0 represents a model that made all predicts perfectly. An area of 0.5 represents a model as good as random.

Interactive Dashboard



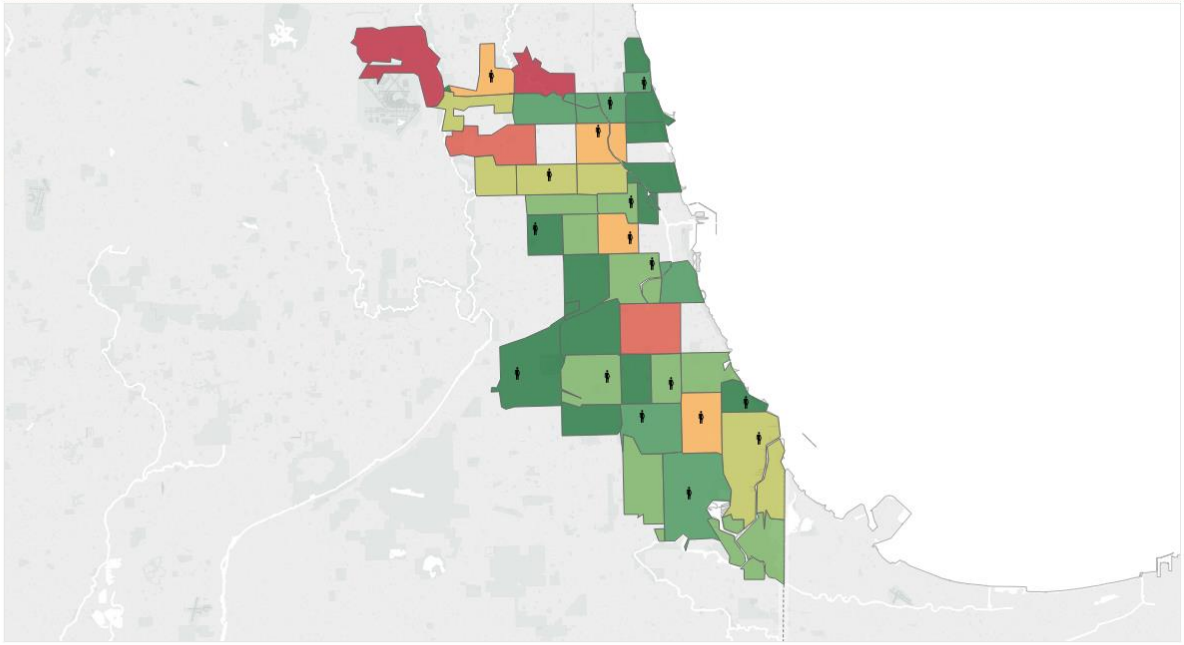
Year of Trap Date
2014

Species
Pipiens Restuans

# of WNV
Positive Traps
0 7

Indicates
Senior
Center

Number of Pipiens Restuans Traps with West Nile Virus by Zip Code - 2014



Details

Mobile Application

Shiny app: https://capstone498-wi19-team54-wnv.shinyapps.io/WNVDashboard/