Evan Meeks
November 6, 2022

Predicting West Nile Virus Rates in Chicago

Project Report - A Capstone Element for Springboard's Data Science Program

Table of Contents

1. Introduction

The West Nile virus (WNV) is the leading cause of mosquito-borne disease in the continental United States. Around 20% of those infected experience mild symptoms, while around 0.6% develop life-threatening complications. (The Center for Disease Control, 2022)

The City of Chicago experienced a city-wide increase of WNV cases and began trapping mosquitos across the city to track and identify hotspots, cases, and trends. This collection phase spanned a 6-year period before the city issued a public cry for help in examining the data and creating a plan for the city to act upon.

1.1 Objective

The objective of the following analysis is to:

- Explore and analyze the data on WNV cases from Chicago, IL.
- Identify features that contributed to WNV case rates.
- Develop a Machine Learning model that accurately predicts the time and location of WNV cases.
- Provide recommendations to the City of Chicago on how to proactively address and prevent WNV outbreaks.

The code for this project can be found here

1.2 Significance

By thoroughly examining the WNV case dataset, we will identify features that affect outbreaks, which can then be used to help predict when and where the cases will occur. This information can then be used by the City of Chicago to help both curb outbreaks and prevent them to begin with.

2. Data

2.1 WNV Cases in Chicago

The data was posted by the City of Chicago to Kaggle, an online Data Science / Machine Learning community and dataset repository. The data was originally posted as a competition in 2013 and can be found here.

Two files from this dataset were used for this project: the "train.csv" and "weather.csv". The train document consisted of 10,506 rows and 12 columns, or features, and consisted of data regarding the date, location, case status and species of mosquito. The weather file consisted of 2,944 rows and 22 features. This file was the aggregate of two weather stations in the city, and described the weather conditions such as temperature, dew point, wind speeds, and precipitation totals at each location daily.

3. Data Cleaning and Wrangling

The purpose of Data Cleaning and wrangling steps are:

- To ensure that all features are of the correct data type.
- To ensure missing values are properly imputed (filled).
- To prepare the dataset for Exploratory Data Analysis and Statistical Analysis.

3.1 Data Type Correction

Most of the columns were of the wrong data type needed for this project and had to be converted. Many of these columns defaulted to "Object" type, where Date Time, Integer, and Floats were preferable. The Date Time format allows for dates to be processed as such, while Integer is a number type, allowing for mathematical operations and other processing. The Float datatype is like integer but allows for higher precision by introducing decimal value storage. The following table details which columns were converted to what:

Column	Data Type
Date	DateTime
Tavg	Integer
Depart	Integer
WetBulb	Integer
Heat	Integer
Cool	Integer
PrecipTotal	Float
StnPressure	Float
SeaLevel	Float
AvgSpeed	Float

Fig 3.1.1

3.2 Incorrect Value and Missing Value Imputation

This dataset was relatively clean and complete, with no completely "blank" values. Some of the issues that had to be corrected were textual placeholders for missing numerical data, as well as zeros put in where a numerical value should otherwise exist. Each column was handled individually, where the solutions ranged from forward filling missing values, to calculated averages or replacing zeros. A combination of native and custom functions were used to accomplish this goal.

3.2.1 Numeric Features

The numeric features host the block, latitude and longitude, and much of the weather data such as temperatures, dewpoint, wet bulb, etc. Missing values in these categories were imputed using their respective medians.

3.2.2 Categorical Features

The categorical features host the date, species, and WnvPresent columns. The date was broken into its day/month/year components, and the months were converted from their numerical value to standard names (for ease of reading.)

4. Data Visualization and Analysis

4.1 Location

The location inside the city proved to be quite impactful, as the latitude and longitude were the 2nd and 4th strongest indicators in the model, with the city block being 6th. This information will prove useful to the City when creating an attack plan and determining the best use case of City resources. The following table shows the 10 highest WNV present rates throughout the data collection process across the city.

Block No.	Street Name	Year	WNV Cases	Street	% Count
46	**N MILWAUKEE AVE**	2013	7	27	25.93
61	**W FULLERTON AVE**	2013	8	33	24.24
82	**S KOSTNER AVE**	2013	8	35	22.86
42	**W 65TH ST**	2013	7	31	22.58
40	**E 130TH ST**	2007	5	24	20.83
71	**N HARLEM AVE**	2013	7	34	20.59
50	**S UNION AVE**	2007	6	30	20
36	**N PITTSBURGH AVE**	2007	7	35	20
58	**N PULASKI RD**	2013	6	31	19.35
65	**N OAK PARK AVE**	2007	6	31	19.35

Figure 4.1.1

(Placeholder for Zip Code Analysis)

4.2 Weather Data

Weather data analysis began on a yearly scale by comparing the average temperature each year against the number of WNV cases. It was immediately evident that warmer temperatures provided higher likelihood of outbreaks. Less volatile temperature fluctuations resulted in more WNV cases, while cooler temperatures and or wider temperature ranges resulted in fewer cases. When looking deeper into weather trends, it was noticed that the minimum, maximum, and average temperatures were all left-skewed whereas the depart was more normally distributed, lending to the idea that temperatures during the collection period were generally higher.

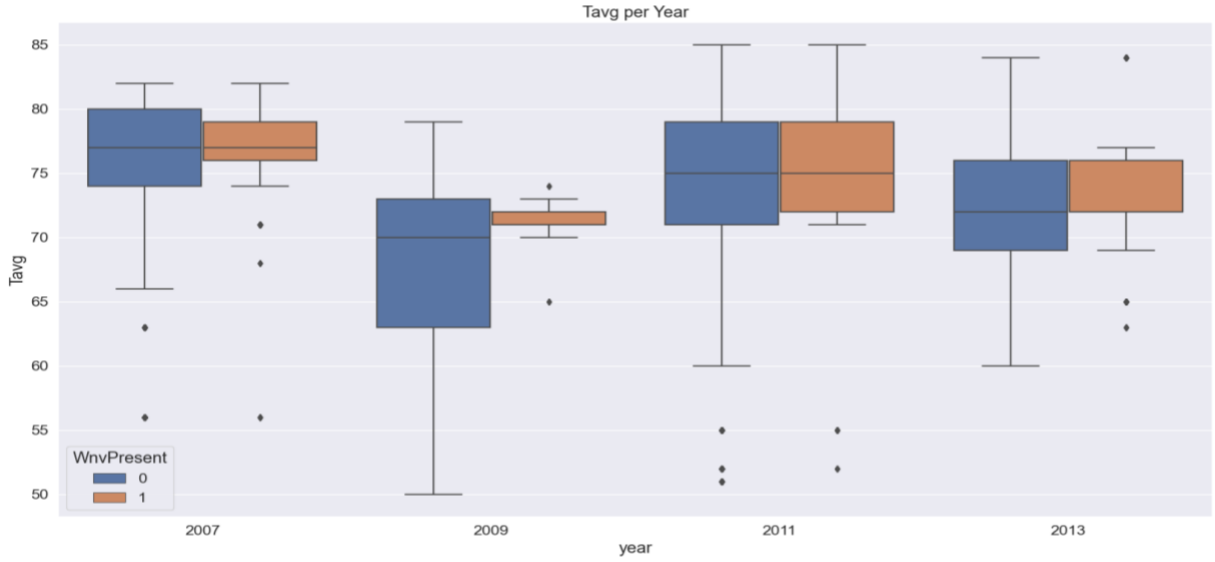

Figure 4.2.1

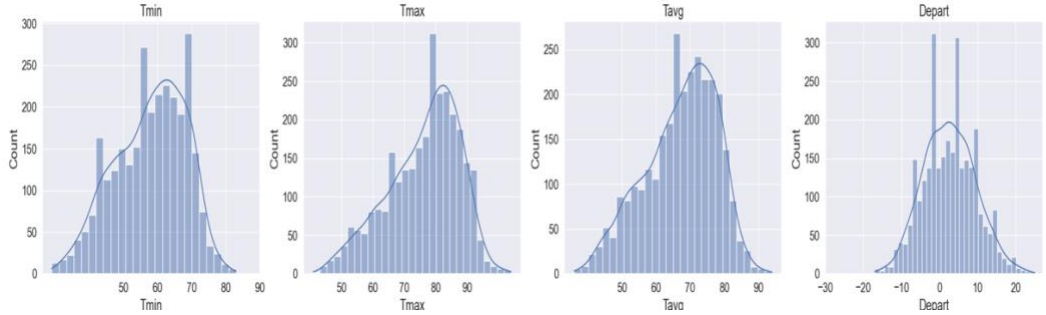

Figure 4.2.2

4.3 Datetime Attributes

As temperature and sunrise/sunset typically follow the season and or month, each of these features were assumed to have at least some dependence on one another. August and September proved to be the worst months for cases, and were also the 1st and 3rd hottest, respectively. The date element of each attribute was broken into day/month/year and renamed, and the sunrise/sunset data was dropped.

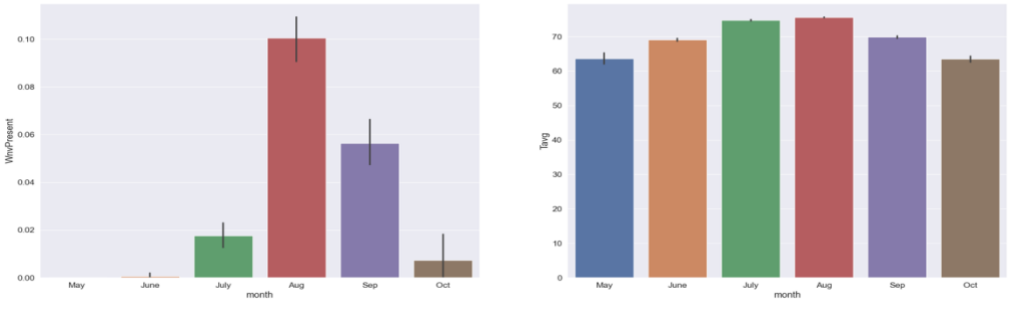

Figure 4.3.1

4.4 Case Rate by Species

Out of 7 species found throughout the city, only 2 species had case rates higher than 2%. Those with lower percentages were dropped from this analysis. Note that several traps lumped the two species together and reported it as one, thus necessitating an extra column for analysis seen in figure 4.3.1. The two species in question, Culex Pipens and Culex Restuans. As the City is only interested in curbing WNV cases overall and not species-specific trends, the remainder of the analysis considers the number of mosquitos as a whole, but not the species individually. The sheer number of mosquitos caught proved to be the single largest influencer in the models—an outcome unsurprising given the nature of probability.

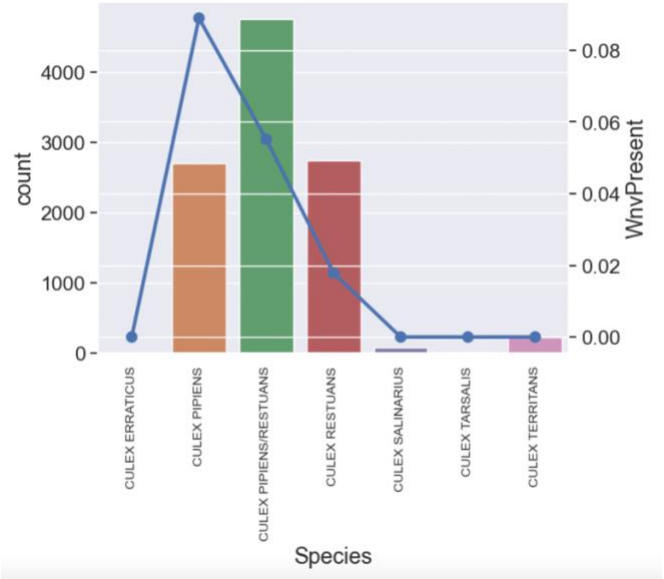

Figure 4.4.1

5 Model Development and Selection

The data between the two weather stations was averaged together before being merged with the training data. From here, the Species, Month and Year columns were encoded for processing inside of different models.

5.1 WOE , IV and VIF

Weight of evidence (WOE) and Information value (IV) are simple, yet powerful techniques to perform variable transformation and selection. (Krishnan, 2018). WOE/IV variables were created and filtered for IV values between 0.01 and 0.08. The resultant features were then used to filter the dataset, and then further examined for multicollinearity using the variance inflation factor (VIF). Once the VIF was calculated, it was used to then limit the dataset further, only including VIF values less than 5.

5.2 Data Modeling

After evaluating WOE/IV data, three standard models were selected for comparison: Random Forest Classifier, Logistic Regression and XGBoost. The model was selected purely based on it's AUC score, a component of the AUC-ROC Curve. AUC - ROC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. The higher the AUC score, the better the model is at predicting (Narkhede, 2018). For this analysis, an AUC score threshold of 0.5 was required.

5.2.1 Random Forest Classifier

A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting (SciKit Learn, 2022). After having instantiated a new instance of the Random Forest Classifier, experimenting with hyperparameters began to achieve the highest possible AUC score. This model came in second place with an AUC score of 0.57

5.2.2 Logistic Regression

In statistics, the logistic model is a statistical model that models the probability of an event taking place by having the log-odds for the event be a linear combination of one or more independent variables. Logistic regression is estimating the parameters of a logistic model (Wikipedia, 2022). This model was used out-of-the-box to achieve the worst AUC score of all at 0.5, and was subsequently dropped from consideration.

5.3.3 XGBoost

XGBoost is an improvement to the Gradient Boosting framework. It gives a prediction model in the form of an ensemble of weak prediction models, which are typically decision trees. The XGBoost model often outperforms Random Forest Classifiers and is widely accepted as one of the most accurate machine learning models. After tuning parameters, this model achieved an AUC score of 0.76, leaps and bounds ahead of the runner up. This model was consequently chosen for this analysis.

6. Conclusion and Recommendations

In this project, we examined the effects of weather, location, and time of year on the case rate of the West Nile Virus in Chicago. Through this analysis, a model was developed that accurately predicts the presence of WNV 86% of the time. Based off our findings, the following recommendations have been made to both treat and prevent further WNV case outbreaks.

1. Most cases occurred east of longitude 87.75 and south of latitude 41.95. This observation shows that many of the outbreaks occurred closer to Lake Michigan, where humidity is higher, and temperatures are warmer. Monitor this area more closely for standing water and treat as necessary. Consider increased monitoring between the 40 and 80 blocks of Chicago as these saw the highest rates in the most recent year of data available.

2. The month of July was the heaviest influencer on model performance out of any month, with June following close behind. This indicates that the weather during this time is most conducive to the gestation and spread of the virus. As such, the city should devote most of its resources to these months to kill off the mosquito population and prevent reproduction.

3. Wind speed and direction also affected the model's outcome. Expect higher mosquito counts during times of low wind or in areas protected from it.

-

Appendix

Appendix I: Original Feature List

Train.csv	Weather.csv
Date	Station
Address	Date
Species	Tmax
Block	Tmin
Street	Tavg
Trap	Depart
AddressNumberAndStreet	DewPoint
Latitude	WetBulb
Longitude	Heat
AddressAccuracy	Cool
NumMosquitos	Sunrise
WnvPresent	Sunset
	CodeSum
	Depth
	Water1
	SnowFall
	PrecipTotal
	StnPressure
	SeaLevel
	ResultSpeed
	ResultDir
	AvgSpeed

Bibliography

Krishnan, S. (2018, April 28). *Medium*. Retrieved from medium.com: https://sundarstyles89.medium.com/weight-of-evidence-and-information-value-using-python-6f05072e83eb#:~:text=Weight%20of%20evidence%20(WOE)%20and,the%20logistic%20regression%20modeling%20technique

Narkhede, S. (2018, June 26). *Understanding AUC - ROC Curve*. Retrieved from Towards Data Science: https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

SciKit Learn. (2022, 11 10). *sklearn.ensemble.RandomForestClassifier*. Retrieved from SciKit Learn: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

The Center for Disease Control. (2022, June 2). *West Nile Virus*. Retrieved from Centers for Disease Control and Prevention: https://www.cdc.gov/westnile/index.html

Wikipedia. (2022, 10 30). *Logistic regression*. Retrieved from Wikipedia: https://en.wikipedia.org/wiki/Logistic_regression