Mosquitoes and West Nile Virus in Chicago

Philip Bradfield - Parisa Yarandi - Ritika Bhasker

January 2017

General Assembly

DC-DSI-3

Abstract

Chicago wants to eliminate the threat of West Nile Virus due to mosquitoes. Given data about weather, time of year, and the location of mosquitoes traps, where should the city spray to most efficiently reach their goal?

Introduction

While west nile virus (WNV) is a disease that can cause severe illness it is fortunate that it does not spread via person to person contact. The primary vector for spreading the disease is mosquito bites. The city of Chicago takes this threat seriously and for many years has worked to learn about the spread of the disease by gathering data about mosquitos from traps located at many locations across the city.

Problem Statement

The daily data includes location of the traps and the number of WNV bearing insects in each trap. This data, along with weather data and an understanding of the life-cycle of a mosquito, can help the city know when is the best time and place to spray pesticides to prevent a flare up of WNV in Chicago.

Given the data provided to us, we wanted to accurately predict when and where the city of Chicago should spray for West Nile, while at the same time being cognizant of limited resources to do so.

Literature Review

There are numerous scientific publications that discuss the tracking of mosquitos. We refer you to “Predicting Culex pipiens/restuans population dynamics by interval lagged weather data” by Karin Lebl Katharina Brugger and Franz Rubel (see https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3660179/) in the journal Parasites & Vectors. This paper influenced our choices of input variables, as it determined that the length of day was a key indicator of West Nile Virus carrying mosquitoes.

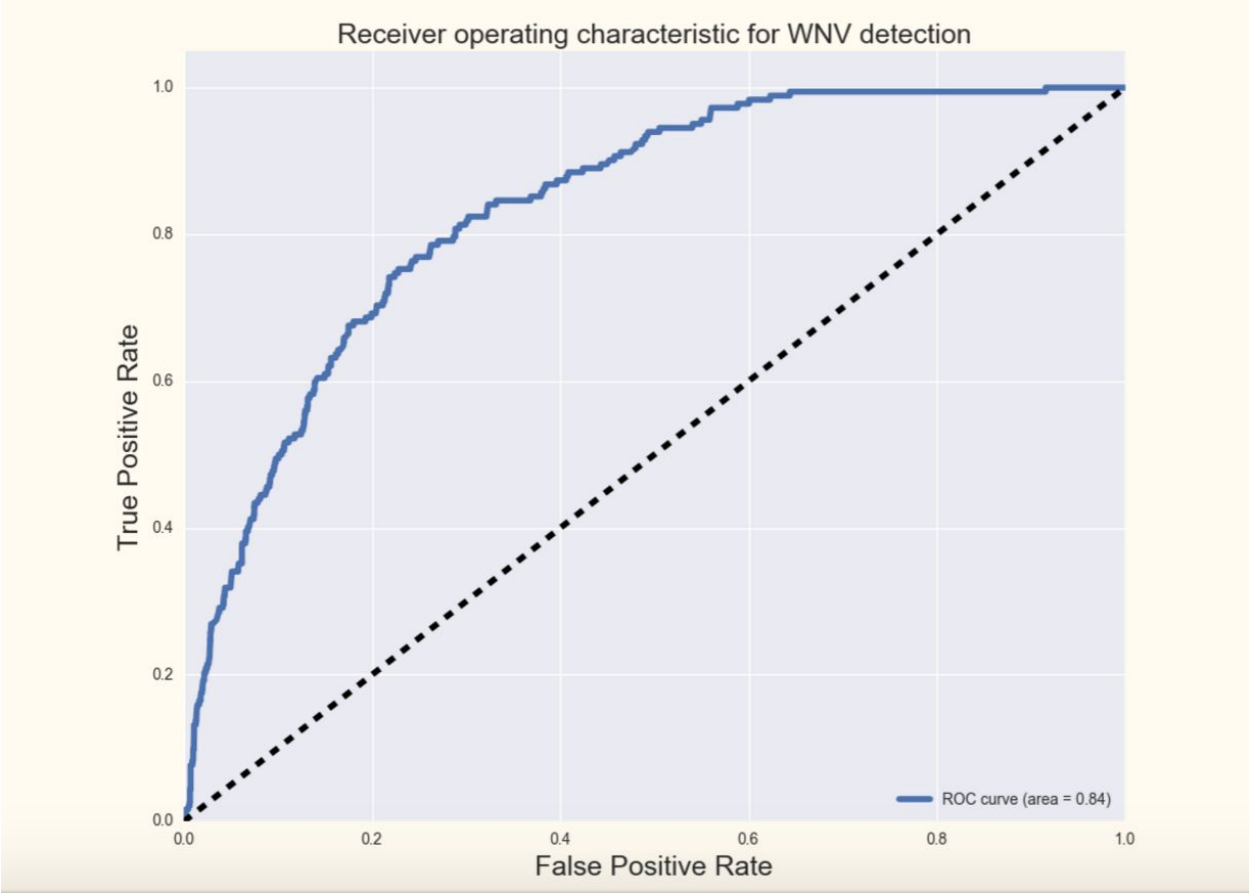
Model Used

To best target WNV mosquitoes, we built a classification algorithm that will help the city of Chicago conserve resources while mitigating the threat. The model we used is named XGBoost (‘eXtreme Gradient Boost’) which is a variety of a decision tree classifier. The variables for the model: precipitation, wind speed, species, length of day, trap location (long./lat.), temperature variance and maximum temperature. The target: If trap has a WNV mosquito - yes or no.

Results

The feature that ended up having the most importance was the length of day followed

by location. The ROC-AUC curve (see below) for this model has a score of 0.84.



This project was originally a Kaggle competition modeling problem. We submitted our solution to see where we would have placed in the competition, our ranking was about 580 (and second highest in our class). Considering that we had a fairly short amount of time and were competing against thousands of experienced data science teams, we were fairly happy with that outcome.

I can not say enough good things about the two people I worked with on this project. Ritika and Parisa were a pleasure to work with. This was a great team. I did the initial research and found the paper that clued us to use the length of day as an important input variable. I wrote the code for that input variable. Ritika and Parisa did rest of the modeling.

The icing on the cake for this project was the presentation we gave to the class. I had our team practice the presentation at least four times. My teammates were troopers and put up with me on that aspect. As a result, it went very smoothly and we won the prize for the best presentation. Gotta love it!! Thanks Ritika and Parisa!! Truly awesome teammates!