West Nile Virus Prediction

Purpose of the Project:

The purpose of this project is to analyze the occurrence of mosquito-borne disease called the West Nile Virus in the City of Chicago and provide recommendations on the best measures to preemptively control it.

DataSet:

The dataset available for this project contained trained data with 10K rows with information on date, location, number and type of mosquitoes and prevalence of west nile virus at the particular location. There was additional test data that contained close to 100k rows of data but was missing the target feature of virus presence. So, for our analysis we ignored this data.

The most important dataset was the weather data from two stations at Chicago's two airports.

Data Preparation and Feature Engineering:

Since there was weather data for each day from two weather stations, we made an assumption that the most accurate weather data for any location would be from the station that is closest to the location. Thus we calculated the distance between the location of the train data and the weather station and merged the weather data to the train data that came from the closest station.

Given that the life cycle of mosquitoes, temperature and precipitation can have a delayed impact on the presence of the virus, we created the time lag data on these variables at 1, 2, 7 and 14 days. The correlation of the virus presence was greater on these time lag data than the original data showing our hypothesis was valid. We further checked for multicollinearity using the VIF score and ignored columns that had a VIF score of more than 15. This resulted in 17 total relevant features for our model building. Furthermore the data had target feature heavily skewed, at a ratio of 19:1. We balanced the data using a resampling function.

Modeling:

We considered 6 different machine learning models and calculated their AUC score to determine which would fare out best. Out of the 6, we picked two models, Random Forest and