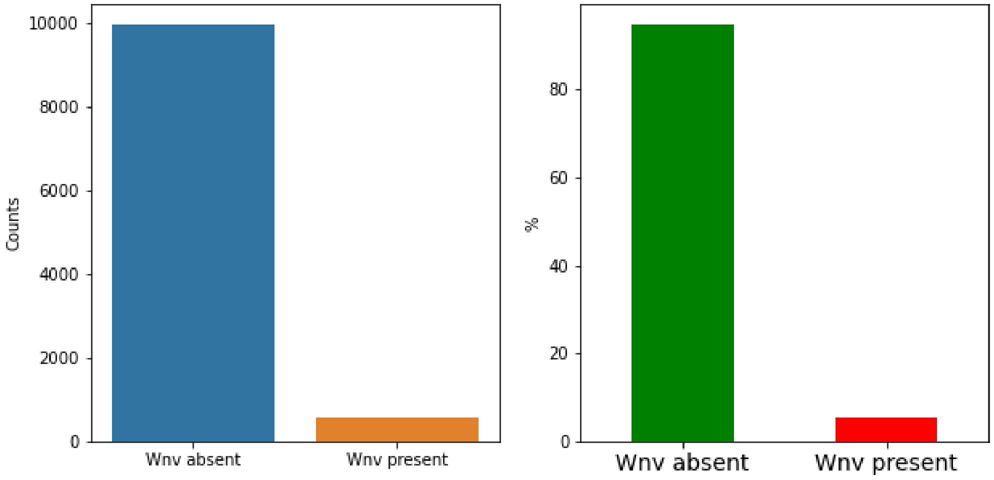
Capstone Report
Predicting the Presence of West Nile Virus

Introduction

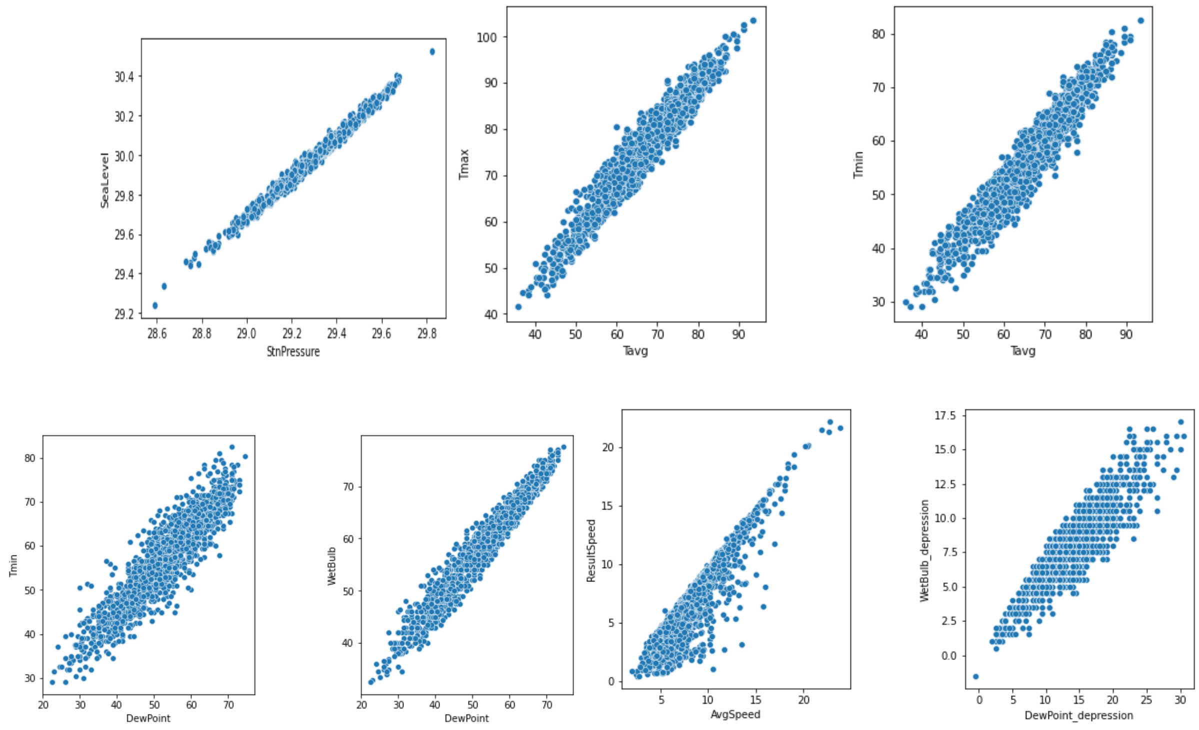
West Nile Virus can be transmitted to humans through infected mosquitos and causes symptoms ranging from a persistent fever, serious neurological illness as well as death in around 20% of those infected. Chicago reported the first human case of the virus in 2002. In order to prevent transmission, the City of Chicago and the Chicago Department of Public Health (CDPH) established a comprehensive surveillance and control program. To better allocate resources to prevent the transmission, there is a need to more accurately predict the presence of the virus in a given time, location and species.

Exploratory Data Analysis

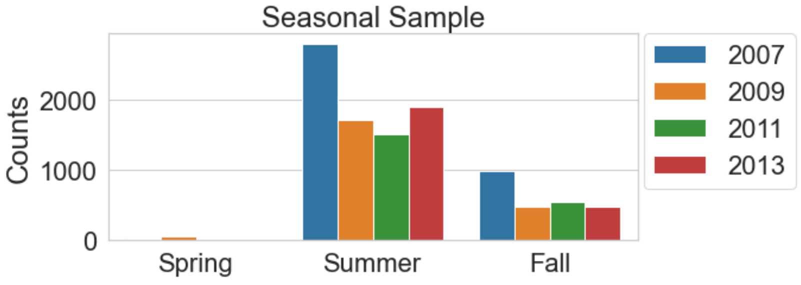
The data is available in Kaggle and has been provided by the Chicago Department of Public Health (CDPH). It consists of 2 sets of data, one with location attributes and the other pertaining to weather. Both the datasets were combined and used in this project to build the model. The GIS data contains 10506 rows and 12 columns and contains information such as date, location attributes, species type and a filed indicating the presence or absence of the virus. Data was collected in the city of Chicago every alternate year starting 2007 to 2013 from 136 locations. Locations are identified using the latitude and longitude, address, street and blocks. Mosquitoes were trapped in each of these locations during May to October. The virus was observed only in two species – Culex Restuans and Culex Pipiens. From the figure below, it is also observed that the virus was present in inly a fraction of the total mosquitoes caught indicating unbalanced data with a wide disparity in the samples collected.

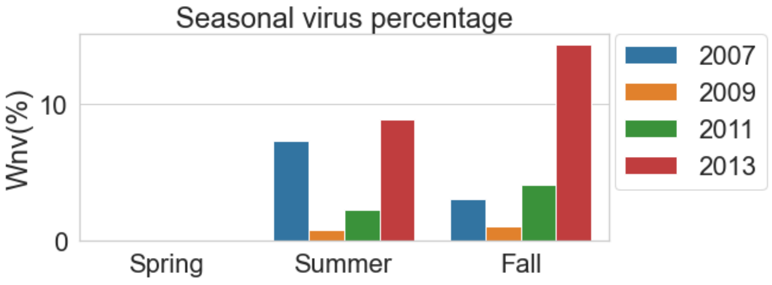


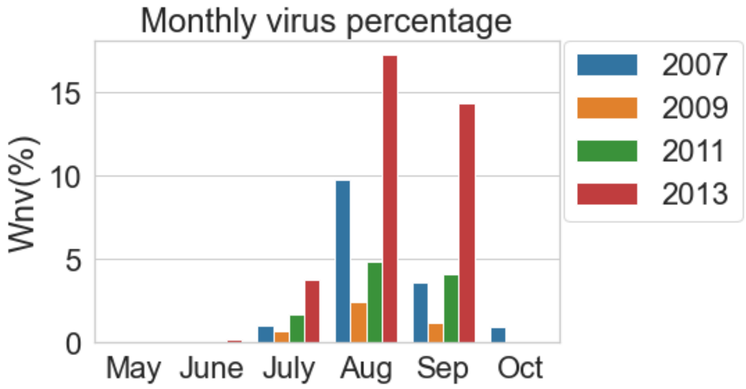
The weather dataset contains 2944 rows and 22 columns containing information recorded by NOAA from two weather stations. Attributes such as temperature, pressure, sunrise and sunset times, precipitation etc., were recorded. Correlations between many of these features were observed.



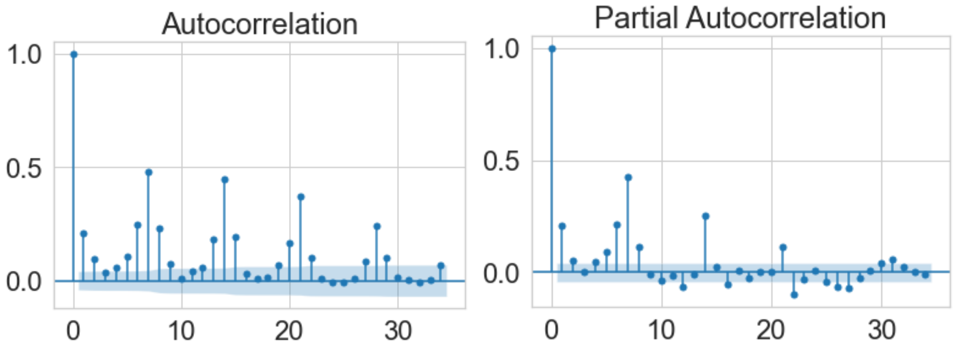
Feature Selection
The data from the two weather stations were averaged and null values were either filled or removed. Additional features were calculated from the existing features. Most of the correlated features were not retained to account for and multi- collinearity issues. The two main datasets were then merged to study the relation between weather on the observation of the virus in the mosquito samples. It was observed that although more samples were collected in 2007, the highest percentage of virus was detected in 2013. It was also observed that the winter season is missing in the data which could be due to the fact that mosquitoes do not survive in the cold. Also, the virus was not detected in the spring season. August was seen to be the month when the highest percentage of virus was detected.







Time series analysis was performed on the percentage of virus seen daily and it was observed that the correlation with 7, 14 and 21 day lag were significantly higher. As a result, additional features indicating these lags were added resulting in a total of 45 features.



Mosquito species which is a categorical column was coded using one hot encoding and those that did not contain the virus were dropped. All other date and object columns were dropped.

Preprocessing

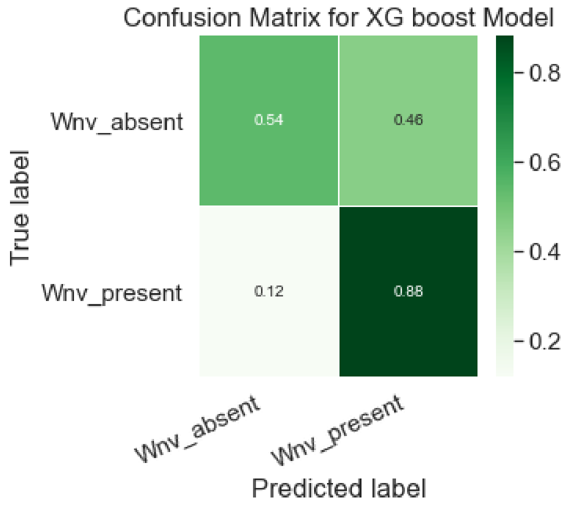
Information Value or IV was used in feature selection. Those features thar exhibited statistics in the range of 0.1 to 0.8 were selected for model building. Further, a technique called Variance Inflation Factor (VIF) was used to reduce multicollinearity between the features. A VIF of greater than 5 indicates extreme multicollinearity and is avoided. The number of features are brough down to 11 with the application of these 2 techniques. Inclusion of heat index and temperature related features is reasonable since mosquitoes are known to survive and breed at higher temperatures. The fourteen-day lag is also another useful feature since the presence of the mosquitoes for longer periods could lead to highest presence of the virus . Month is another feature that is included since significant variations by month are observed where August sees the highest presence whereas it is virtually nonexistent in October.

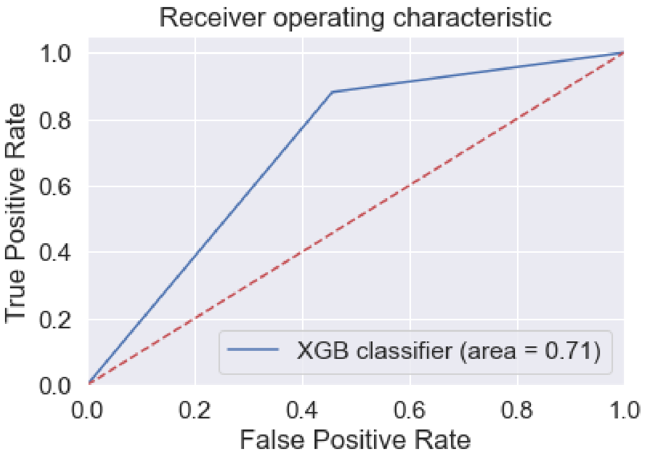
Modelling

The data is split into 70/30 for training and testing sets. The XGBoost (eXtreme Gradient Boosting) classifier algorithm which is a supervised learning technique was used to build the model on the training set. The grid search method was used where multiple model parameters were tuned with a five-fold cross validation to predict the presence or absence of the virus.

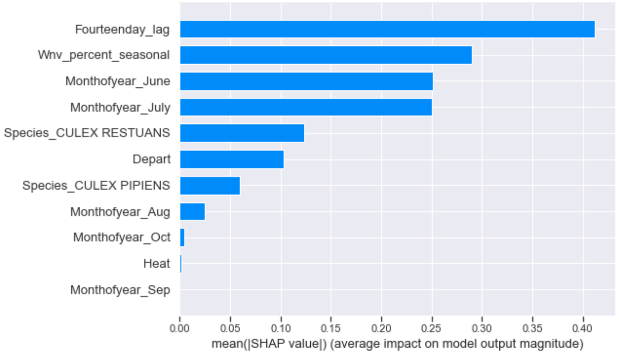
Results

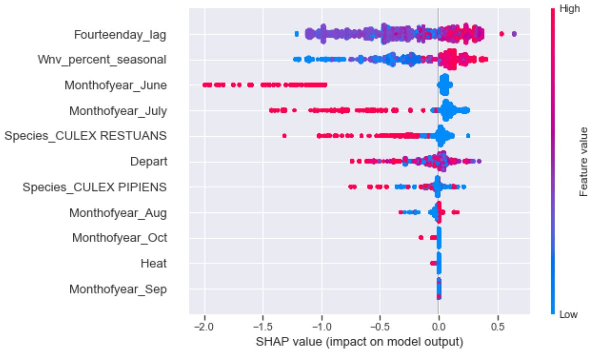
A lower score of 0.71 is obtained when the model is applied to the test set but performs better than a random classifier with a AUC of 0.5. In this study, a false negative, where the presence of virus is incorrectly classified would be considered more risky than a false positive. Therefore, the recall metric would need to be high to indicate a good model. The confusion matrix generated below shows a recall of 0.88 for the model indicating an efficient model for prediction of the presence of the virus.





The summary and shap plots below show the features and their contribution to the model's output. The fourteen day lag seems to be the most important feature affecting the classification decision with a positive correlation, followed by the seasonal observation.





Discussions & Conclusions

The risk associated with an incorrect classification of the presence of the virus would have greater impact than a false positive. For this reason, a higher recall value of the model is required. The recall value observed is 0.88 which makes the model and efficient classifier. The 14-day lag feature with a positive correlation seems to be an important indicator which would need to be considered by the City of Chicago and the CDPH when designing preventive measures and controls to tackle any outbreaks. Season and month have also been found to contribute to the presence of the virus, with summer season and especially the month of August having the highest percentages. This would be the months when more precautions should be taken to avoid infections. One limitation of this project is that the spraying data was not considered in model building. Proactively spraying the areas could have an affect on the presence of mosquitoes and the virus and should probably be considered in model building to obtain a more efficient model.