Predicting the Presence of West Nile Virus

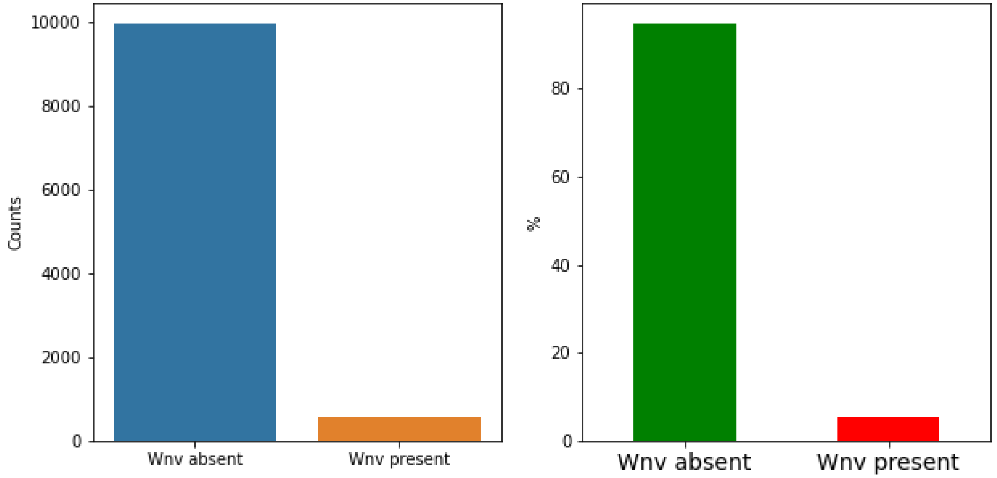
Capstone Project

Introduction

- West Nile Virus - transmitted to humans through infected mosquitos
- Symptoms - persistent fever, serious neurological illness as well as death
- 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

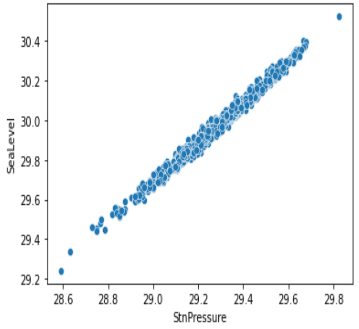
Exploratory Data Analysis

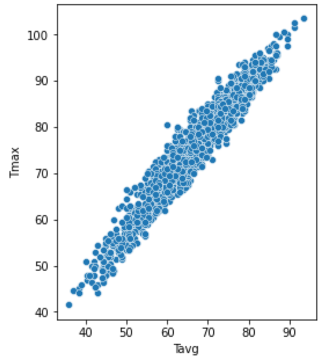
- 2 sets of data - location and weather
- Location Data
  - 10506 rows and 12 columns
  - date, location attributes, species type and a filed indicating the presence or absence of the virus
  - collected in the city of Chicago every alternate year starting 2007 to 2013 from 136 locations
  - Mosquitoes trapped during May to October.
  - virus was observed only in two species – Culex Restuans and Culex Pipiens

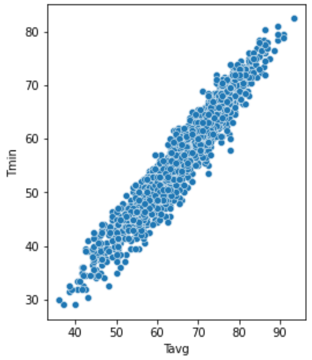


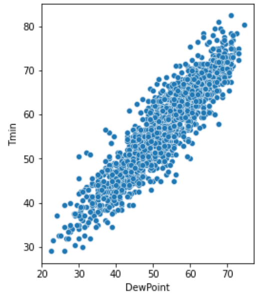
Exploratory Data Analysis

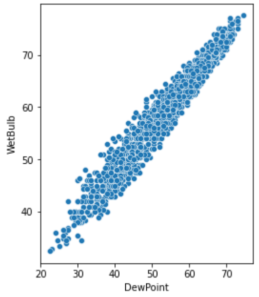
- Weather Data:
  - 2944 rows and 22 columns
  - temperature, pressure, sunrise and sunset times, precipitation etc
  - Correlations between many of these features were observed

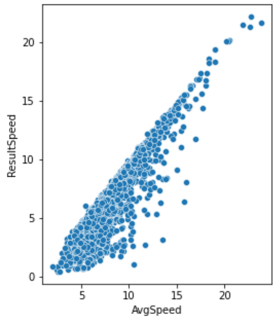


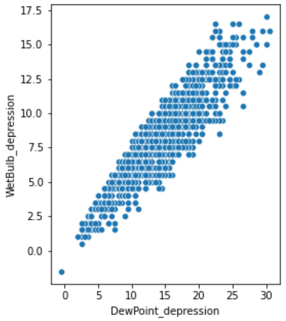






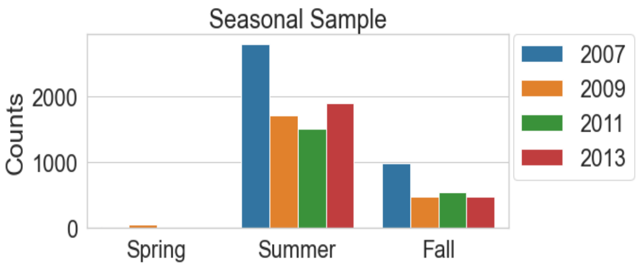


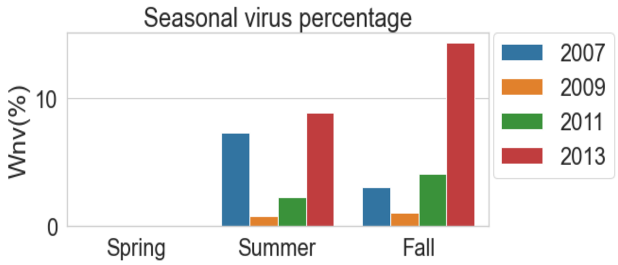


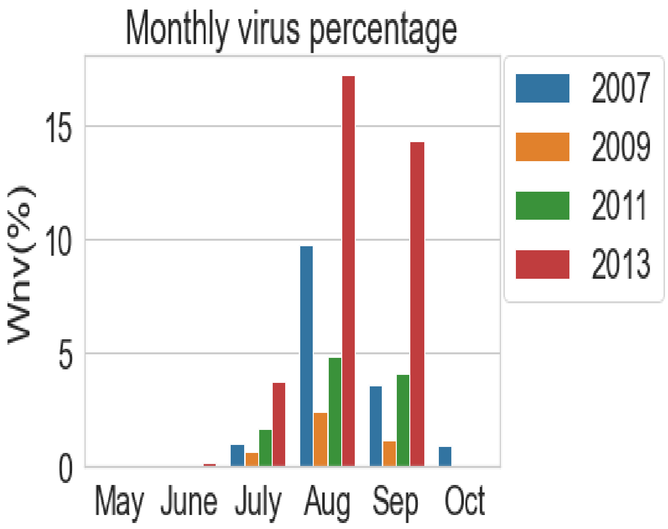


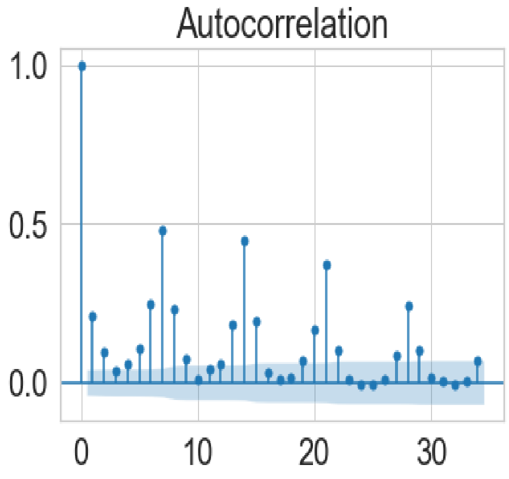
Exploratory Data Analysis

- Higher percentage of virus detected in 2013
- Fall saw higher percentages of virus with August being highest
- Correlation with 7, 14 and 21 day lag







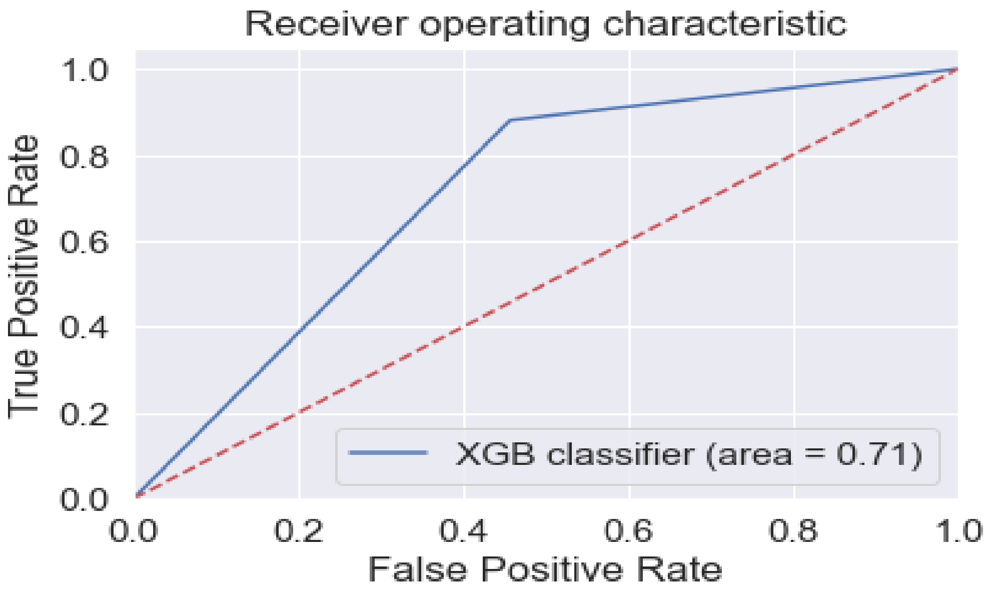


Feature Selection & Preprocessing

- Data from the two weather stations averaged
- Null values were either filled or removed
- Additional features calculated from existing features such as month, season, 7, 14 and 21 days lag
- Correlated features not retained
- One hot encoding used on mosquito species
- Object columns other than latitude and longitude dropped
- Information Value (IV) used in feature selection with range of 0.1 to 0.8
- Variance Inflation Factor (VIF) used to reduce multicollinearity
- 11 features retained for model building with the application of the 2 techniques

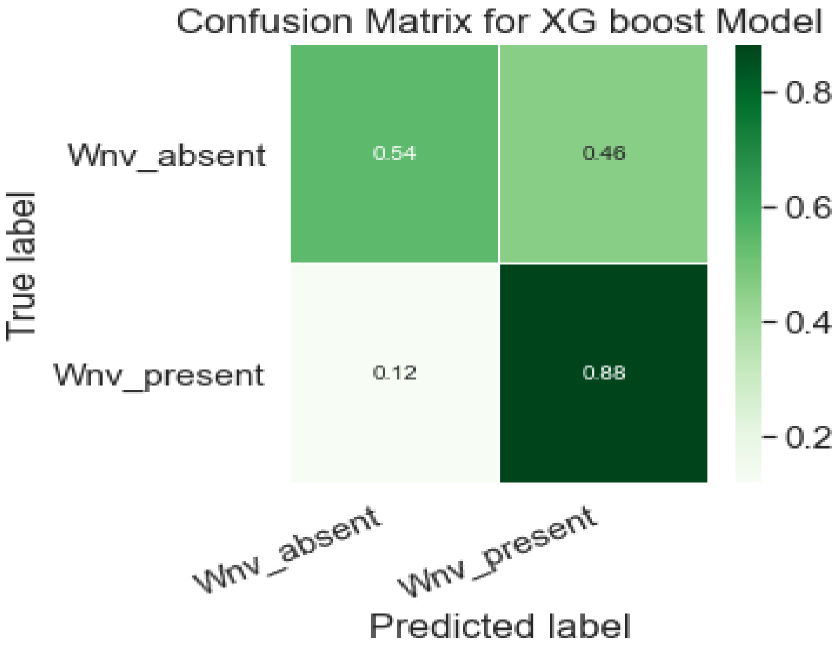
Modelling

- Data split into 70/30 training and testing sets
- XGBoost (eXtreme Gradient Boosting) used for modelling
- Model AUC score of 0.71

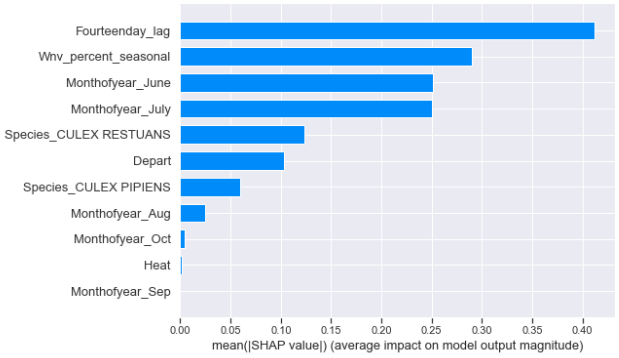


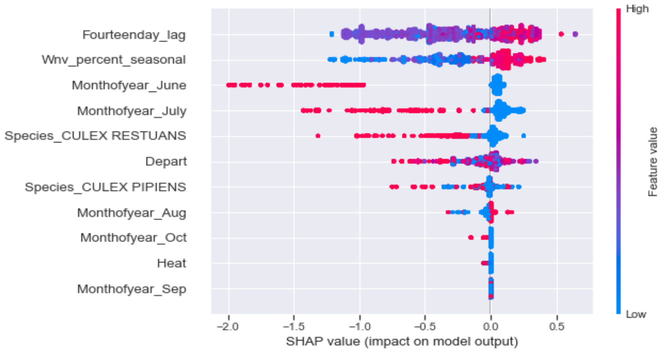
Results

- Recall metric of 0.88



- Fourteen day lag most important feature affecting the classification decision





Discussion & Conclusions

- Higher recall value of the model is required since risk associated with an incorrect classification of the presence of the virus would have greater impact than a false positive.
- Recall value of 0.88 observed indicating an efficient model
- One limitation - spraying data was not considered in model building
- Proactively spraying could affect presence of mosquitoes and the virus
- More efficient model could be probably developed by including spraying data

Recommendations

- The 14-day lag should be considered by the City of Chicago and the CDPH when designing preventive measures and controls to tackle any outbreaks
- The fall season, especially the month of August should be when more precautions should be taken to avoid infections