1 Problem Identification

Through a methodical analysis of Chicago's West Nile Virus data, build a model to predict the presence of West Nile Virus in given a set of conditions in Chicago.

2 The Data

The list of data and data sources being used in building the model for prediction:

- Source: Kaggle - https://www.kaggle.com/c/predict-west-nile-virus/data
  - Training/ Testng Dataset – containing mosquitos trap data across years in selected Chicago locations, contains insight on the presence of West Nile Virus (WNV) in various locations
  - Weather – containing various weather related metrics measured out from 2 different Chicago weather station
  - Spray data – containing locations where spraying has been done to eradicate mosquitos across selected years

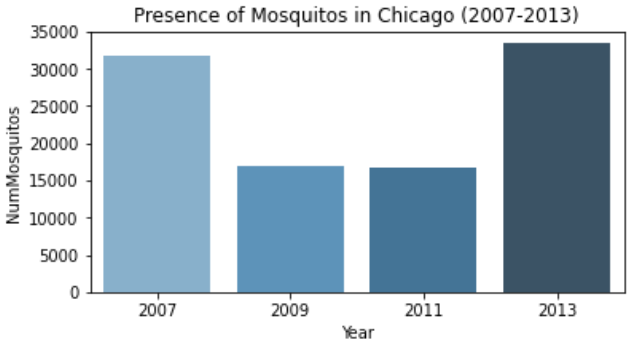
3 Data Wrangling

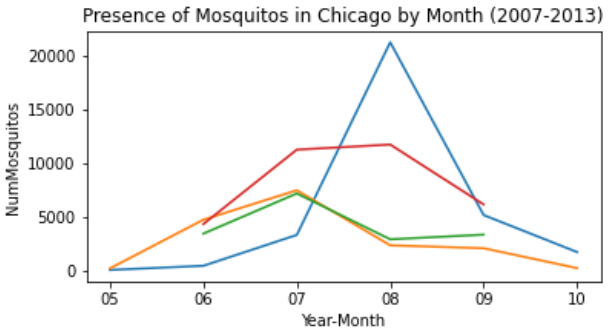
The thought process behind the data cleaning process are as follows:

- Step 1: Cleaning up obvious and generic data issues, including:
  - Parsing the date time data into proper format
  - Identifying and removing any duplicate records
  - Aggregating number of mosquitos as the records were split into a new record once the trap reaches 50 mosquito counts
- Step 2: Cleaning up other problems identified with the dataset, including:
  - Identifying missing values in the weather data, which is represented as 'M'
  - Replacing categorical indicators into numeric format to facilitate future model creation. E.g. "T" for trace rainfall and assigning a trace rainfall value as 0.001
  - Forward filling the missing data from station 1 into station 2 given most missing data is heavily concentrated in station 2
  - Drop columns which may not facilitate further analysis (Depth, Water1, Snowfall, CodeSum)
  - Adding date time indicators such as week, month to support exploratory data analysis process
  - Aggregating the 2 weather station's data by drawing an average between station 1 and station 2 at any given date
- Step 3: Merging various datasets together
  - Merging the training data and weather data together based on date

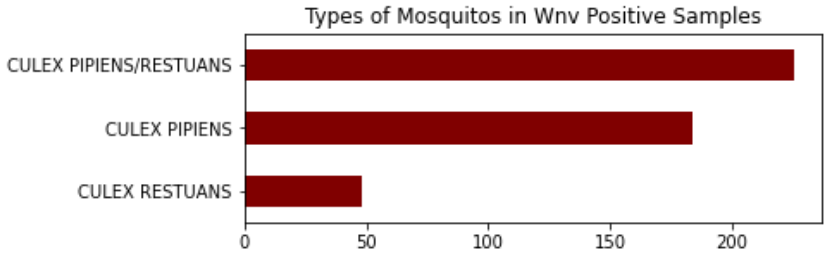
4 Exploratory Data Analysis

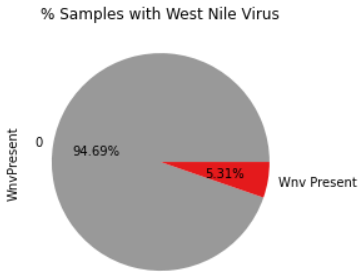
High level analysis were done to understand the patterns of mosquitos and West Nile Virus in Chicago during specific years where data is available.





Data are showing 2007 and 2013 being the most severe years with moquito problems, and moquito count peaks between July and August.





2 breeds of mosquitos causes West Nile Virus and around 5% of samples contain West Nile Virus. This indicates that data is very imbalances and will require re-sampling in pre-processing.

5 Feature Engineering and Pre-Processing

5.1 Feature Engineering

Further work is being done in preparing the data for machine learning. One of the issues identified with the dataset is, some useful information are categorical data. Specific work has been done to:

- Hot code mosquito breeds into numeric features
- Encode the mosquito Trap number into numeric features – as such, location data can be extracted or made possible to be interpreted

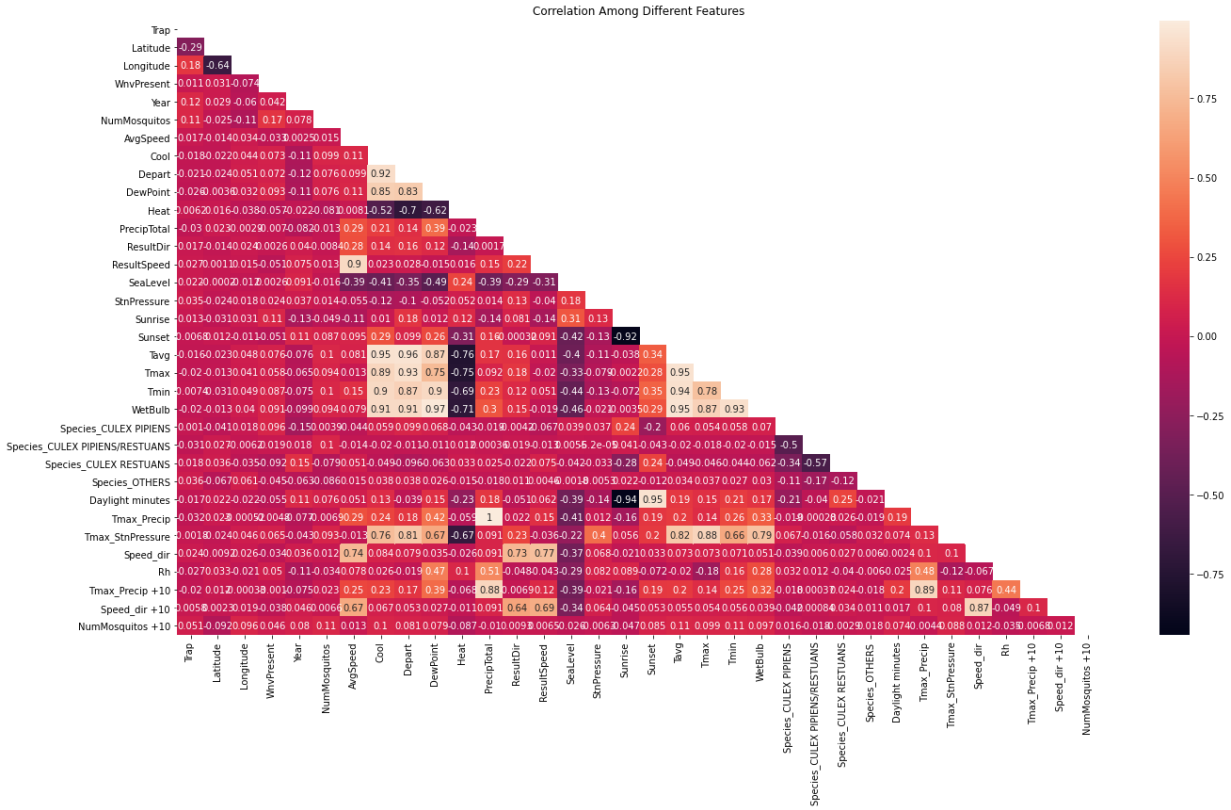
To increase the performance of the machine learning model, additional features has been engineered. These features are selected first based on desktop research, which suggests that growth of mosquitos are influenced by factors such as temperature, humidity, availability of water source etc. These are the final additional features built:

- Transforming Sunrise Sunset numeric numbers into day light duration in minutes
- Maximum Temperature and precipitation
- Maximum Temperature and pressure
- Windspeed and wind direction
- Relative Humidity

- 10 day time lagged features were also created for Maximum temperature & precipitation; windspeed/ direction and number of mosquitos.

10 day has been chosen as the duration of the time lag as the natural cycle of West Nile Virus causing mosquitos is typically between 7-10 days.

After features were being added to the dataset, a simple heat map has been created to understand the correlation between the features –

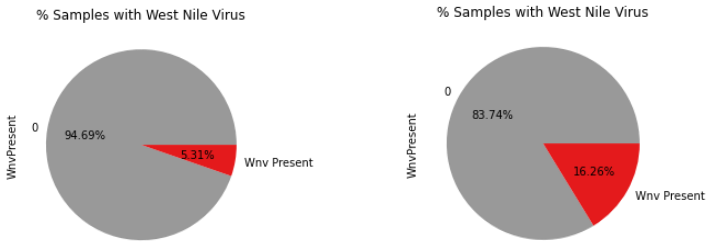


Results from the heat map is suggesting possible collinearity issues between some features, which needs to be addressed prior to building the train test split.

5.2 Under Sampling

The presence of Wnv samples in the original dataset is heavily imbalanced, hence resampling is required to improve the balance of the dataset. Specifically, two actions have been taken:

i. Removing rows of records where mosquitos identified does not cause Wnv
ii. Randomly selecting 30% of the records where Wnv is not present



These techniques has improved the balance of the data from 95% of negative class to 5% of positive class, to 84% negative class to 16% positive class.

5.3 Removing Multicollinearity

Variance inflation factor (VIF) has been used to identify and remove collinearity issues among the list of features. Below are the list of final features list to be used in modelling.

	VIFactor	features
0	2.537111	Trap
6	2.040188	Speed_dir +10
5	1.472204	Species_CULEX RESTUANS
4	1.356172	Species_CULEX PIPIENS
1	1.328338	NumMosquitos
7	1.304704	NumMosquitos +10
3	1.204422	PrecipTotal
2	1.108977	Heat

5.4 Train/ Test Split

A 75:25 train test split has been built, with the test data being scaled with the MinMaxScaler since the data is not normally distributed.

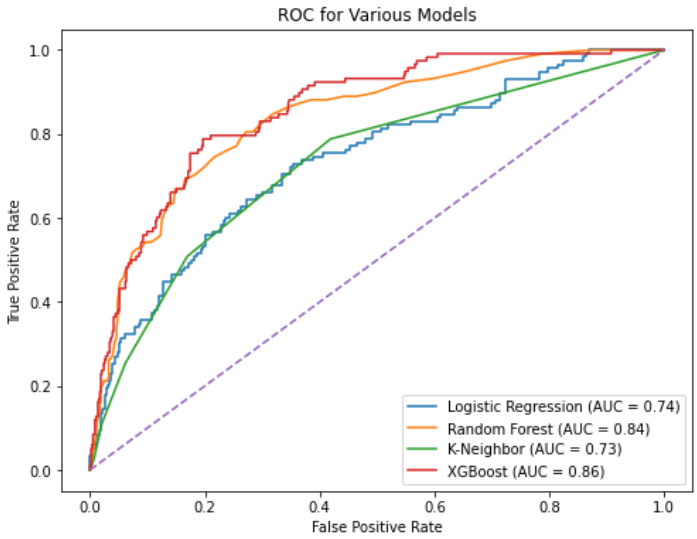
6 Modelling

Various models have been experimented in the process:

- Linear Regression
- Random Forest
- K-Nearest Neighbor
- XGBoost Classifier

These algorithms are chosen as the nature of the issue we are resolving is both a prediction and a classification problem. Several metrics has been used to illustrate each model's performance on Wnv and below figure summarizes the performance of each model:

	Metrics/ Model	Logistic Regression	Random Forest	K-Nearest Neighbor	XGBoost
0	Accuracy Score (Test data)	0.835	0.846	0.824	0.858
1	ROC Score	0.519	0.634	0.596	0.664
2	F1 Score	0.079	0.407	0.326	0.468
3	Precision Score	0.625	0.578	0.455	0.629
4	Recall	0.042	0.314	0.254	0.373



While all models are showing reasonable accuracy scores above 0.8 for all models and, the ROC and F1 scores are showing a different picture. Laying all metrics side by side, XGBoost Classifier is the best performing model with the highest accuracy, AUC and F1 scores. As such, the model will be further fine turned and to identify what are the optimal parameters.

7 Model Optimization & Evaluation

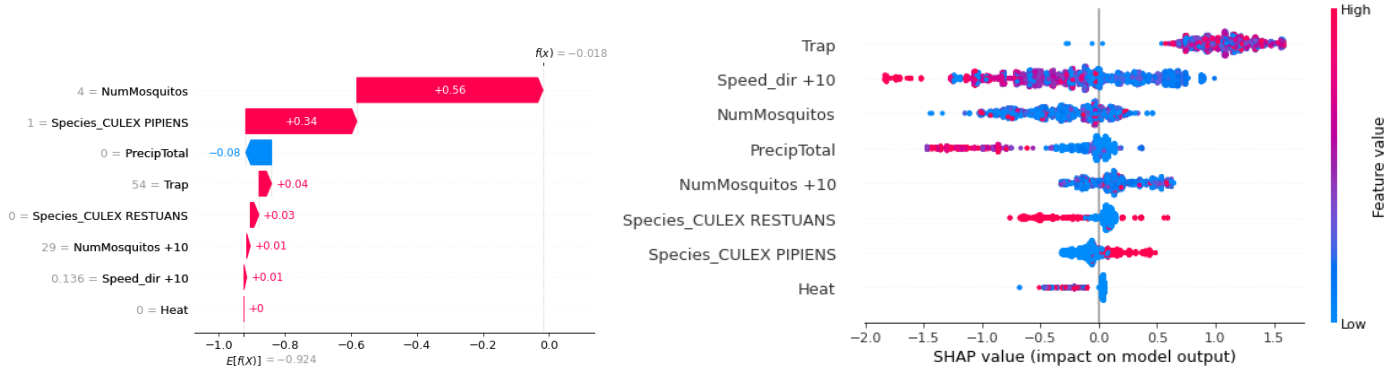
In order to identify the optimum parameters for the model, GridsearchCV has been used to find the best optimized combination of:

```
'learning_rate': (0.01, 1.0),
'n_estimators': [500, 800, 1000],
'min_child_weight': [1, 5, 10],
'max_depth': [5, 10, 15],
```

Results from the GridsearchCV is suggesting the best parameters for the model is - `{'learning_rate': 0.01, 'max_depth': 5, 'min_child_weight': 10, 'n_estimators': 500}`

To estimate the predictability of the machine learning model on new data, a 5 fold cross validation has been done. Result shows the accuracy of the model with cross validation is at 83.53%, meaning the model performs well even with data outside of the training set.

In order to further interpret and understand the model, SHAP analysis has been used to outline which features holds the most important key in helping with predicting Wnv.



From the chart, it can be observed that the species of the mosquitos (particularly if Culex Pipens exists in a mosquito tra sample), trap location and numbers of mosquitos (as of the prevalence of mosquitos in the area) are the core factors, and key predictors to Wnv.

8 Conclusion & Future Improvements

Among all the models used to predict the occurance of West Nile Virus, XGBoost Classifier, with parameters - `{'learning_rate': 0.01, 'max_depth': 5, 'min_child_weight': 10, 'n_estimators': 500}` yields the best predictive results. The final XGBoost model yields an AUC score of 0.86 with cross-validation accuracy score at 0.8353.

However, the model does have potential limitations, as it assumed:

- The environment where sample has been collected is reltively static over the period from 2007 to 2013
- The 2 weather station's data would largely be similar, and a simple forward fill and extracting the mean would not distort the weather data

For future areas of improvements, other data could be studied and understand if it would facilitate in refining the predictive model:

- Landscape/ environmental data, open construction sites, near rivers, old swimming poos etcs
- Test the model on other years in Chicago to provide insights on the models' predictive accuracy