West Nile Virus Detection

Data driven approach to prevention and eradication of the virus

Project for Chicago Municipality (CM) and Chicago Department of Public Health (CDPH)
By Eran Schenker

Executive Summary

Background

- West Nile virus (WNV) is most commonly spread through infected mosquitos.
- 20% of people who are infected, develop symptoms ranging from fever, to serious neurological illnesses and death.
- In 2002, first human cases of WNV were reported in Chicago. By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program.
- Every week mosquitos in traps across the city are tested for the virus. The results of these tests influence when and where the city will spray pesticides.
- Given weather, location, testing, and spraying data, Chicago Municipality (CM) and CDPH asked (through Kaggle) to predict when and where different species of mosquitos will test positive for West Nile virus.

Resource: https://www.kaggle.com/c/predict-west-nile-virus

Background

- West Nile virus (WNV) is most commonly spread through infected mosquitos.
- 20% of people who are infected, develop symptoms ranging from fever, to serious neurological illnesses and death.
- In 2002, first human cases of WNV were reported in Chicago. By 2004 the City of Chicago and the Chicago Department of Public Health (CDPH) had established a comprehensive surveillance and control program.
- Every week mosquitos in traps across the city are tested for the virus. The results of these tests influence when and where the city will spray pesticides.
- Given weather, location, testing, and spraying data, Chicago Municipality (CM) and CDPH asked (through Kaggle) to predict when and where different species of mosquitos will test positive for West Nile virus.

The goal

"To create a more accurate method of predicting outbreaks of West Nile virus in mosquitos to help CM and CPHD to more efficiently and effectively allocate resources towards preventing transmission of this potentially deadly virus."

Resource: https://www.kaggle.com/c/predict-west-nile-virus

Are the datasets informative for prediction?

“Train” (main) Dataset

- **Features:**
  Train dataset has informative features (e.g. number of mosquitos, species)
- **Assumption:**
  For this project's purposes let's assume for now that the 'NumMosquitos' feature IS NOT 'target-linked'* and is an integral part of the features provided by Chicago Municipality to predict WNV occurrences.
- **Prediction:**
  - Initial prediction efforts with this raw dataset shows that accuracy is deceivingly high ~95%. That’s because rate of WNV is very low (~5%).
  - This means that without feature engineering and data enrichment from other datasets we can get at most, a model that is deceivingly sensitive (sensitivity=75%) and entirely not precise (Precision=4%).

* In this context of this dataset ’Target-linked’ means that when deploying our model to predict WNV, we wouldn't have the number of mosquitos available to us because it is determined along side the detection of the virus – hence “target link”

Are the datasets informative for prediction?

“Spray” (Supplementary) Dataset

- **Instances:**
  The Spray dataset’s low number of relevant instances (observations) and lack of significant relationship to WNV, deems it un-informative to our purpose to enrich the main “Train” dataset, so we can ignore it

“Weather” (Supplementary) Dataset

- **Features:**
  After conducting rigorous feature engineering, the newly engineered weather features became highly informative which provides strong prediction power for WNV cases

Building a Statistical Model for optimal prediction of WNV

- **Assumption** –
  - Chicago Municipality (CM) and CDPH Would like to detect the highest number of WNV cases possible, rather than avoid a false alarm (falsely detecting WNV where there isn't).
  - This is because of the high cost related to undetected WNV cases (due to public health implications) as compared to the lower cost related to false alarm (the cost which involves over-spraying).

Building a Statistical Model for optimal prediction of WNV

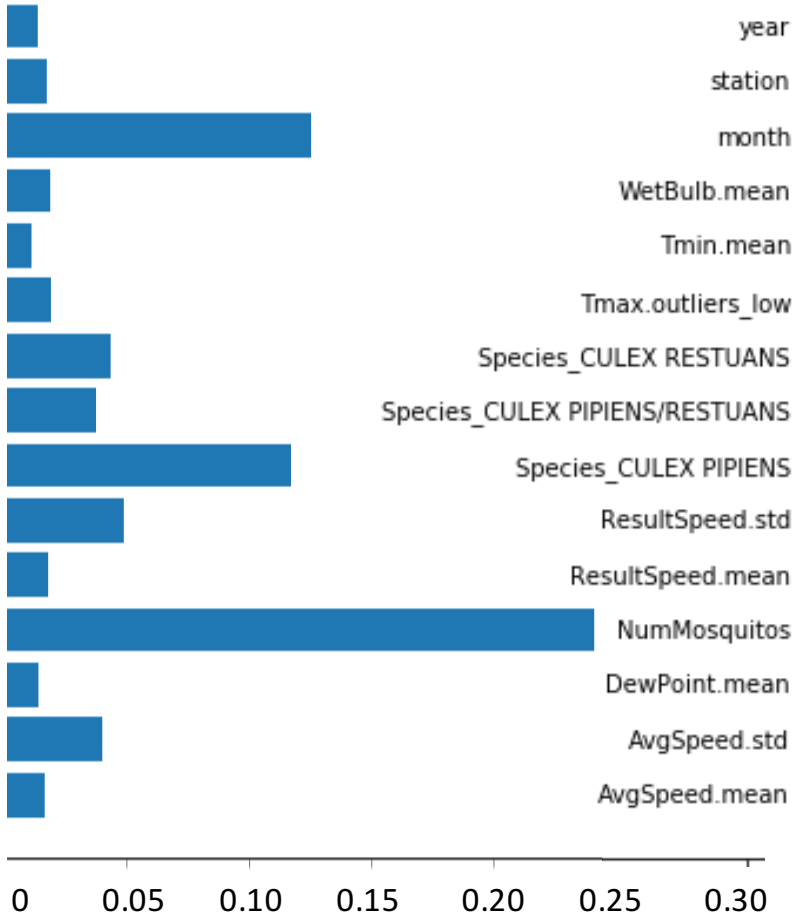
- **Assumption** –
  - Chicago Municipality (CM) and CDPH Would like to detect the highest number of WNV cases possible, rather than avoid a false alarm (falsely detecting WNV where there isn't).
  - This is because of the high cost related to undetected WNV cases (due to public health implications) as compared to the lower cost related to false alarm (the cost which involves over-spraying).

- **Method** –
  - We will run a classification model (Random Forest Classifier) to predict WNV cases
  - Then we'll optimize the model depended on the prediction score we get (AUCROC score)
  - Finally we'll set our recommendations based on the the most sensitive test we are able to provide, given a cost-benefit analysis (of undetected WNV vs over-Spraying)

Summary of the Proposed Model and Final Recommendations

- **What the model is revealing to us about WNV factors?**

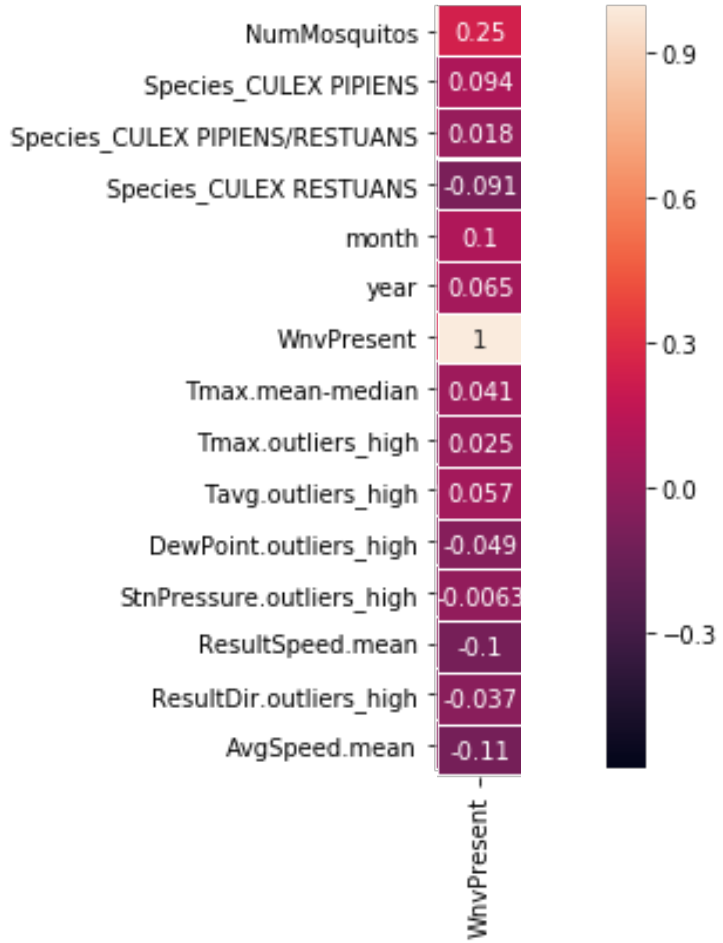
- **8 out of 13** of the most important factors are **engineered weather features**
- Not surprisingly, the most important factor is **Number of Mosquitos**
- Whether the specie is **CULEX_PIPIENS** or not is important, as well as which **month** it is (e.g. peak of summer)
- Weather factors include **wind speed** (most important probably due to strong winds effecting mosquito activity)
- Next in importance are **humidity** factors (i.e. wetbulb & dewpoint)
- **A significantly cold day** during otherwise stable 2 weeks is important (probably because it interrupts mosquito/virus activity)
- **Some years** have more cases of WNV than other years (probably due to specifically cold/hot years)



Summary of the Proposed Model and Final Recommendations

- **What the model is revealing to us about WNV factors?**

- A heatmap of correlations can give a better sense of how these factors are linked to WNV.
- All factors involving wind have negative correlation with WNV response i.e.
  more WINDY -> Less WNV.
- Although a small effect, when we look at Dewpoint, we see that the
  more HUMID -> Less WNV
  (which is aligned with known facts about WNV which prefers dry conditions)
- "Culex_Pipiens" specie is the most indicative of WNV response i.e.
  more Culex_Pipiens -> more WNV,
  and "Culex Restuans" is indicative of NO-WNV response i.e.
  more "Culex Restuans" -> less WNV



Summary of the Proposed Model and Final Recommendations

- **What the model is revealing to us about WNV factors?**

Summary of factors relationship to WNV:

(+++) strong positive relationship

(++) moderate positive relationship

(-) mild negative relationship

Factor	Related to WNV
Number of Mosquitos	+++
CULEX_PIPIENS (specie)	++
Month	++
High Temperatures	+
CULEX_RESTUANS (specie)	-
Wind	-
Humidity	-

Summary of the Proposed Model and Final Recommendations

- **Final actionable recommendations for Chicago Municipality and Department of Public Health**

- The sensitivity of the model we are offering ranges from detecting 60% to detecting 100% of WNV cases
- Depended on the cost of false alarm (e.g. spraying areas that are not infested), we could offer the right package for Chicago Municipality and CDPH purposes.
- We offer 2 possible packages:
    i. If it's affordable to "over-spray" 4 times the number of necessary spraying (i.e. only 1 out of 5 spraying efforts are justifiable), we predict eradication of 90% of WNV of occurrences in Chicago (lowering its rate from 5% to 0.5%).
    ii. Alternatively If it's affordable to "over-spray" only 2.5 times (i.e. 2 out of 5 spraying efforts are justifiable), then we predict eradication of 60% of WNV occurrences (lowering its rate from 5% to 2%).
- We recommend to take into consideration the inhabitants population density of the areas in order to asses the importance of eradicating the mosquitos. This could provide important clues about the right choice of over-spraying and WNV detection rate discussed previously.

Summary of the Proposed Model and Final Recommendations

- **Final actionable recommendations for Chicago Municipality and Department of Public Health**

- We suggest to take preemptive measures as well as spraying infested locations. Since June has the highest occurrences of WNV, whilst not being the hottest, driest and least windy, we could assume that the high occurrences are due to reactive measures to eradicate mosquitos been taken too late in the season (only after locations become infested).
- We propose to invest efforts in prevention, starting from increasing awareness of the general population (e.g. to not leave exposed untreated water such as ponds and pools in their yard)
- Additionally, since one specie (Culex Pipiens) stands out as a carrier with higher rates of WNV, We suggest to increase efforts in preventing the proliferation of this specie in particular. This could be done with more specific and environmentally friendly measures (e.g. "targeted biological mosquito control") which would make the process of lowering occurrences more efficient and less costly

Supplementary Slide - Project Discussion and Final Notes

- AUC under the ROC curve is ~89% so the model is robust enough to allow flexible choice of thresholds (choosing the most fitting recall vs precision rates)
- The model is based on the assumption that the number of mosquitos is a valid info that could be collected and used for prediction. The model's performance would change if we exclude the 'Number of Mosquitos' feature. At that scenario we will use all other features to predict the number of mosquitos, and then use that prediction as a new feature in a new classification model, to make a new prediction for the WNV occurrences. We can expect the performance of this model to be lower.
- The spray data was implemented with the hypothesis that recently sprayed sites would effect the occurrences of WNV. after EDA, occurrences seem to not vary (on average) and since the number of observations taken from recently sprayed sites were low (79), these observations were not excluded from the dataset with the rational that they would not effect the model's performance.
- The data could be enriched further with publically available data, e.g. Implementing data on locations of exposed water reservoirs, lakes and other bodies of fresh water which might be correlated to increased mosquito population.