West Nile Virus Prediction

Audrey, Akhila, Shaun

Table of Contents

The West Nile Virus

- Leading cause of mosquito-borne disease
- Potentially fatal
- No Vaccine

Problem Statement

Problem Statement

Project Aim

- Predicting *West Nile Virus* (WNV) in Chicago

- Utilising weather features

Purpose

- Assist CDC and CPHD in combating WNV

Data Cleaning

Train
Test

1) Incorrect dtype

2) Drop (Redundant)

3) Drop (Duplicates)

4) Drop (Not in Test)

Spray

1) Incorrect dtype

2) Drop (Redundant)

3) Drop (Duplicates)

Weather

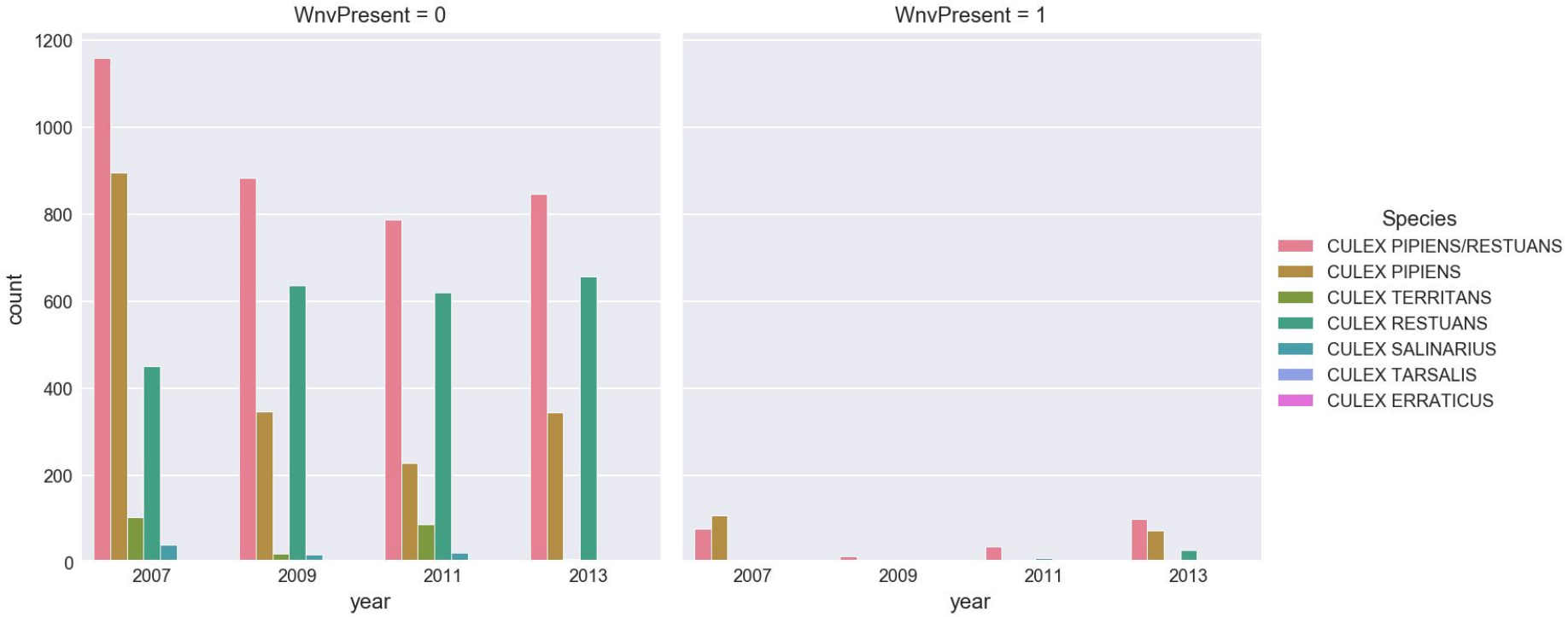
1) Incorrect dtype

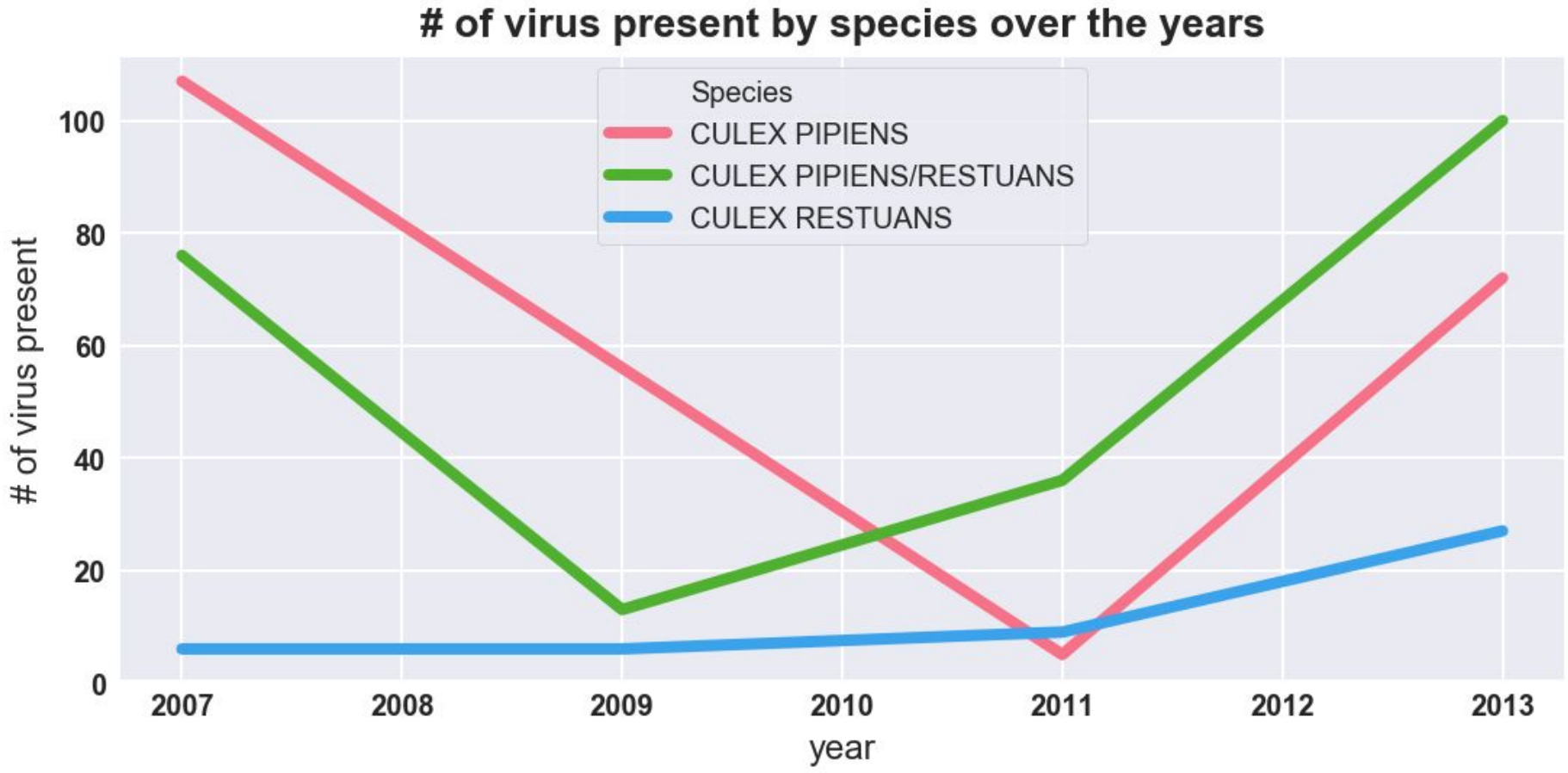
2) Drop (Station 2)

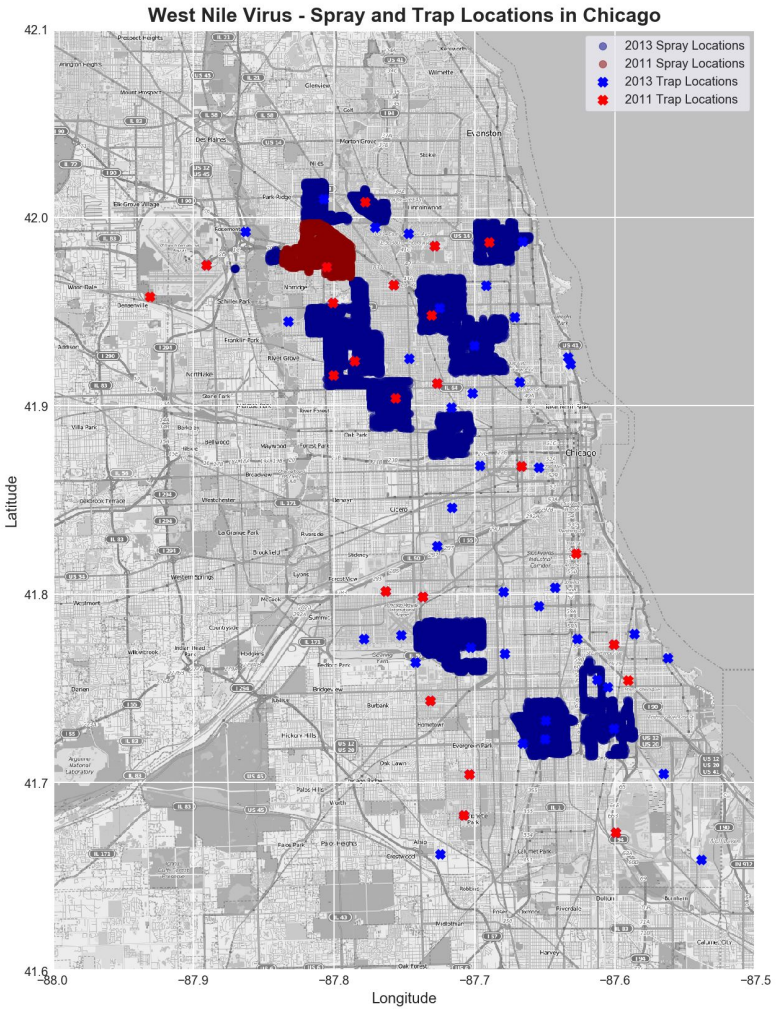
3) Drop (Missing Info)

4) Impute (“M”, “T”)

EDA


WnvPresent = 0
WnvPresent = 1
count
year
year
2007
2009
2011
2013
0
200
400
600
800
1000
1200
Species
CULEX PIPIENS/RESTUANS
CULEX PIPIENS
CULEX TERRITANS
CULEX RESTUANS
CULEX SALINARIUS
CULEX TARSALIS
CULEX ERRATICUS


# of virus present by species over the years
Species
CULEX PIPIENS
CULEX PIPIENS/RESTUANS
CULEX RESTUANS
# of virus present
0
20
40
60
80
100
2007
2008
2009
2010
2011
2012
2013
year


West Nile Virus - Spray and Trap Locations in Chicago
2013 Spray Locations
2011 Spray Locations
2013 Trap Locations
2011 Trap Locations
Latitude
42.1
42.0
41.9
41.8
41.7
41.6
Longitude
−88.0
−87.9
−87.8
−87.7
−87.6
−87.5
Evanston
Chicago

Preprocessing

Preprocessing

Features Engineered

1. Dummy variables for species created in both train and test set
2. Parsed dates into year, month,week of year, month of year
3. Coded hot and wet conditions based on dew point and average temperature
4. Merged weather dataset to train set for modeling on Date

Dropped

1. Spray set is omitted for now due to the lack of information over the years.
2. Temperature related & Dewpoint features

Modelling

Models Tried

1. **Logistic Regression**
2. **Gradient Boost Classifier**
3. **Gradient Boost Classifier with GridSearchcv**
4. **Random Forest Classifier**
5. **XGBoost Classifier**
6. **Decision Trees**

Modeling Results

Model	Hyper-parameters	Train Set AUC Score	Test Set AUC Score	Kaggle Score
Logistic Regression	-	**0.74**	**0.78**	**0.50**
Gradient Boost Classifier	-	**0.90**	**0.87**	**0.56**
	'learning_rate': 0.08, 'max_depth': 2, 'n_estimators': 100	**0.87**	**0.87**	**0.49**
Random Forest Classifier	-	**0.99**	**0.78**	**0.52**
XGBoost Classifier	n_estimators = 500	**0.96**	**0.87**	**0.53**
Decision Trees	-	**0.99**	**0.58**	**0.52**

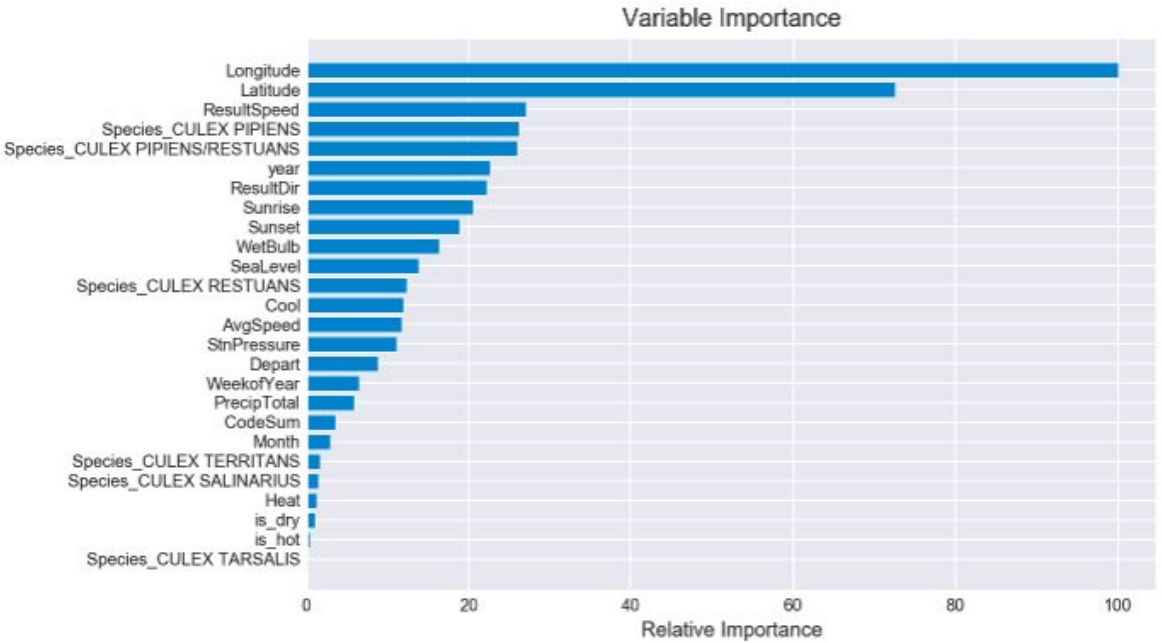
Model Evaluation

Baseline score vs. Chosen model score

Model	Train Set	Test Set
Logistic Regression	**0.74**	**0.78**
Gradient Boost Classifier	**0.90**	**0.87**

Heavily imbalanced dataset

Important Features



Insights & Conclusions

1) **Least variance between test and training scores >> not overfit**
2) **Feature importance suggests the prevalence of WN virus in certain locations for specific species**
3) **Time of the year with longer days influences virus presence**
4) **Features highlighted to be used to determine where and when to focus for spraying efforts**

Limitations & Recommendations

Limitations & Recommendations

LIMITATION	Model scored significantly lower on out of sample Kaggle set *>> probably due method of creating internal validation set vs. Kaggle's approach*
RECOMMENDATIONS	Elaborate feature engineering with time series functions on weather dataset Include dummified variable for top traps *NumMosquito* feature even with its high correlation to virus presence had to ignored Availability of spray information for more years would help the model score better

Cost-Benefit Analysis

Cost-Benefit Analysis

	Spraying	Not Spraying
Costs	● Chemical spray (Zenivex) ● Labour work ○ Approx. $0.75 per acre of spray	● Medical cost ○ $33,143 per inpatient ○ $6,317 per outpatient ○ $18,097 per patient spent time in a nursing home ● Productivity loss ○ $58,935 per personal income

Source:
http://www.gfmosquito.com/wp-content/uploads/2013/06/2013-North-Dakota-Bid-Tabulation.pdf
https://www.statista.com/statistics/205235/per-capita-personal-income-in-illinois/
https://www.cdc.gov/westnile/resources/pdfs/data/WNV-Disease-Cases-PVDs-by-State-2018-P.pdf

Cost-Benefit Analysis

	COST	BENEFIT
Spraying	• Vector control cost **Total = ~ $144K per person**	**Human Life Saved**
Not Spraying	**Human Life Loss**	• Medical bills (~$46,530 per person) • Productivity loss ~$58,935 per person) **Total = ~$108K per person**

Total $ Cost > Benefits ?

Cost-Benefit Analysis

	Spraying	Not Spraying

Benefits	• **Human Life Saved**
	  ○ Improved quality of life
	  ○ Increased workplace productivity
	  ○ Savings in hospital bills
	  ○ Attract visitors → economic benefits
		• **Human Life Loss**
		  ○ Long term/Wider mental health issues
		  ○ Lower quality of workforce and reduce output productivity
		  ○ Negative impact on tourism and entertainment sectors

Human Benefits of spraying outweighs the Costs!

Cost-Benefit Analysis

Year	Acre coverage*	Projected Annual Costs @ $0.75 per Acre
2020	~16m	~$12m
2021	~14.4m	~$10.8m
2022	~13m	~$9.75m

*Source:
https://www.cdc.gov/westnile/statsmaps/cumMapsData.html
https://www.cdc.gov/westnile/vectorcontrol/aerial-spraying.html
(assumption: 10% decrease in levels of pesticide coverage annually)

Conclusion and Recommendations

Conclusion & Recommendations

Since humans are gregarious and following CDC's mission statement, we must think of a regional approach to WNV vector control..... Spray!

- Consider strategically spraying at locations with highest infections
- Conduct spraying during hotter months i.e. August and September
- Educate and promote the public to:
  - Use insect repellent
  - Wear long-sleeved shirts and pants
  - Take steps to control mosquitoes indoors and outdoors
    - I.e. Remove standing water where mosquitoes could lay eggs

Questions?