West Nile Virus

Predicting Occurrence of WNV

Marvin, Chee Ming, Wenyan, Nicky

Background

The WNV is the leading cause of mosquito-borne disease in the continental United States. In view of this outbreak, the Chicago Department of Public Health (CDPH) has set up a surveillance and control system to trap mosquitoes and test for the presence of WNV.

Problem Statement

To help CDPH predict when and where different species of mosquitoes will test positive for WNV, thus effectively allocate resources towards preventing transmission of this potentially deadly virus.

Our project also aims to determine the best strategy for controlling the spread of WNV, as well as analyzing the various trade-offs that need to be made in implementing our model.

Contents

01. Data Cleaning & EDA

- Impute missing values, drop features with low variance, correct wrong entries
- Create weather and lag features

02. Preprocessing & Modeling

- Feature transformation
- 8 classifiers used and compared

03. Model Evaluation

- ROC-AUC score
- Recall Score

04. Cost Benefit Analysis

- Analysis

Data Cleaning
EDA

01.

Data Cleaning

01.

Impute Missing Values

- Average temp
- Depart temp
- Wetbulb
- Heat
- Cool
- Depth
- Water
- Snowfall

02.

Correct wrong data formats

- Incorrect sunset hours

03.

Station 2 has no sunset/sunrise data

- Forward fill

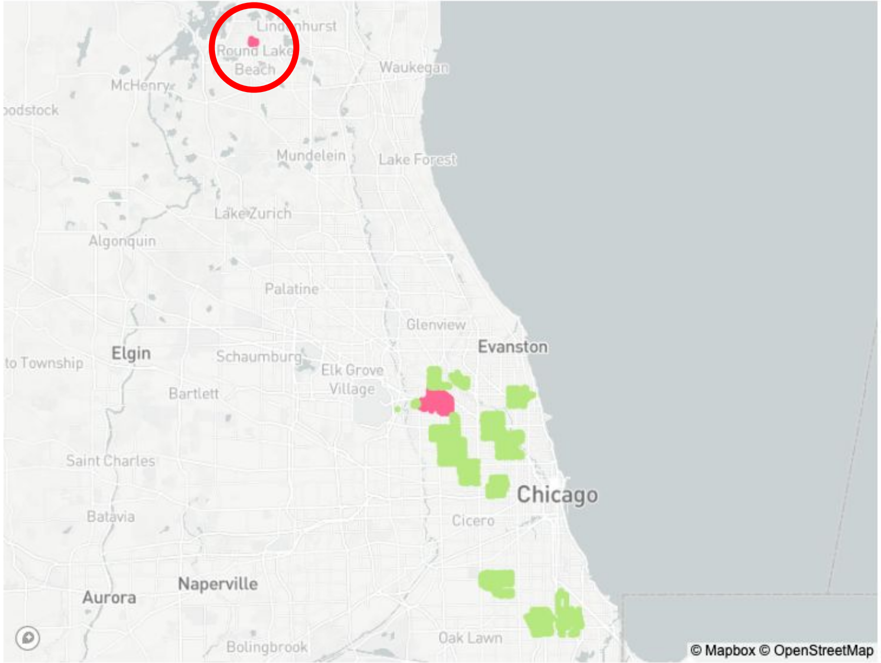
04.

Drop Duplicates:

- Spray column

Outliers in spray dataset

- Location not in city of Chicago. (latitude = ~42, longitude = -88 and occuring on 2011-08-29)
- Outliers removed from the dataset.



EDA & Feature Engineering

EDA/ Feature Engineering: Meteorological Features

- National Center for Biotechnology Information (NCBI) suggests that these weather factors are most pivotal in predicting mosquito growth and WNV transmission.
  - Average temperature
  - Temperature fluctuation (tmax-tmin)
  - Light intensity (operationalized by daylight hours, sunset - sunrise)
  - Relative humidity (formula explained)
  - Rainfall/precipitation rates

- Average temperature & precipitation rates are already included in the data and the rest will be engineered.

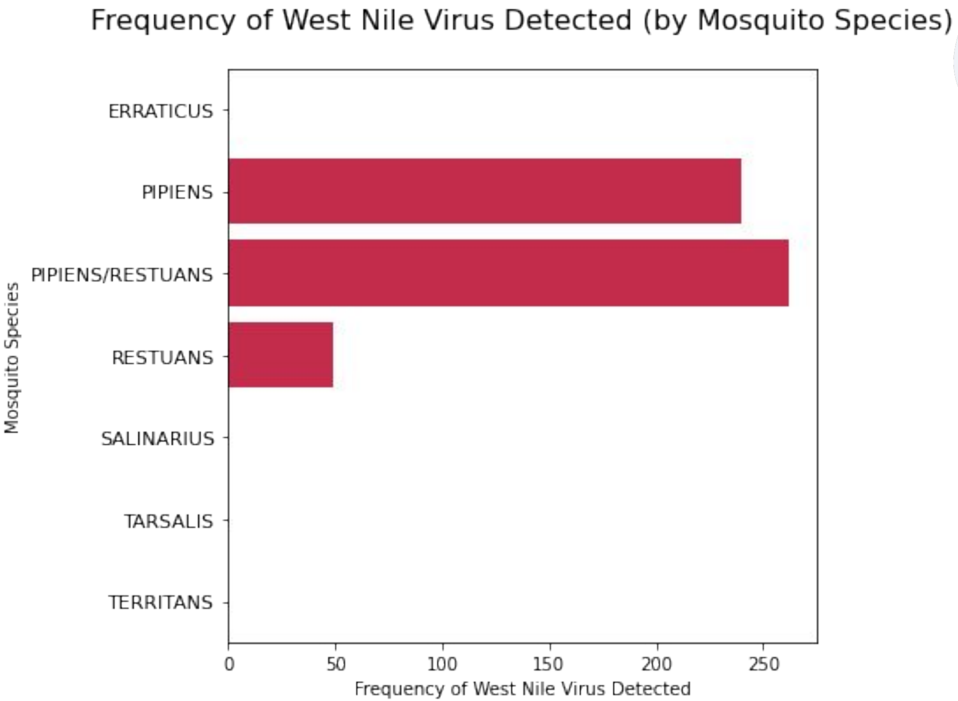
- Relative humidity: Computed using Metpy library with air temperature and dew point temperature as input parameters.

EDA/ Feature Engineering: Culex Species

- NCBI:
  - More important than meteorological features
  - Most virulent species: Pipiens, Tarsalis & Territans

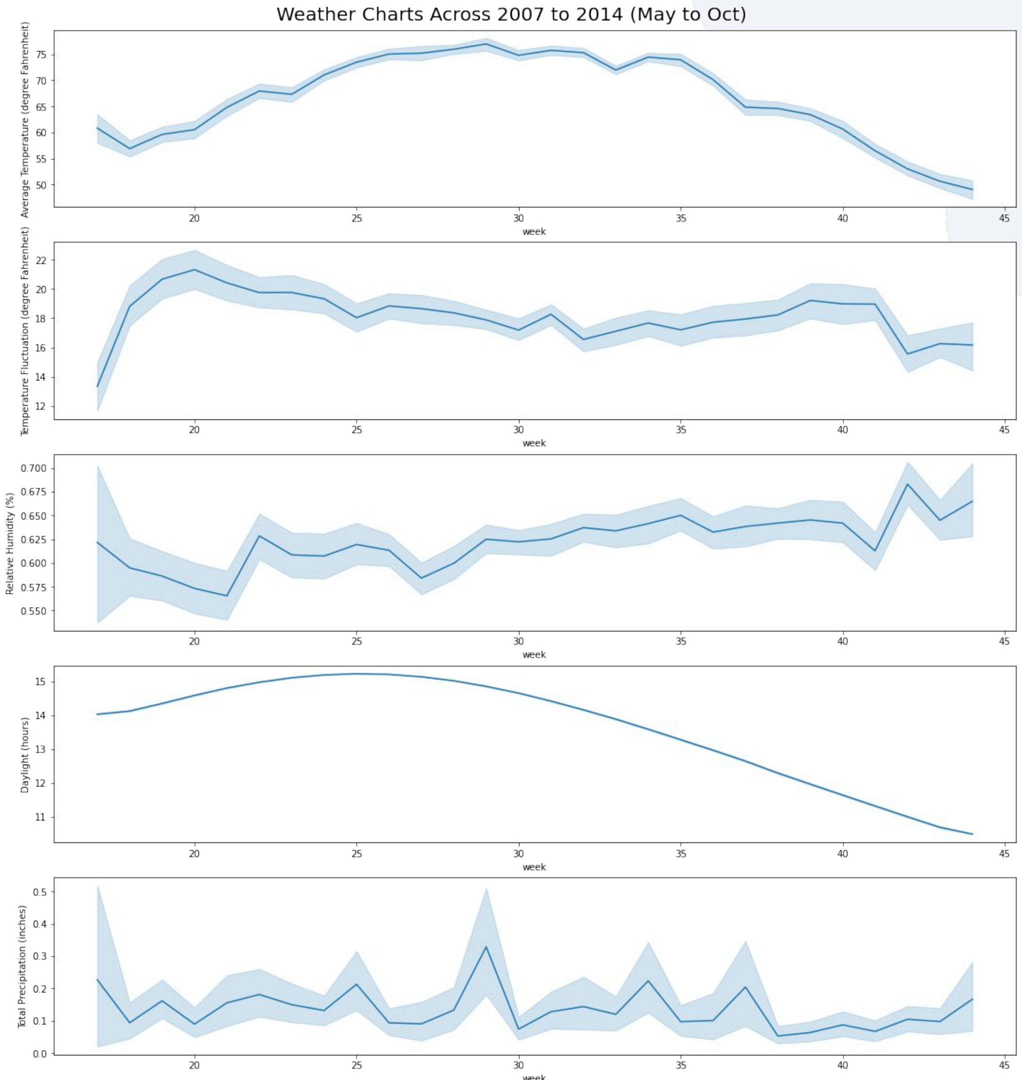
- EDA: 3 most WNV-positive species: Pipiens, Pipiens/Restuans & Restuans

- Because of the conflict of information, we decided to One-Hot encode mosquito species instead of ordinal encoding



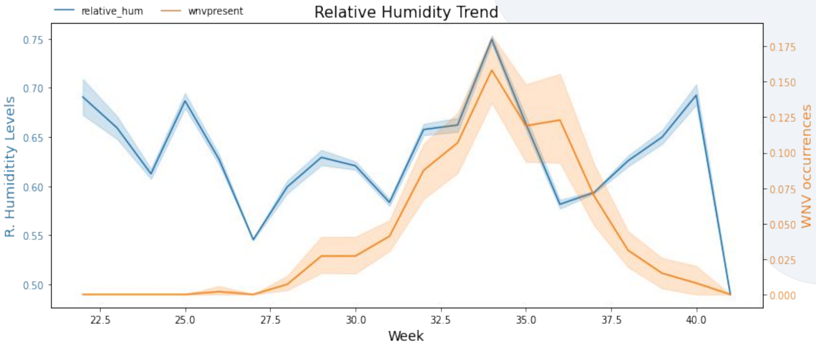
EDA/ Feature Engineering: Weather features

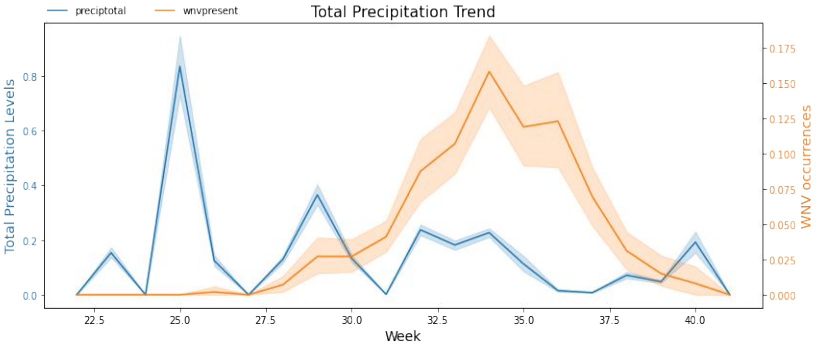
- Weather features against time (week of the year)
- Consistent readings across the years for average temp and daylight hours (narrow range in the CI)
- Relative hum and total precipitation have larger variance across the years.

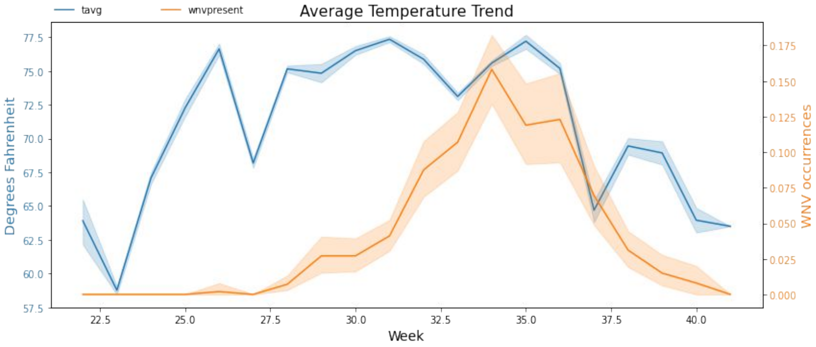

Weather Charts Across 2007 to 2014 (May to Oct)
Average Temperature (degree Fahrenheit)
50
55
60
65
70
75
Temperature Fluctuation (degree Fahrenheit)
12
14
16
18
20
22
Relative Humidity (%)
0.550
0.575
0.600
0.625
0.650
0.675
0.700
Daylight (hours)
11
12
13
14
15
Total Precipitation (inches)
0.0
0.1
0.2
0.3
0.4
0.5
20
25
30
35
40
45
week

EDA/ Feature Engineering: Lead-lag features

- Lead indicator - features that occur first and give early indications of performance
- Analysis from this section suggest that precipitation & temperature are leading indicators, while WNV occurrences are extremely responsive to changes in relative humidity from week 31-37.
- Engineered new features are “time-lagged” i.e. for observation record on week N, include also week N-2 data.

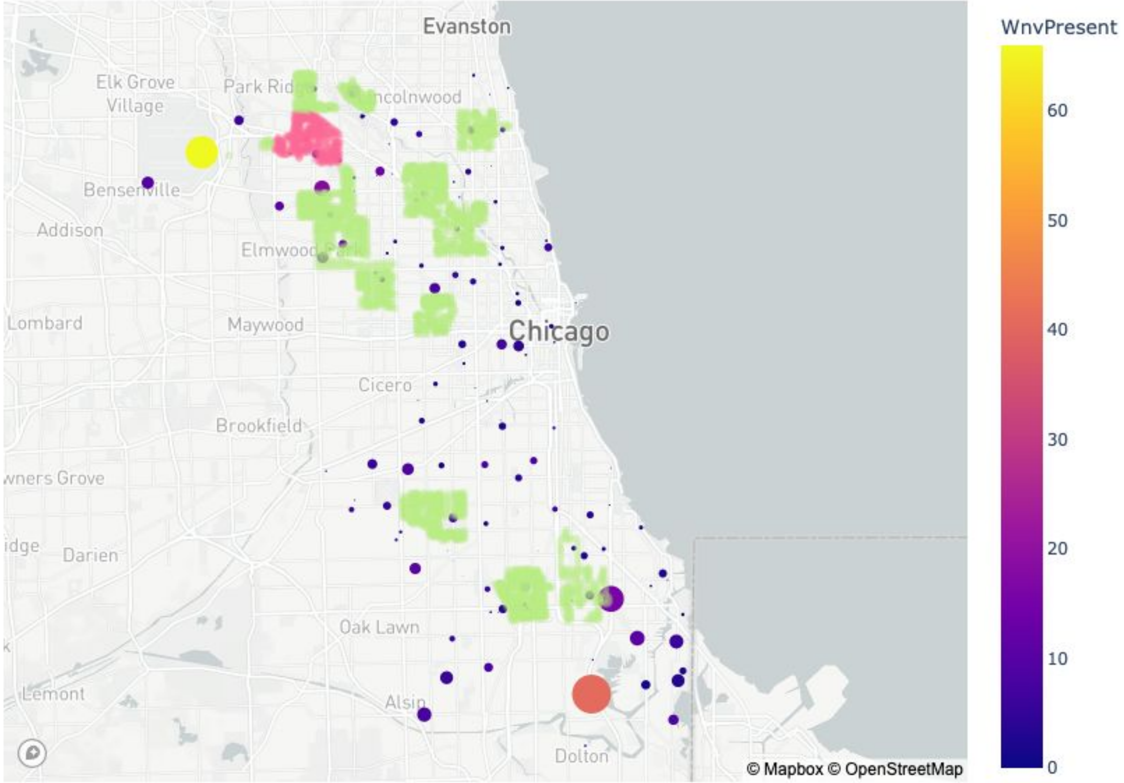






EDA/ Feature Engineering: Mosquitos traps and vector control effort

- More Spray effort on the northern area
- Occurence of WNV scattered across Chicago. Lesser on the central area
- 2 large clusters:
  O'Hare International Airport
  Lake Calumet



Legend:
Pink: spray effort in 2011
Green: spray effort in 2013
Circle: location of traps

Modelling & Tuning

02.

Modelling Process

1 Preprocessing

- One Hot Encode categorical features
- Standard scaling for numerical data
- Train-Test Split

2 Baseline Model

Dummy Classifier used to get a baseline score

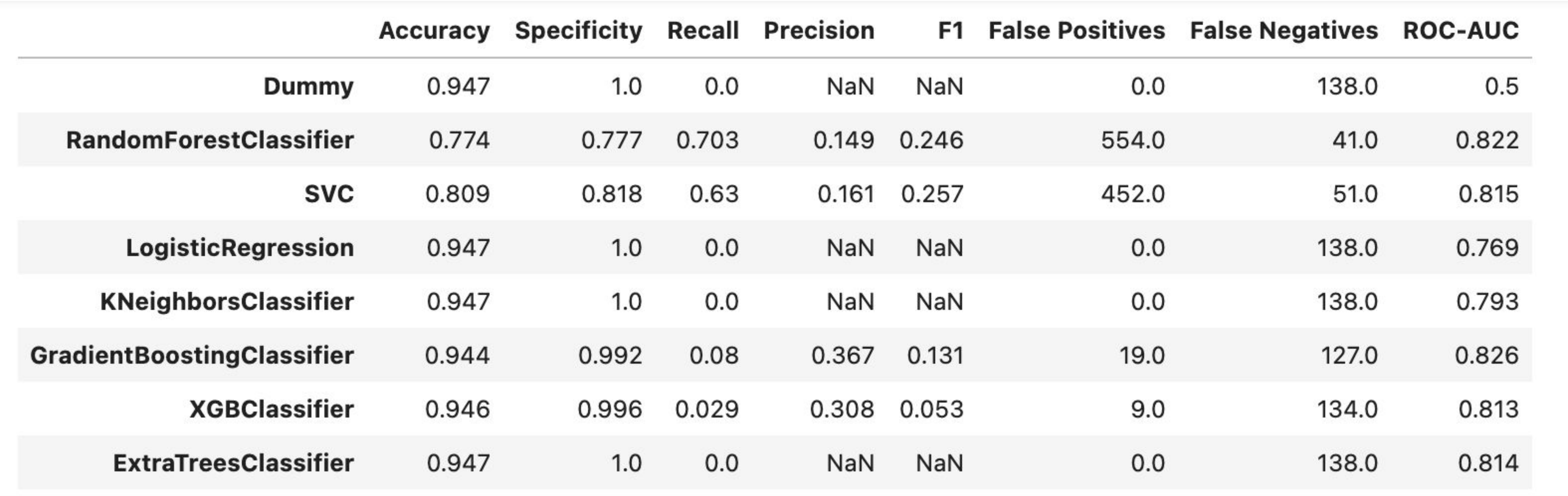
3 GridSearch through 7 Models

Random Forest, Support Vector, Logistic Regression, K Nearest Neighbors, Gradient Boosting, XGBoost, Extra Trees

4 GridSearch with SMOTE

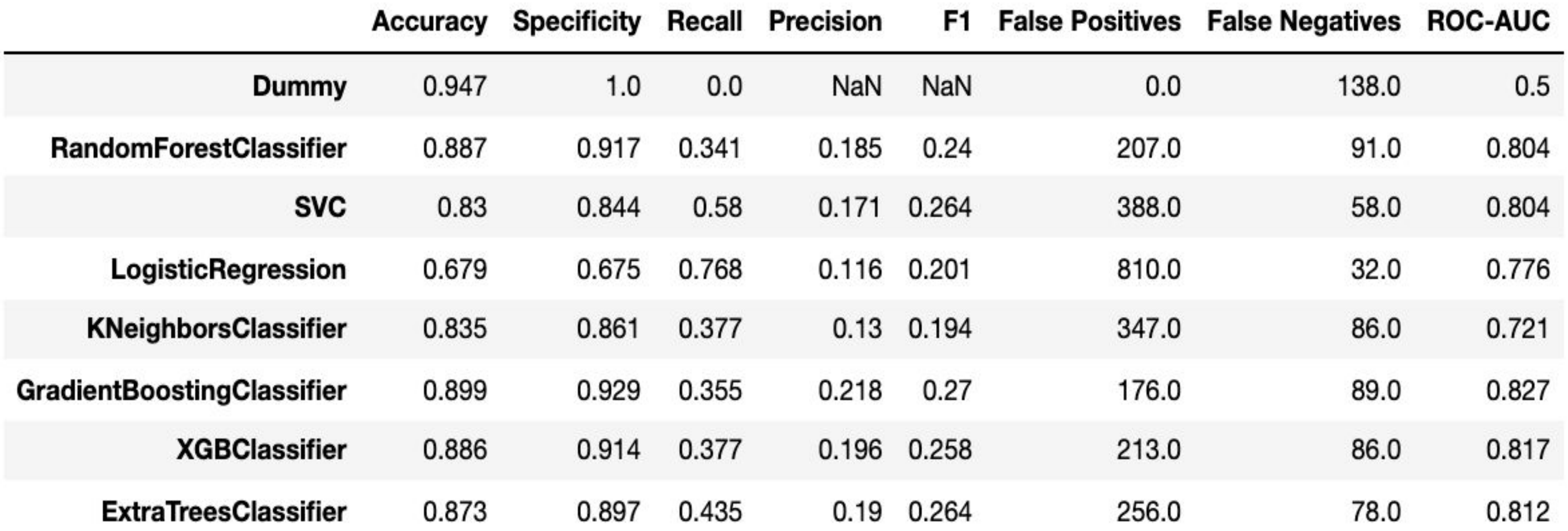
Use oversampling of minority class to overcome class imbalance

Summary of Results (without SMOTE)



	Accuracy	Specificity	Recall	Precision	F1	False Positives	False Negatives	ROC-AUC
Dummy	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.5
RandomForestClassifier	0.774	0.777	0.703	0.149	0.246	554.0	41.0	0.822
SVC	0.809	0.818	0.63	0.161	0.257	452.0	51.0	0.815
LogisticRegression	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.769
KNeighborsClassifier	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.793
GradientBoostingClassifier	0.944	0.992	0.08	0.367	0.131	19.0	127.0	0.826
XGBClassifier	0.946	0.996	0.029	0.308	0.053	9.0	134.0	0.813
ExtraTreesClassifier	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.814

Summary of Results (with SMOTE)



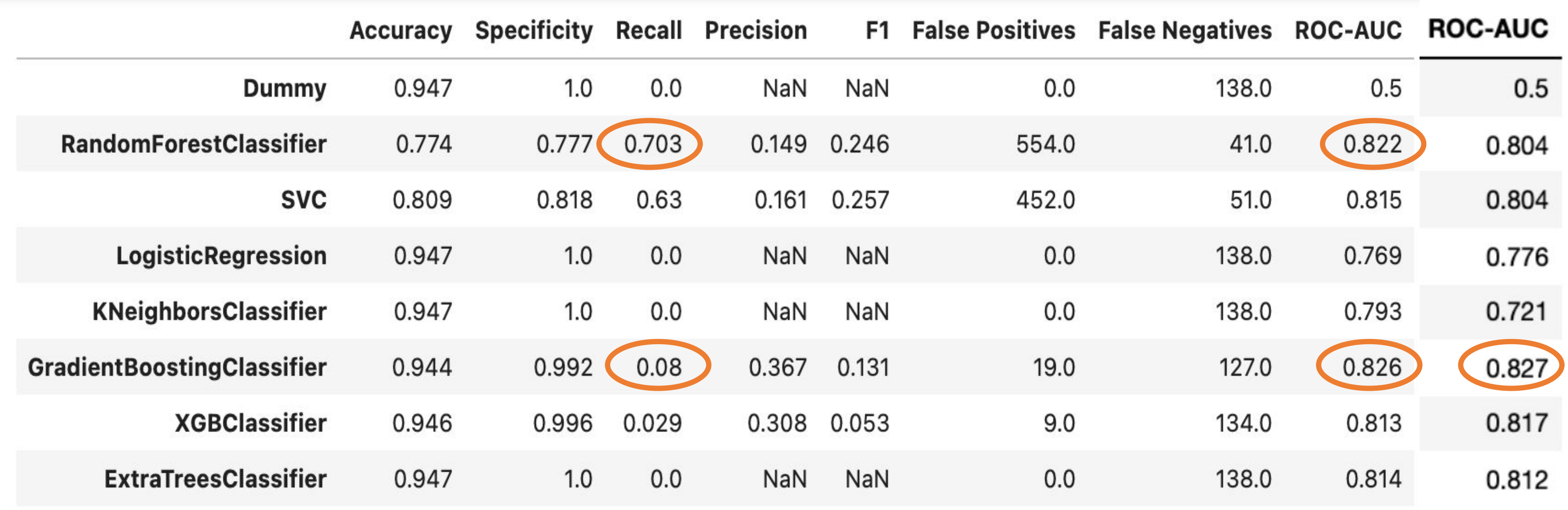
	Accuracy	Specificity	Recall	Precision	F1	False Positives	False Negatives	ROC-AUC
Dummy	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.5
RandomForestClassifier	0.887	0.917	0.341	0.185	0.24	207.0	91.0	0.804
SVC	0.83	0.844	0.58	0.171	0.264	388.0	58.0	0.804
LogisticRegression	0.679	0.675	0.768	0.116	0.201	810.0	32.0	0.776
KNeighborsClassifier	0.835	0.861	0.377	0.13	0.194	347.0	86.0	0.721
GradientBoostingClassifier	0.899	0.929	0.355	0.218	0.27	176.0	89.0	0.827
XGBClassifier	0.886	0.914	0.377	0.196	0.258	213.0	86.0	0.817
ExtraTreesClassifier	0.873	0.897	0.435	0.19	0.264	256.0	78.0	0.812

Model Evaluation

03.

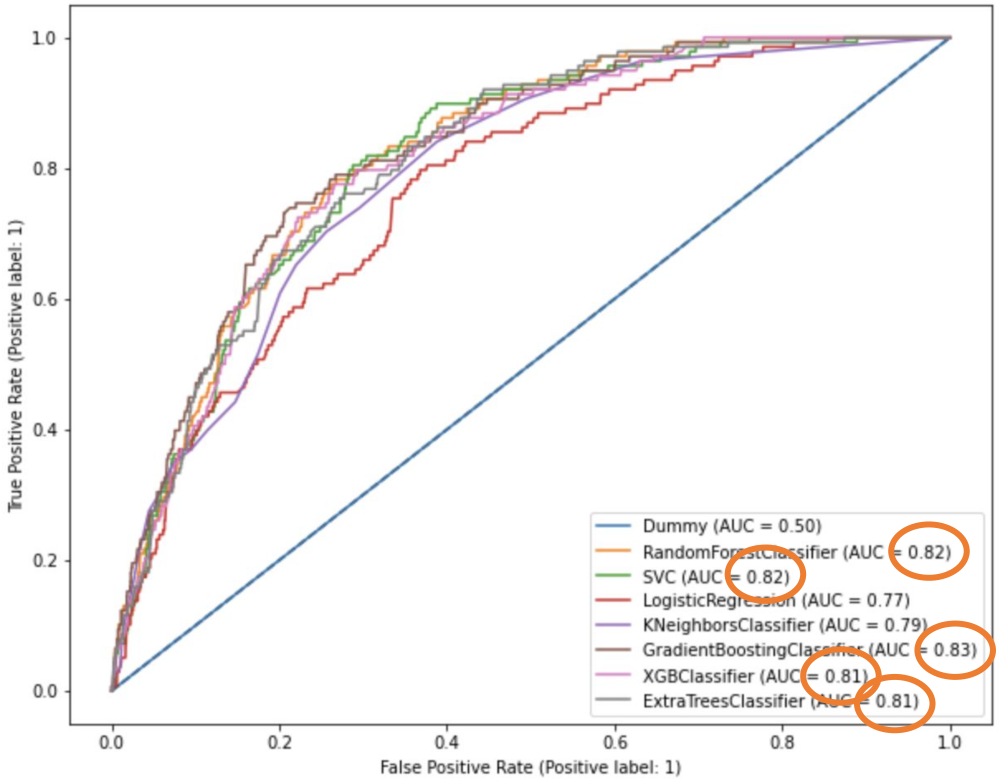
Summary of Results (without SMOTE)

with SMOTE

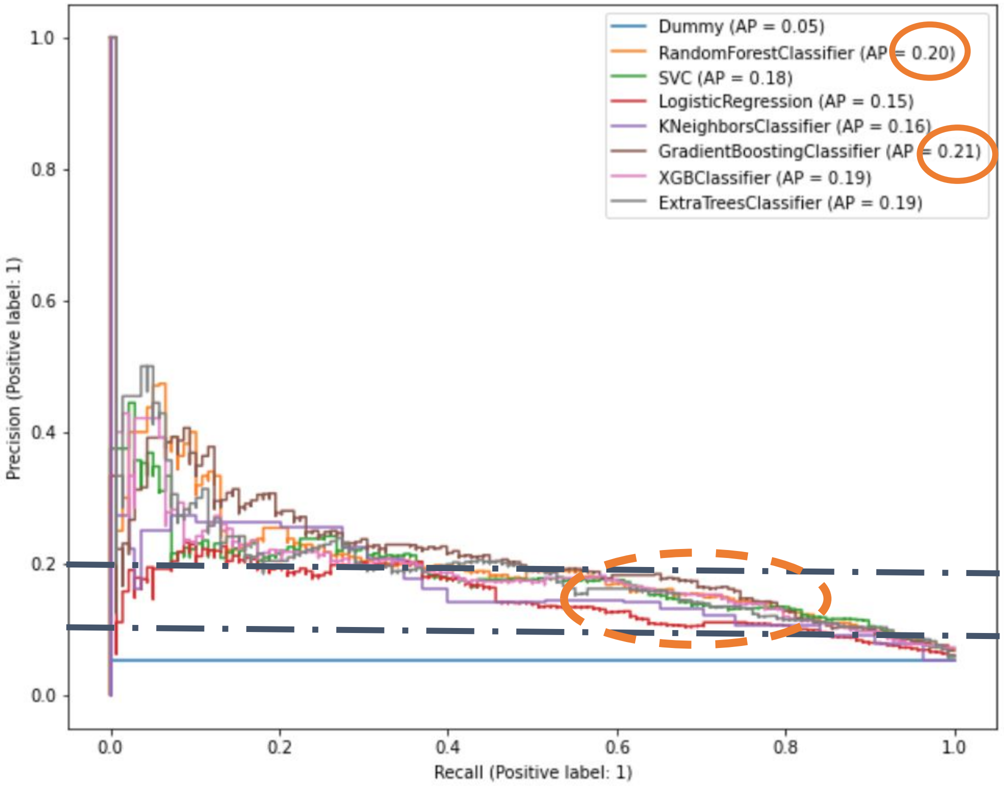


	Accuracy	Specificity	Recall	Precision	F1	False Positives	False Negatives	ROC-AUC	ROC-AUC
Dummy	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.5	0.5
RandomForestClassifier	0.774	0.777	0.703	0.149	0.246	554.0	41.0	0.822	0.804
SVC	0.809	0.818	0.63	0.161	0.257	452.0	51.0	0.815	0.804
LogisticRegression	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.769	0.776
KNeighborsClassifier	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.793	0.721
GradientBoostingClassifier	0.944	0.992	0.08	0.367	0.131	19.0	127.0	0.826	0.827
XGBClassifier	0.946	0.996	0.029	0.308	0.053	9.0	134.0	0.813	0.817
ExtraTreesClassifier	0.947	1.0	0.0	NaN	NaN	0.0	138.0	0.814	0.812

ROC Curve


1.0
0.8
0.6
0.4
0.2
0.0
True Positive Rate (Positive label: 1)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (Positive label: 1)
Dummy (AUC = 0.50)
RandomForestClassifier (AUC = 0.82)
SVC (AUC = 0.82)
LogisticRegression (AUC = 0.77)
KNeighborsClassifier (AUC = 0.79)
GradientBoostingClassifier (AUC = 0.83)
XGBClassifier (AUC = 0.81)
ExtraTreesClassifier (AUC = 0.81)

Precision-Recall Curve



To achieve a recall of 0.6-0.8 (catch 60%-80% of actual WNV positive), we can get a precision of 0.1-0.2 (for every 10 predicted as positive, 8-9 are false alarms).

Production Model

1 Classification Algorithm

Random Forest Classifier

2 Feature importance

- Daylight
- Longitude & Latitude
- Average and Lagged temperature

3 Metric Score

ROC AUC: 0.822
Recall: 0.703
Precision: 0.149

4 Possible improvement

Use neural network model

Cost Benefit Analysis

04.

Cost Benefit Analysis: Pesticides

Cost factors to consider

- Type of pesticides used
- Frequency of spraying
- Volume needed

Adulticides vs Larvicides

- Larvicides are 3-22x more expensive, with Altosid Liquid Larvicide Concentrate SR-20 2X2.5 being the most expensive ($4470/gallon).
- Adulticides: $170-300/gallon

Recommendations

- Minimize pesticide costs: use only adulticides
- Minimize transmission: use a combination of adulticides and larvicides

Cost Benefit Analysis: Spray methods

Cost factors to consider

- Culex Restuans more likely to transmit WNV
- Culex Pipiens more likely to get infected

Mosquito life cycle

- Culex mosquitoes have a life cycle of approx. 7-10 days (larvae to adulthood)
- Two weeks lag between mosquito growth and virus transmission

Recommendations

- Target highly infectious areas (WNV positive) by using drone technology to spray pesticides, because they complete the same task for 10-15% of the price of a helicopter
- Aligned spray with growth cycle

Cost Benefit Analysis: Spray locations

Cost factors to consider

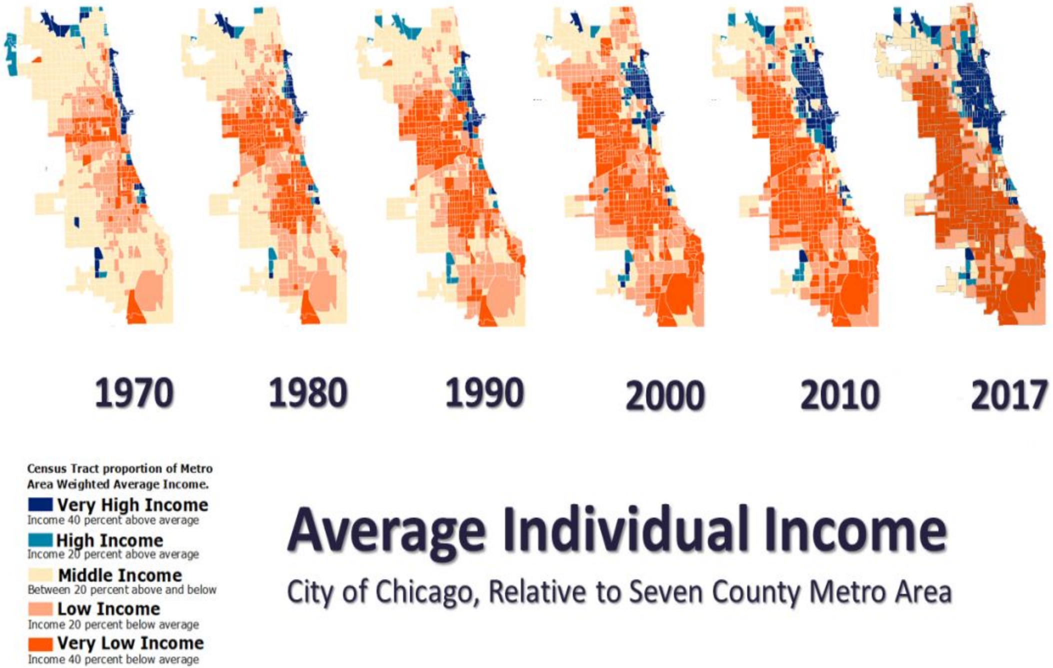
- Acute flaccid paralysis: (median $25K; range $5K – $283K)
- Encephalitis (median $20K; range $4K – $324K)
- Meningitis (median USD 10,556; range USD 0 – USD 260,748

Wealth distribution

- More spraying has been done in wealthier neighborhoods
- Less privileged individuals are less likely able to afford exorbitant costs of insurance and healthcare

Recommendations

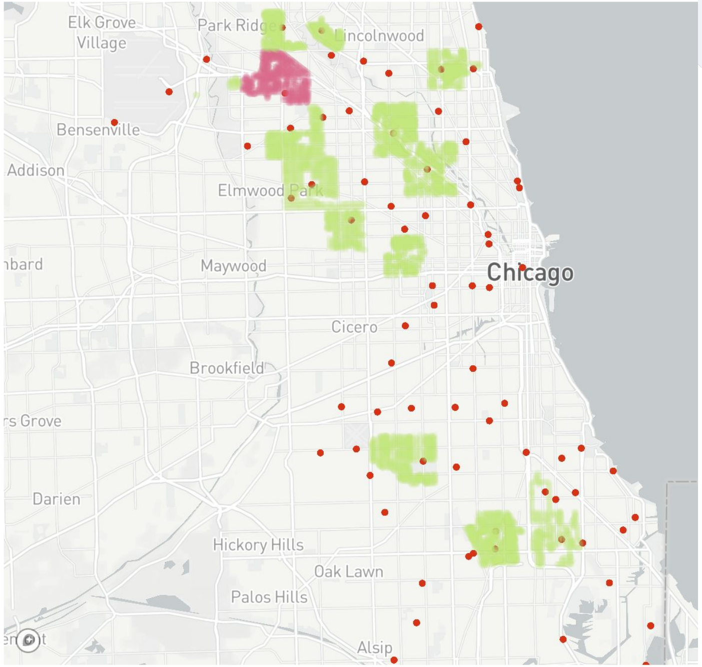
- More prudent to increase spray volumes for lower income neighborhoods


1970
1980
1990
2000
2010
2017
Census Tract proportion of Metro Area Weighted Average Income.
Very High Income
Income 40 percent above average
High Income
Income 20 percent above average
Middle Income
Between 20 percent above and below
Low Income
Income 20 percent below average
Very Low Income
Income 40 percent below average

Average Individual Income

City of Chicago, Relative to Seven County Metro Area

Traps in 2013 & Sprays in 2011 & 2013



Cost Benefit Analysis: Expert Collaboration

Cost factors to consider

- Data can be presented to scientists specializing in mosquito biology and infectious diseases

Expertise on subject matter

- Have a greater mastery of this subject matter and more expertise in feature selection and feature engineering.
- They can utilize their in-depth expertise of mosquito lifecycle, habitats and breeding habits to generate a model with an edge in predicting WNV transmission

Recommendations

- Help to rapidly identify and treat WNV cases before the individual progresses to a more severe stage of the disease, especially in neighborhoods with a high infection rate. This will definitely increase the odds of patient recovery and decrease their medical bills

Thanks !