A Prevention Model for the West Nile Virus Outbreak

Project 4 | Group 3

Ariffin Sarhid | Chris-Anabel Khaw | Joseph Gan | Melvin Chandra

Problem Statement

West Nile Virus has been infecting City of Chicago since 2004 causing serious **neurological illnesses** that can result in **death**.

Leveraging on information such as **weather, locations & species of mosquito,** an accurate method of **predicting outbreaks** of West Nile virus shall be developed to assist the City of Chicago in efficiently and **effectively allocate resources** towards **preventing transmission** of this potentially deadly virus

Metrics of Assessment

ROC AUC

Audience

Chicago Department Of Health

Agenda

- Data Cleaning, EDA & Feature Engineering
- Handling of Imbalance Data
- Data Modelling: Binary Classification
- Cost Benefit Analysis

Data Description

- ~33 variables describing locations, mosquitoes species & weather
- Data collected from 2007 to 2013
- 10, 506 samples of data

Data acquired from https://www.kaggle.com/c/predict-west-nile-virus/data

DATA CLEANING

Good data quality is essential for modelling..

Data Cleaning

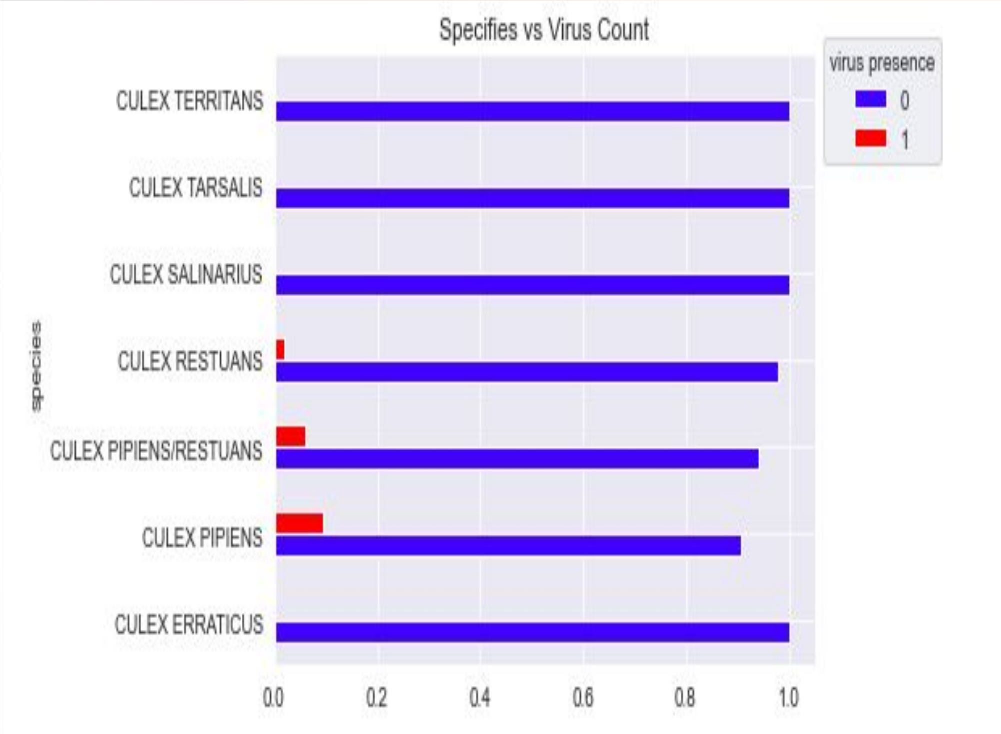
- Removed duplicate values
- Dropped Columns with High Missing Values & Rows with missing values < 10% (99% of data still retained)
- Mean Imputation & arbitrary imputation on missing values in weather dataset
- Combining weather & main dataset on “date” feature

EDA & FEATURE ENGINEERING

Exploring Weather, Locations, Timing, Virus Species to find any correlations with Presence of Virus...

Are Mosquito Species an Indication of Virus Presence?

- There seems to be some suggestion so

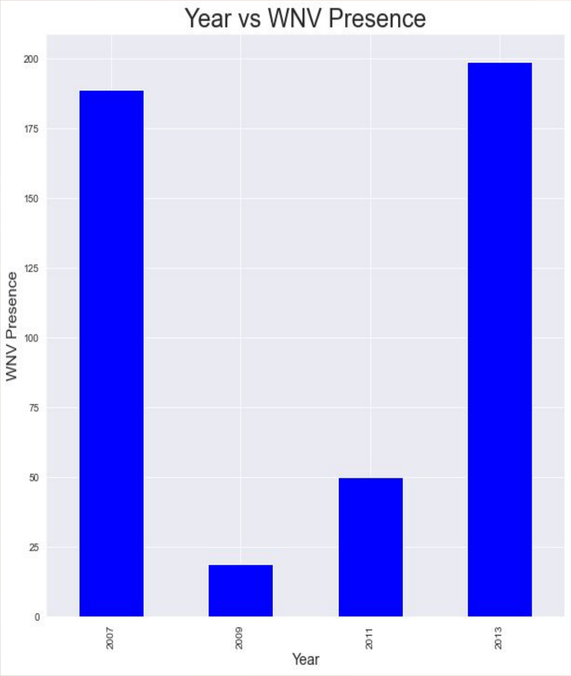


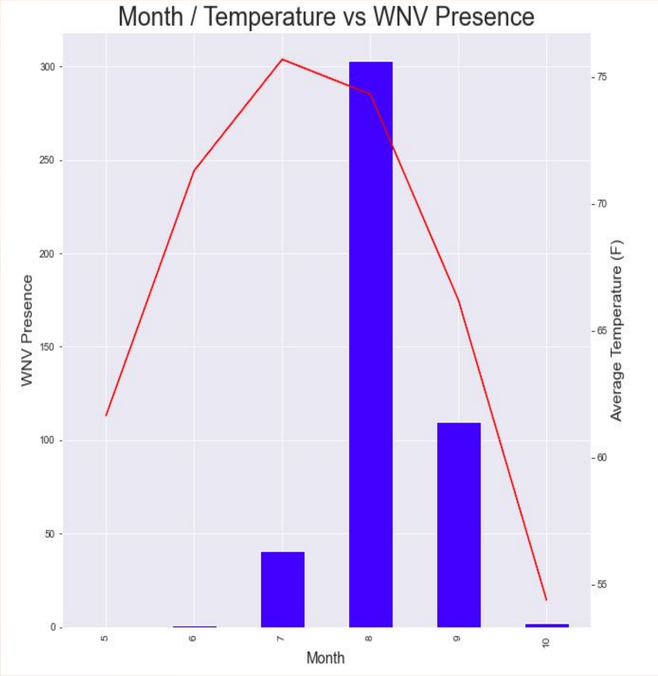
- ‘CULEX PIPENS’: 3
- ‘CULEX PIPENS/RESTUANS’: 2
- ‘CULEX RESTUANS’: 1
- OTHERS: 0

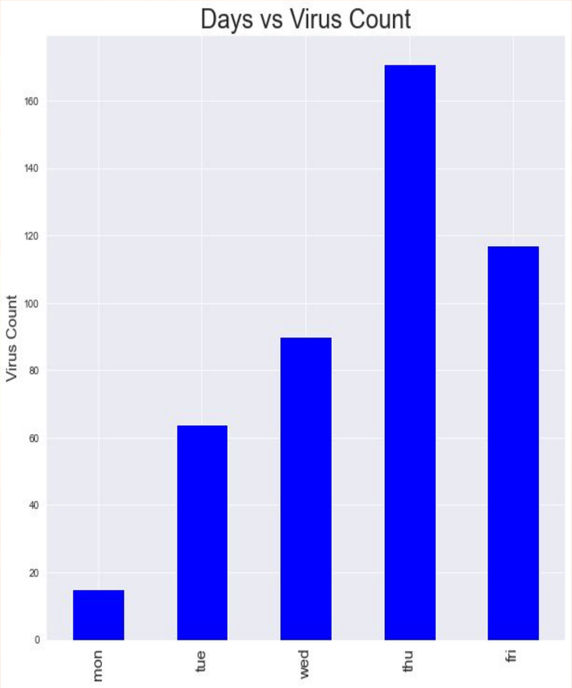
* **Target Guided Label Encoding:** Ranking of Species according to the highest count of virus presence | **Rare Label Encoding:** Grouping of low count of virus presence into one group.

Are Timing & Temperature Good Predictors of Virus Presence?

- There is a correlation between Date/Timing, Temperature and Virus Presence





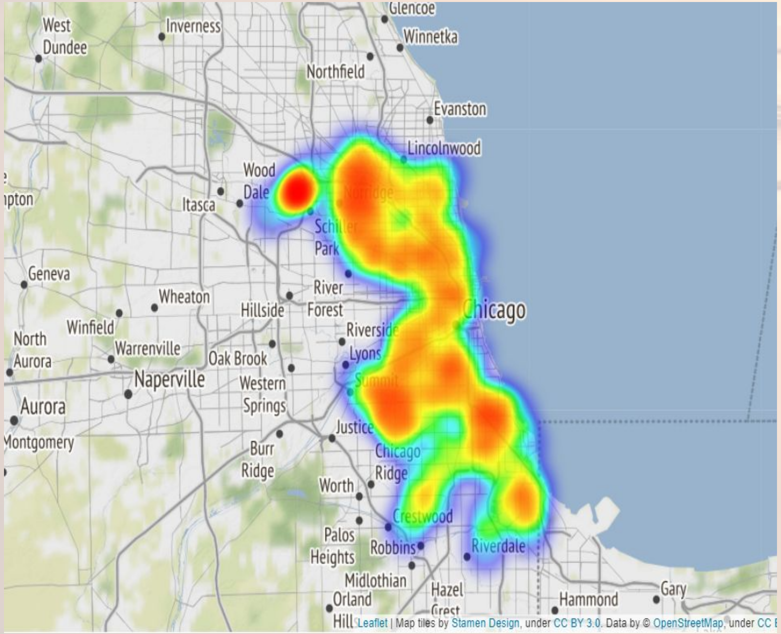


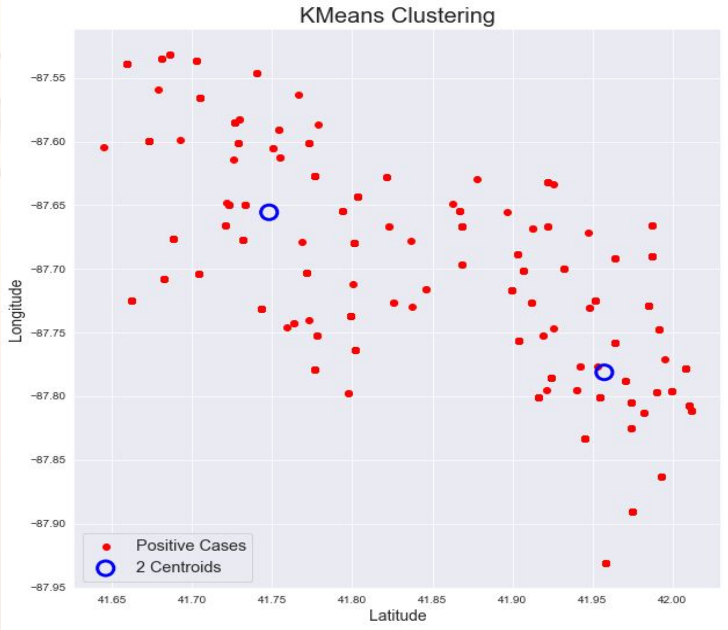
Created Date Features: Year, Month, Day

Does geography hint at the likelihood of Virus outbreak?

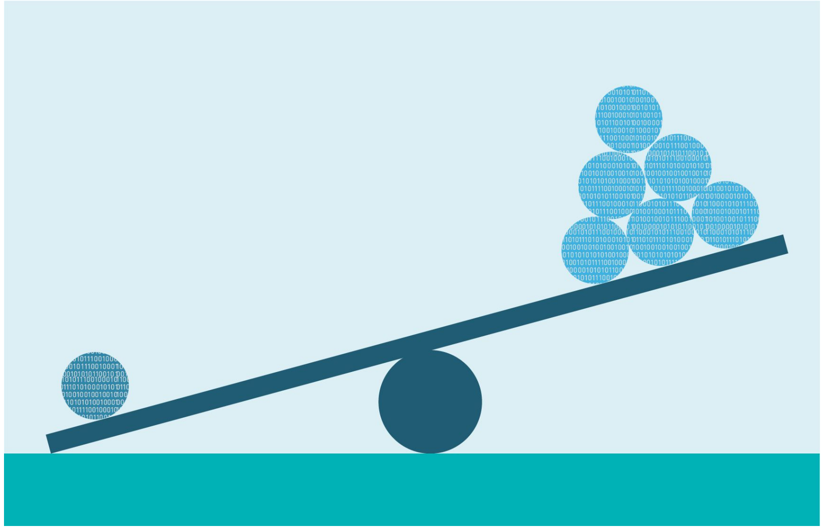
- Indeed so as there were hotspots detected

*Performed **K-means clustering** to find the centroid of the clusters and **c**reate a distance feature to the hotspot centroid*





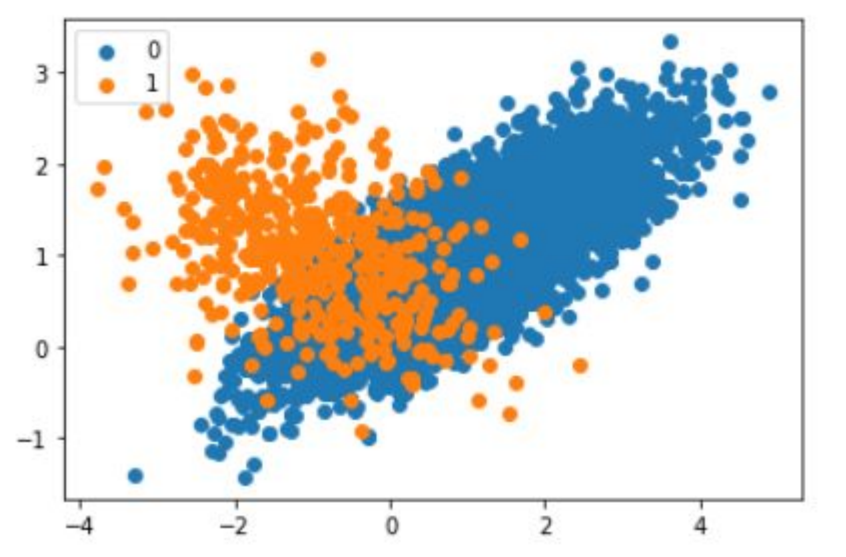
Imbalance Data



Imbalance data will deteriorate a model performance if it's not dealt with...

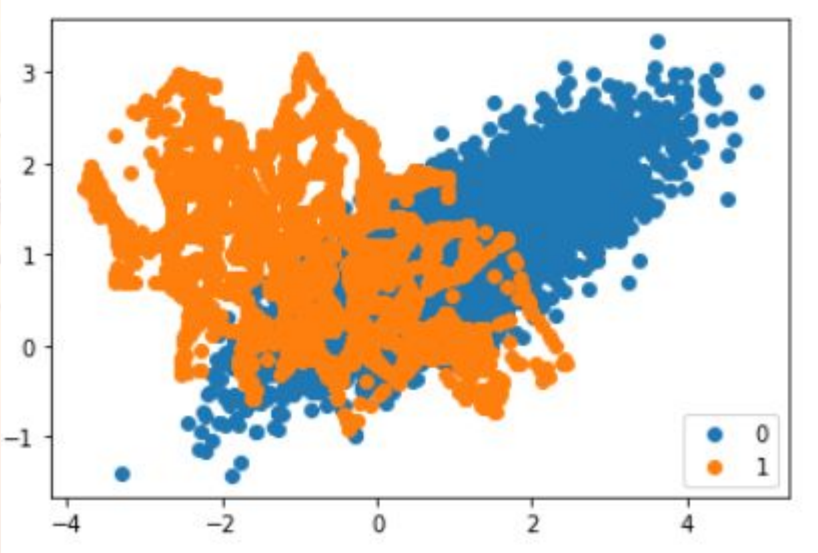
Addressing **Data Imbalance** Through **SMOTE**

- In the pre-modelling process, we found out that the data is heavily imbalanced
- From the visuals, we can see that our target ('wnvpresent') has
  - 95% of 0
  - 5% of 1



Addressing **Data Imbalance** Through **SMOTE**

- We use Synthetic Minority Oversampling Technique (SMOTE) to address the data imbalance
- SMOTE is achieved by duplicating examples from the minority class in the training dataset prior to fitting a model
- SMOTE should be used only on the training set else there will be data leakage and will cause overfitting



Model Selection

Logistic Regression

Random Forest Classifier

Gradient Boosting Classifier

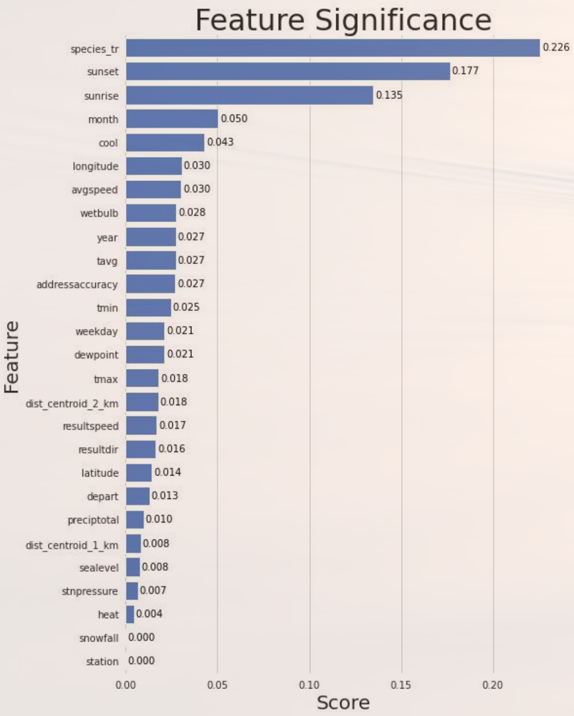
Data Modelling: Binary Classification Problem

- Pipeline:
  - SMOTE
  - GridSearch CV

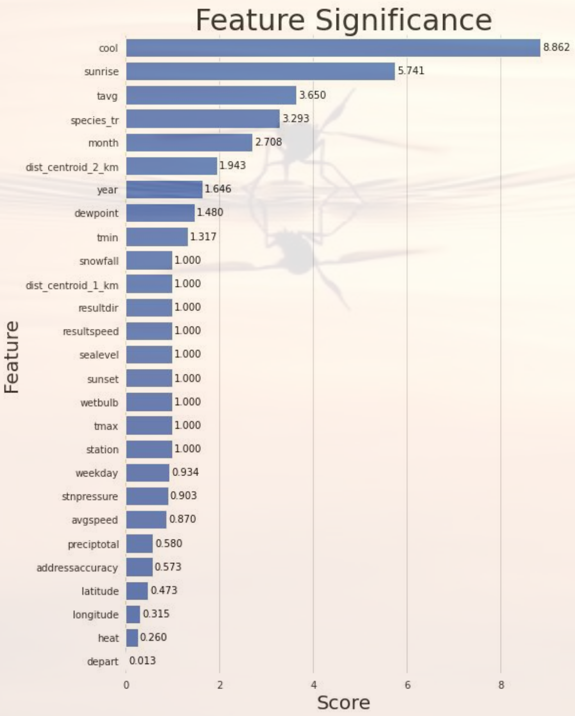
	Metrics	Random Forest Classifier	Logistic Regression	Gradient Boost
Scoring (WNV Present)	ROC AUC Train	0.861	0.818	0.917
	ROC AUC Test	0.828	0.800	0.848
	Precision	0.16	0.13	0.25
	Recall	0.74	0.73	0.46
	F1-Score	0.27	0.23	0.32

Understanding the significance of variables

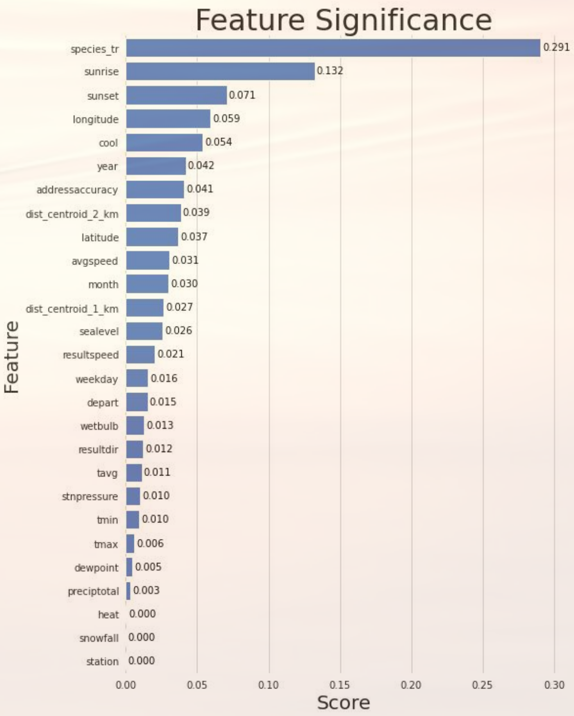
<< Random Forest Classifier >>



<< Logistic Regression >>



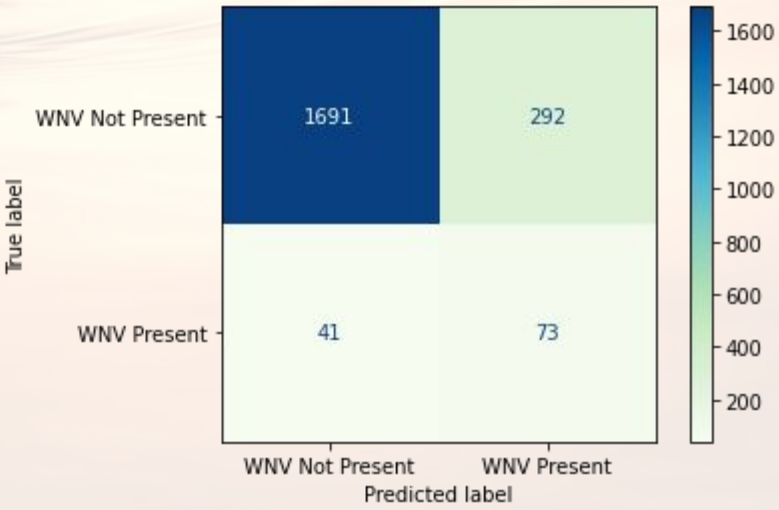
<< Gradient Boost>>



Bias-Variance Tradeoff: Tuning

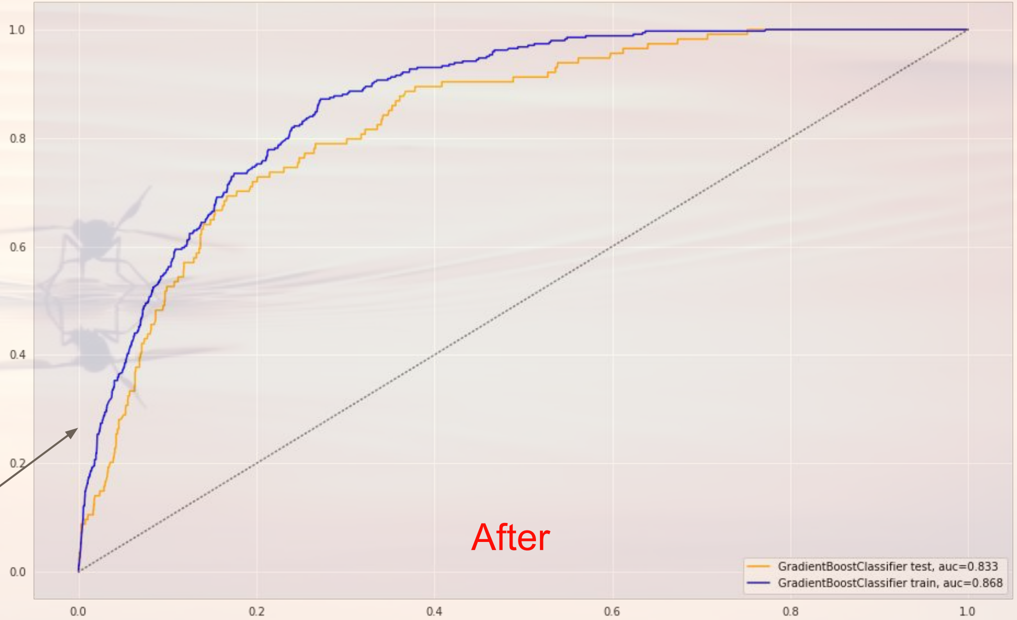
- Remodel with top 10 features
- Reduced no. of trees
- Increased learning rate

	Metrics	Gradient Boost	Gradient Boost (Pruned)
Scoring (WNV Present)	ROC AUC Train	0.917	0.864
	ROC AUC Test	0.848	0.833
	Precision	0.25	0.20
	Recall	0.46	0.64
	F1-Score	0.32	0.30



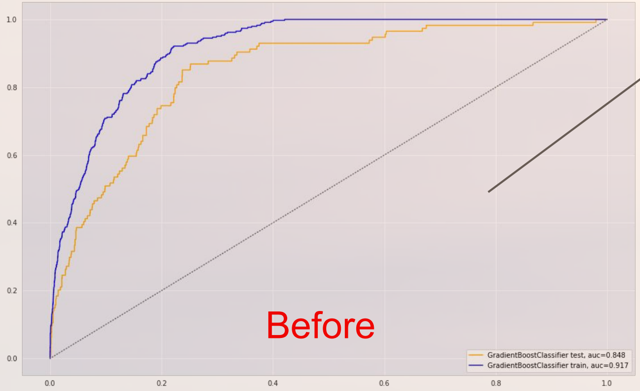
Bias-Variance Tradeoff: Tuning

- Gradient boost model provides highest ROC AUC score
- Gradient boost model has the highest variance
- More hyper parameter tuning
- Limit no. of features
- Kaggle out of sample score of 0.79



Name	Submitted	Wait time	Execution time	Score
kaggle.csv	2 days ago	1 seconds	1 seconds	0.79403

Complete



Cost Benefit Analysis

How many west nile virus cases to financially justify preventive interventions?

Cost Benefit Analysis

Loren Barber et. al, Research on West Nile Virus impact in Sacramento County, 2005

- The total economic impact of WNV was $2.98 million.
- Only 15 cases of West Nile neuroinvasive disease would need to be prevented to make the emergency spray cost-effective

A similar study is performed on the City of Chicago with cross referencing.

Cost Benefit Analysis

Factors considered and Information required

- Cost Impact of West Nile Virus
- Fixed Cost for Vector Control (eg: Plane Rental, Human Resource, etc)
- Variable Cost for Vector Control (eg: Insecticide, Plane fuel etc) (Larger area, Cost increase)
- Effectiveness of Spray

Assumptions

- Inflation across all items are same, hence cost will not be adjusted for inflation
- Medical cost treatment for Sacramento County is the same for City of Chicago
- Cost for Vector Control for Sacramento County is the same for City of Chicago
- Breakdown of Cost of Vector Control between Fixed Cost & Variable Cost

Cost Benefit Analysis

Effectiveness of Spray

Carney et al. (6) has documented that after the Sacramento County emergency spray there are no cases within the spray area.

Hence we assumed that if we were to spray these area in City of Chicago, 100% of the cases can be eliminated.

Cost Table (Reference from Loren Barber et. al)

		Cost
WNV Impact Cost per Case	$	49,565
Vector Control		
Fixed Cost (per Spray)	$	351,631
Variable Cost (spray / km^2)	$	736

Cost Benefit Analysis

Break-even Table

With the WNV Cost Impact per case of ~ $ 50,000, break even case as below
Best case scenario: 8 cases
Worst case scenario: 16 cases

Table:2: Cost Benefit Analysis Summary

	Break Even Table					
	unit	**Best Case Scenario**	25% of chicago Area	50% of chicago Area	75% of chicago Area	**Worst Case Scenario (100%)**
Infected Area	km²	1	147.5	295	442.5	590
Fixed Cost per Spray (A)		$ 351,631	$ 351,631	$ 351,631	$ 351,631	$ 351,631
Variable Cost per spray (B)		$ 736	$ 108,505	$ 217,011	$ 325,516	$ 434,021
Total Spray Cost (A + B)		**$ 352,366**	**$ 460,136**	**$ 568,641**	**$ 677,146**	**$ 785,652**
Treament Cost per Case (C)		**$ 49,565**	**$ 49,565**	**$ 49,565**	**$ 49,565**	**$ 49,565**
Cases Required to break even (A+ B) / C	case	**8**	**10**	**12**	**14**	**16**

Conclusion

- Best Model (Gradient Boost scores 0.79 roc auc on Out of Sample [Kaggle])

- Top Feature Importance to predict Presence of Virus :
    1. Species
    2. Sunrise
    3. Sunset
    4. Longitude
    5. Cool

- Best Case Scenario for Break-even: 8 cases

- Worst Case Scenario for Break-even: 16 cases

Recommendation

- Model to find out the link between presence of virus vs number of infected cases.
- Researching on additional features to improve model accuracy such as number of mosquitoes.
- Revalidate the Assumption made in Cost Benefit Analysis with Data of City of Chicago.