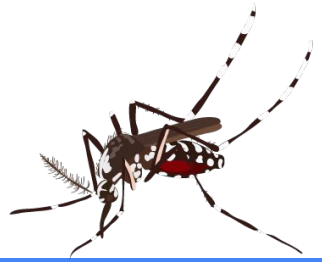
Predicting West Nile Virus

& Cost-Benefit Analysis for Spraying



Project 4

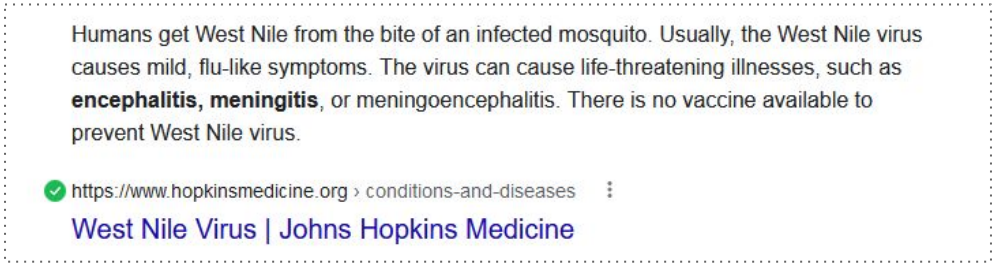
Team 1: Cel, Harry, Xiuting, YX

Introduction & Problem Statement

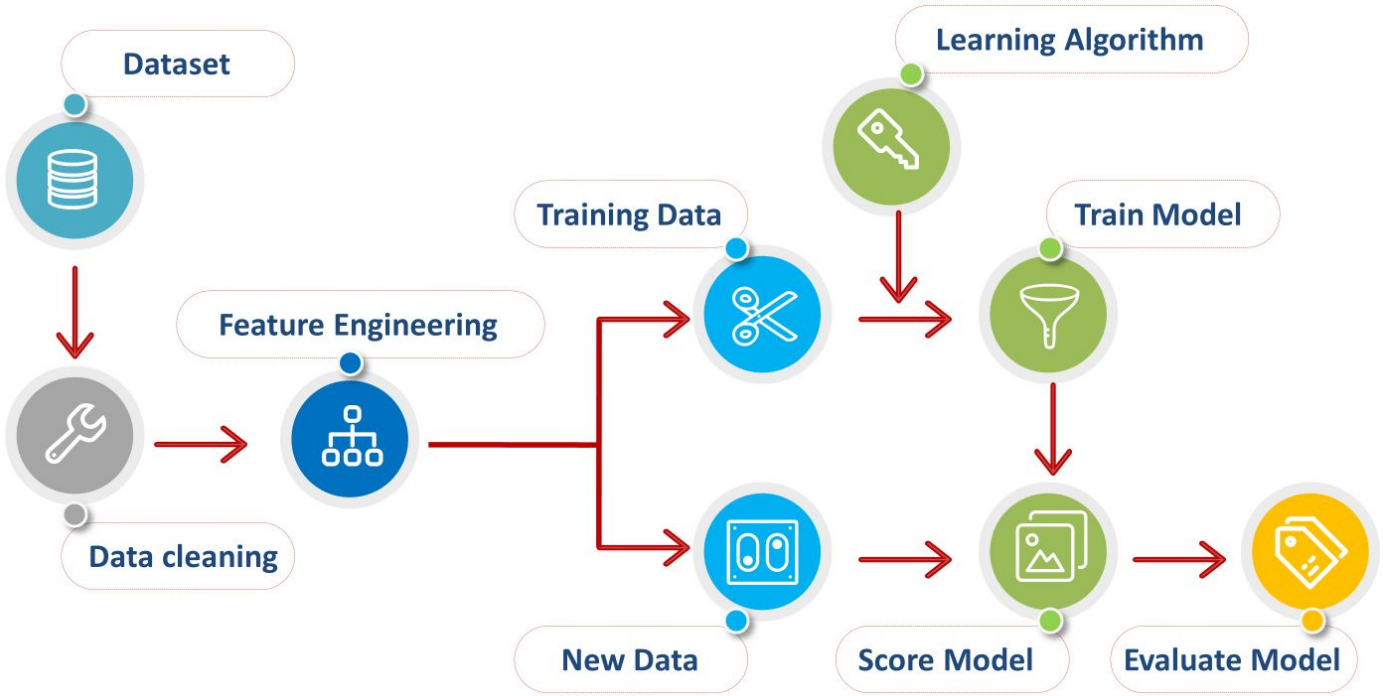
West Nile Virus still plagues citizens in Chicago, Illinois, *today*.

To make Chicago homes safer for all, our team of Data Scientists have

- analyzed available data to **predict likelihood of incidence of the disease**, and
- present a preliminary **cost-benefit analysis** to inform health authorities on Zenivex spray-based interventions



Methodology : Explore, Process, Engineer, Select, Model, Iterate



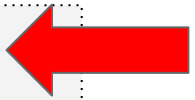
Source: https://medium.datadriveninvestor.com/data-preprocessing-3cd01eefd438

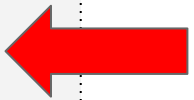
Metric Selection

Kaggle: Area under Receiver Operating Characteristic Curve (AUROC) – Balance between cost-benefits

Precision: Choose to avoid false positives (spray costs priority)
Tp / (Tp + Fp)

Recall: Avoid false negatives (save the people priority)
Tp / (Tp + Fn)





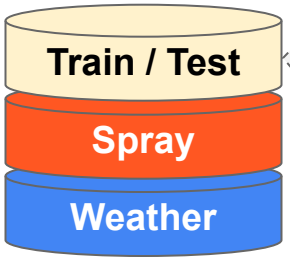
Metric Selection

Kaggle: Area under Receiver Operating Characteristic Curve (AUROC) – Balance between cost-benefits

Recall: Avoid false negatives (save the people priority)
Tp / (Tp + Fn)

Precision: Choose to avoid false positives (spray costs priority)
Tp / (Tp + Fp)

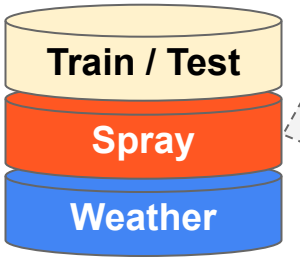
Data



Prima-Facie

- Useful information like mosquito population numbers (trapped) at specific locations and detection of WNV for trapped mosquitoes
- Test data does not contain number of mosquitoes

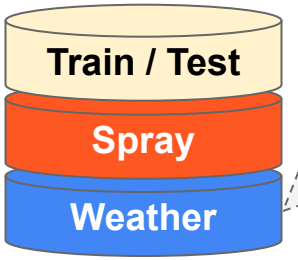
Data



Prima-Facie

- Spray data (some months only) in 2011 and 2013, mainly indicates timing and location of spray
- Drop time - not present in any other dataset
- Drop NAs, duplicates
- Create columns - year, weekofyear
- Change data types - date, latitude, longitude

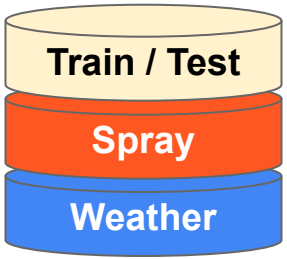
Data



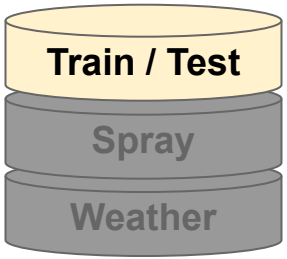
Prima-Facie

- A number of weather variables measured
- Data comes from two stations
- Some data would mirror others (e.g. sunrise / sunset timings, station pressure readings and sea level)

Pre-processing

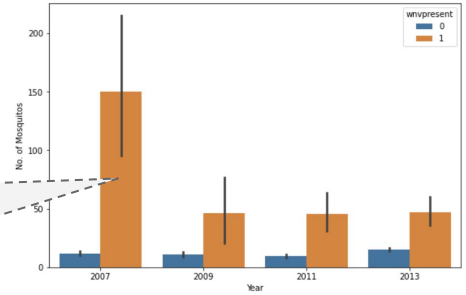


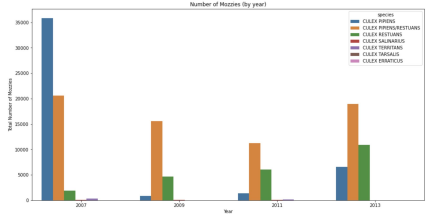
Exploratory Data Analysis: WNV across the years

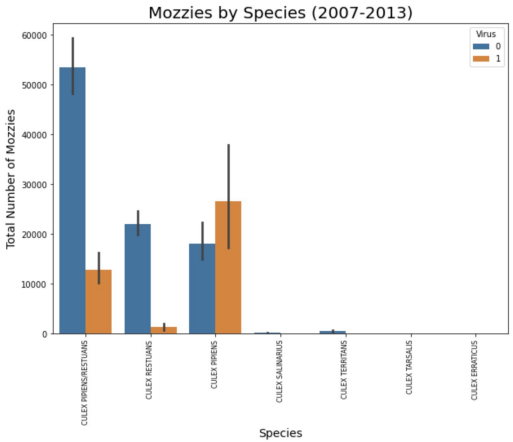


2007 had the highest number of WNV cases

- This spike of WNV also coincided with a high number of culex pipiens and pipiens/restuans species





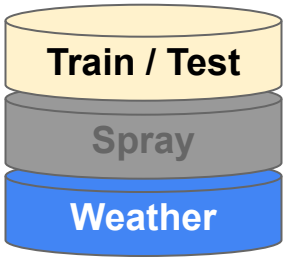



```
Percentage of CULEX PIPIENS/RESTUANS with west nile virus is: 19.39%
Percentage of CULEX PIPIENS with west nile virus is: 59.65%
Percentage of CULEX RESTUANS with west nile virus is: 5.76%
```

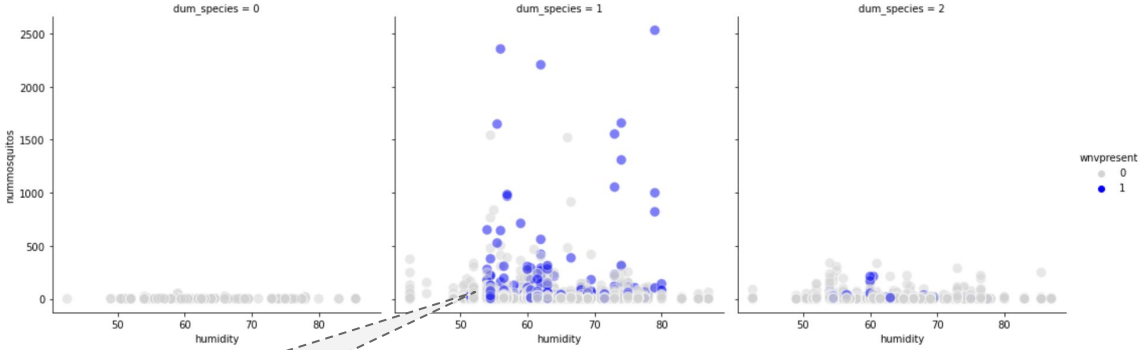
Encoded species based on WNV probability

- 1 for culex pipiens + pipiens/restuans
- 2 for culex restuans

Exploratory Data Analysis: Weather and WNV

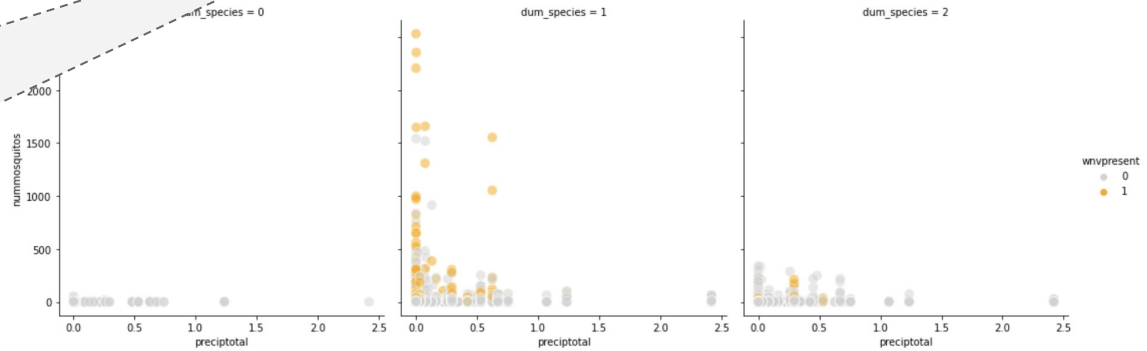


Recall:
1 - culex pipiens + culex pipiens/restuans
2 - culex restuans

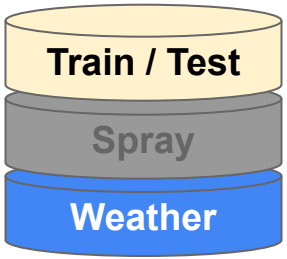


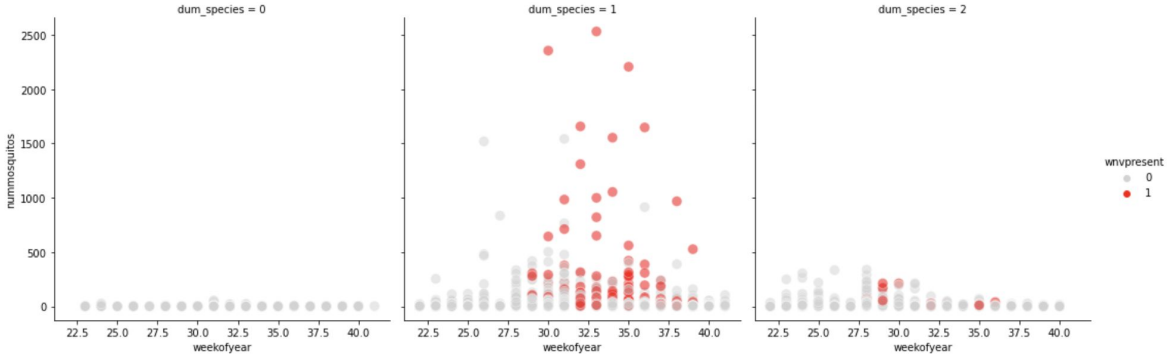
For wet weather features:

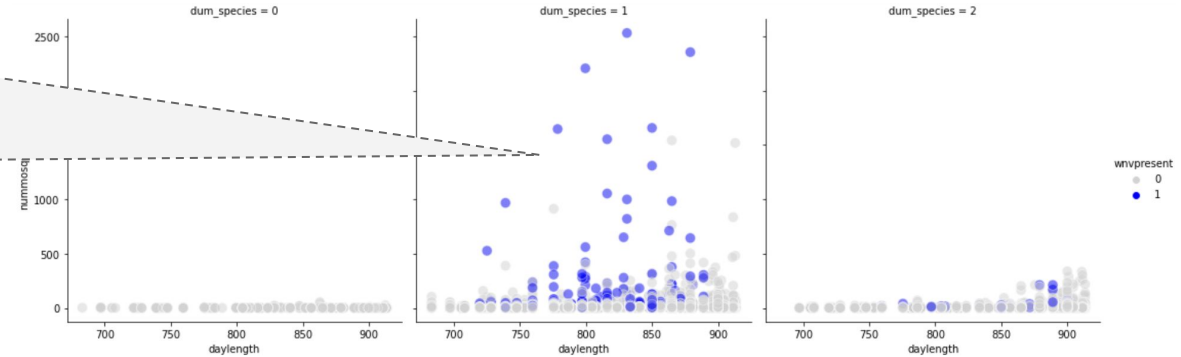
- Propagation of WNV most likely for culex pipiens and pipiens/restuans at humidity range 55-80 ;
- Precipitation 0.0 - 0.5
- But these ranges only accounted for 3-5% of total data



Exploratory Data Analysis: Weather and WNV



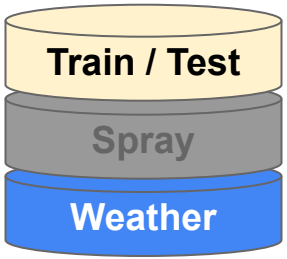




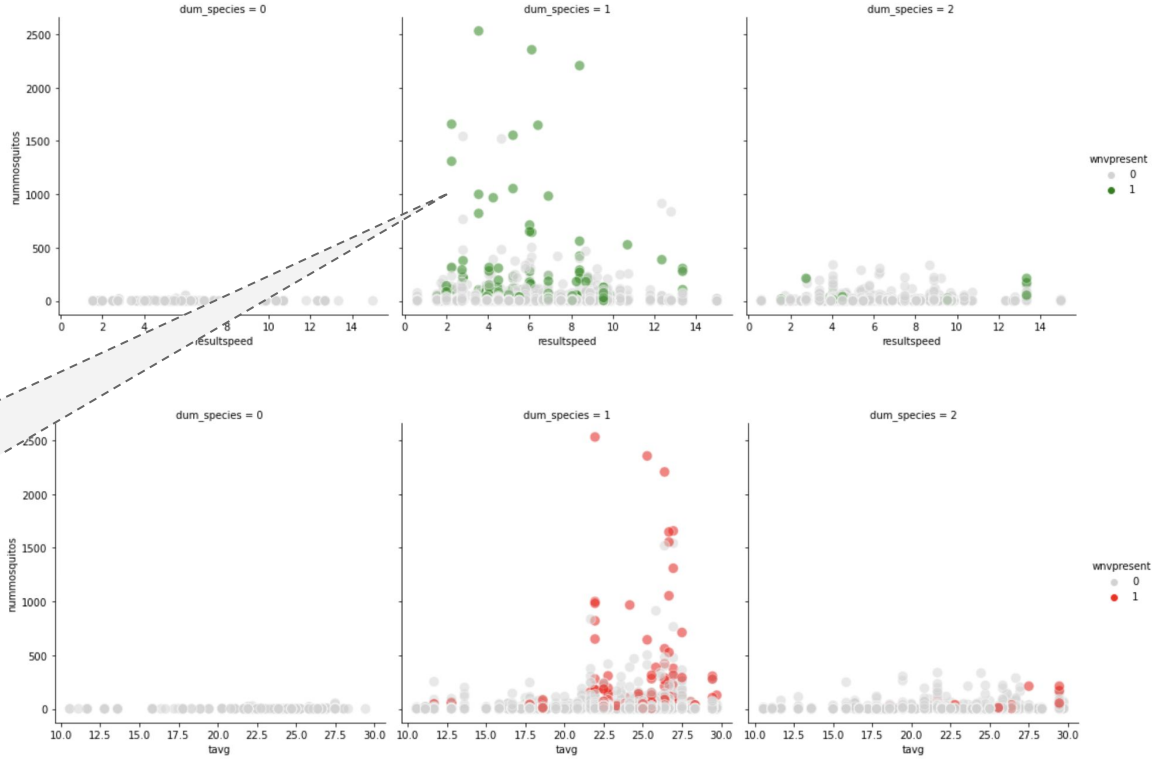
For time related features:

- No significant pattern for culex pipiens and pipiens/restuans
- Propagation of WNV for culex restuans seem to occur more between weeks 28-30, and when daylengths are longer

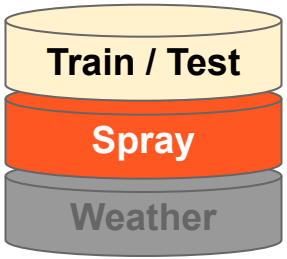
Exploratory Data Analysis: Weather and WNV

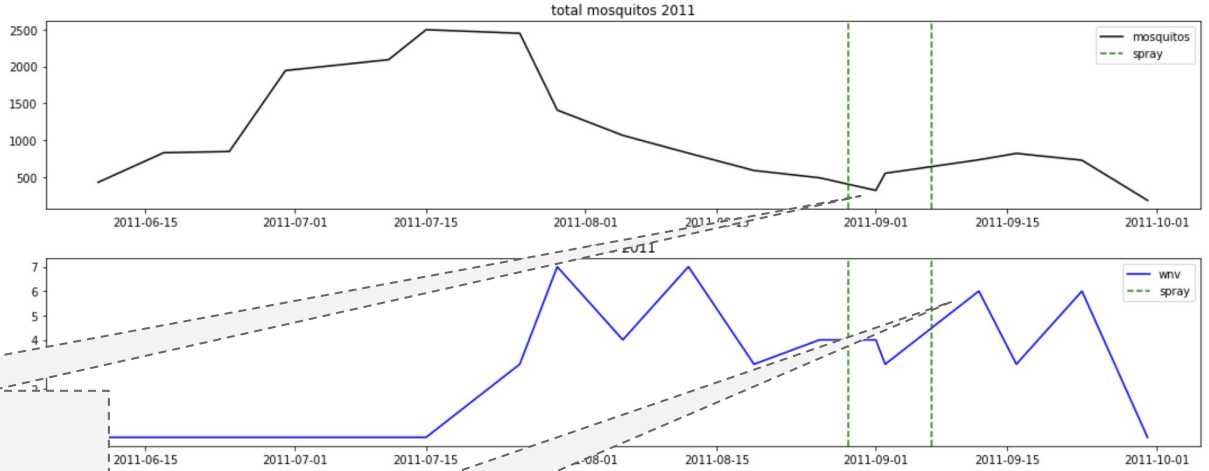


- No significant pattern for wind speed features across all species
- Propagation of WNV seems more likely at higher temperatures (27.5-30 deg C)



Exploratory Data Analysis: Spray Effectiveness (2011)





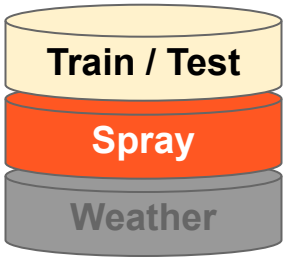
Spray #1

- Killed some mosquitoes
- WNV carriers were constant and only decreased a few days later

Spray #2

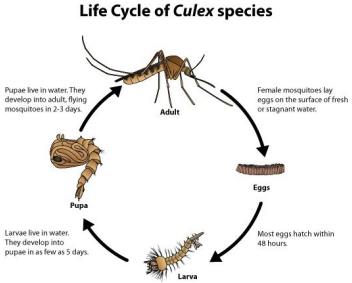
- Little to no effect on total mosquitoes
- WNV carriers increased exponentially compared to the total number of mosquitoes

Exploratory Data Analysis: Spray Effectiveness (2013)



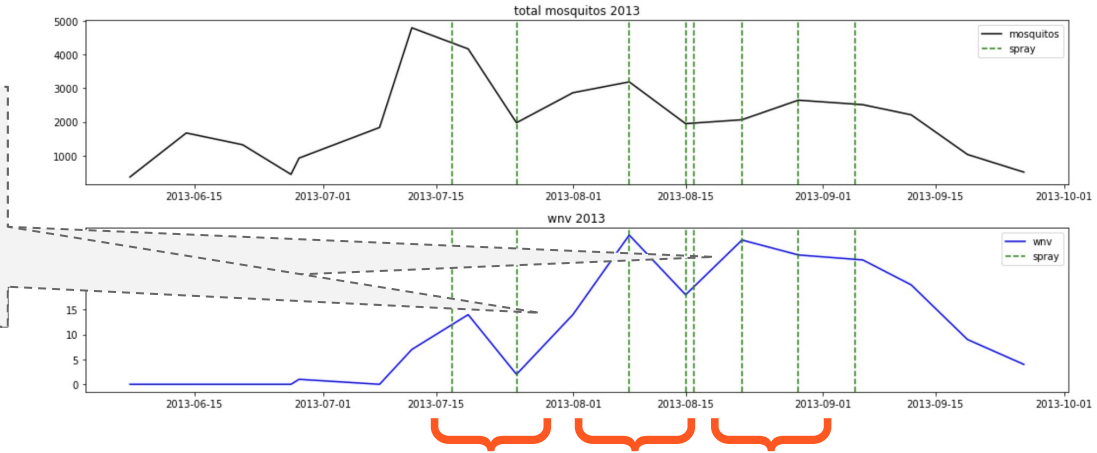
In general,

- Similar pattern to 2011
- Seasonal pattern → corresponds to mosquito life cycle (~14 days)

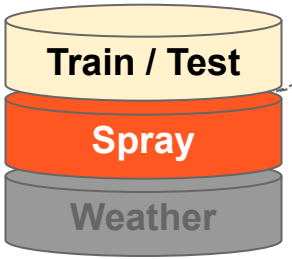


Spraying ✖

- Limited effectiveness on reducing mosquito populations
- WNV carriers still seem to be present and follows the 14 day life cycle



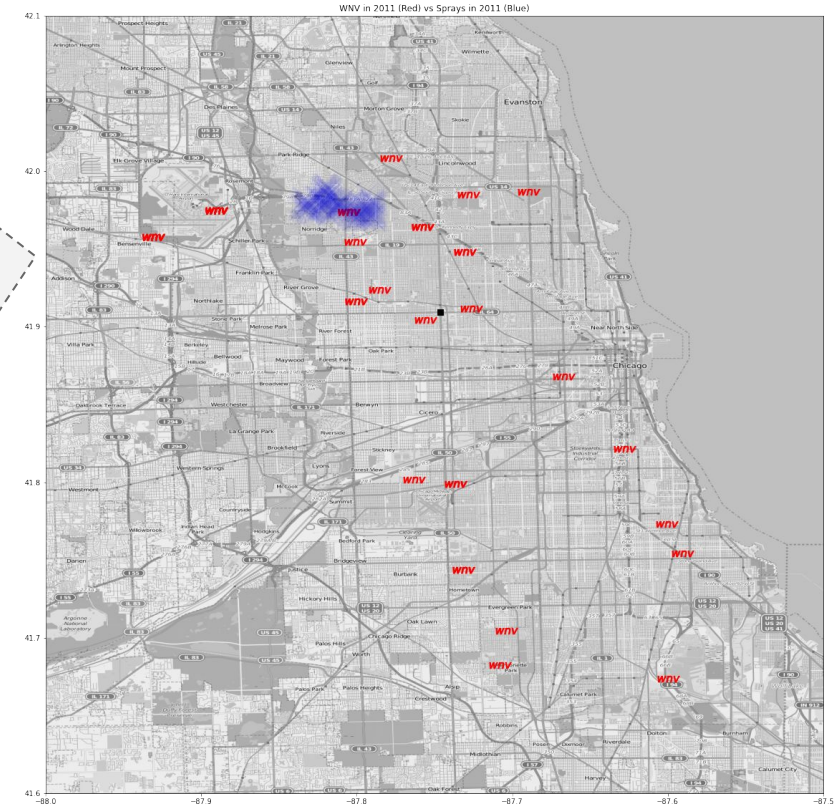
Exploratory Data Analysis: Maps 2011



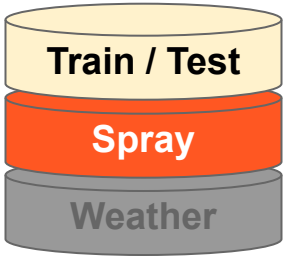
Sources: Train / Spray Data

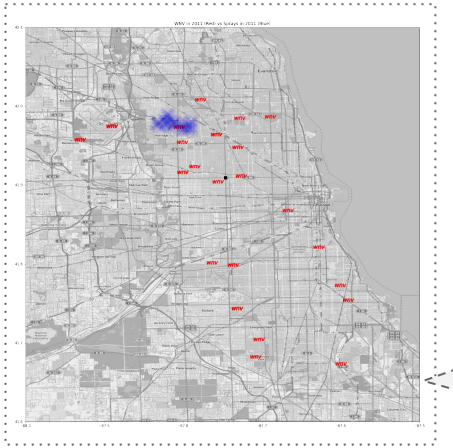
2011 WNV (Red), Sprays (Blue)

- Only one spray area although a number of WNV locations were identified



Exploratory Data Analysis: Maps 2011

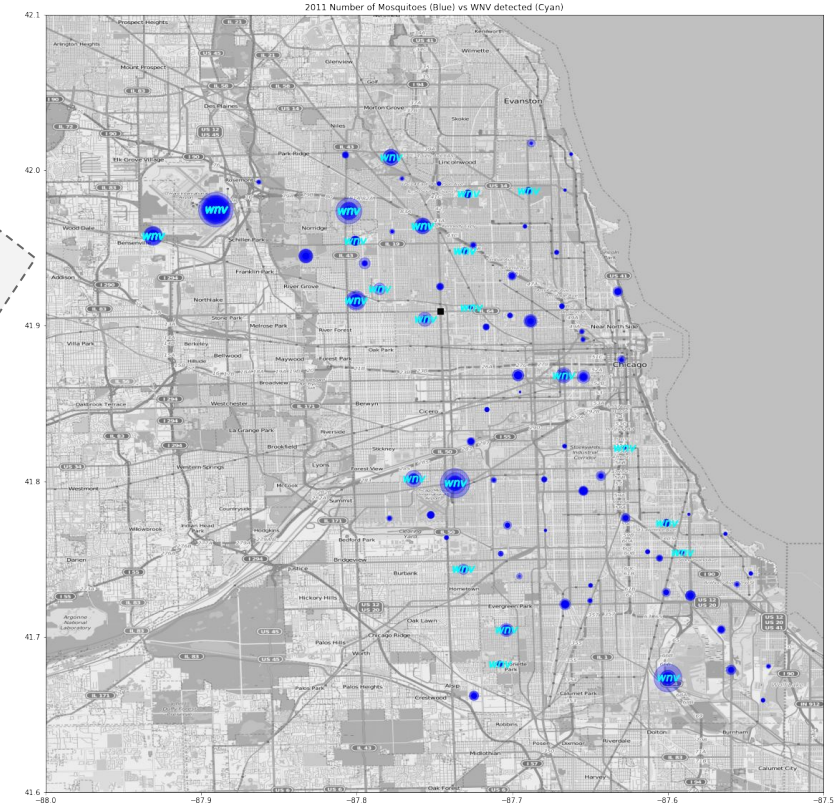




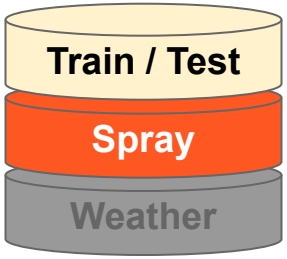
2011 WNV (Cyan), Mozzies (Blue)

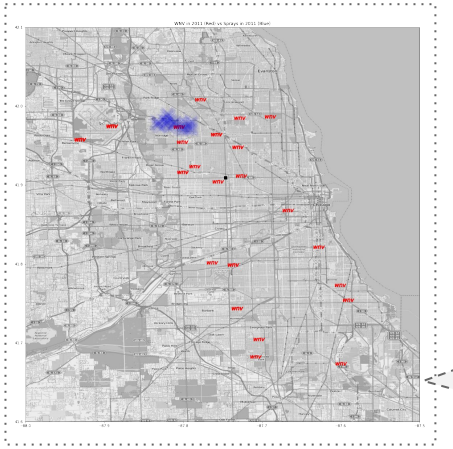
- Many mosquito vectors but not that many with WNV in 2011

2011 WNV (Red), Sprays (Blue)



Exploratory Data Analysis: Maps 2013

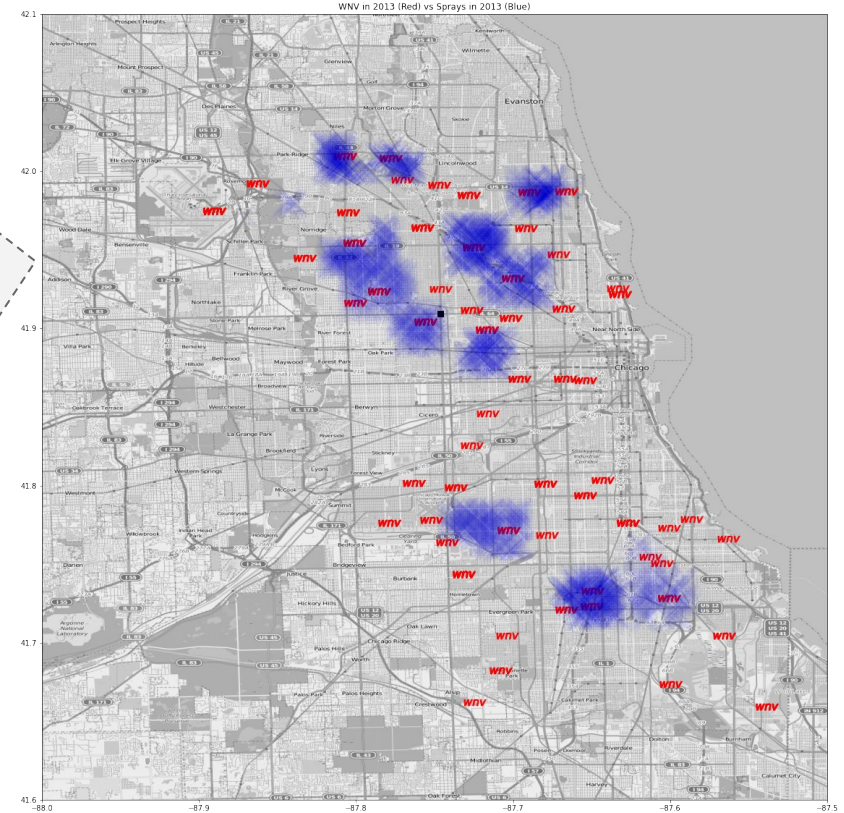




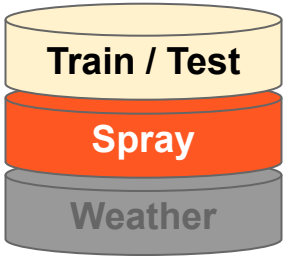
2013 WNV (Red), Sprays (Blue)

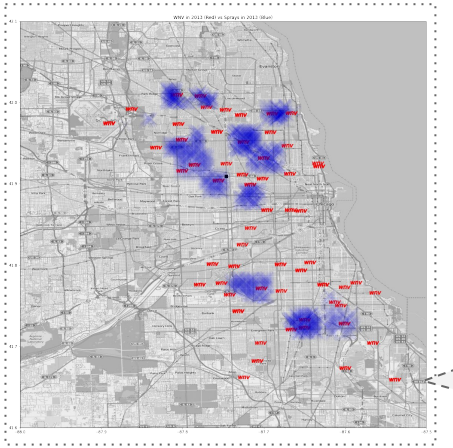
- More spray areas than in 2011

2011 WNV (Red), Sprays (Blue)



Exploratory Data Analysis: Maps 2013

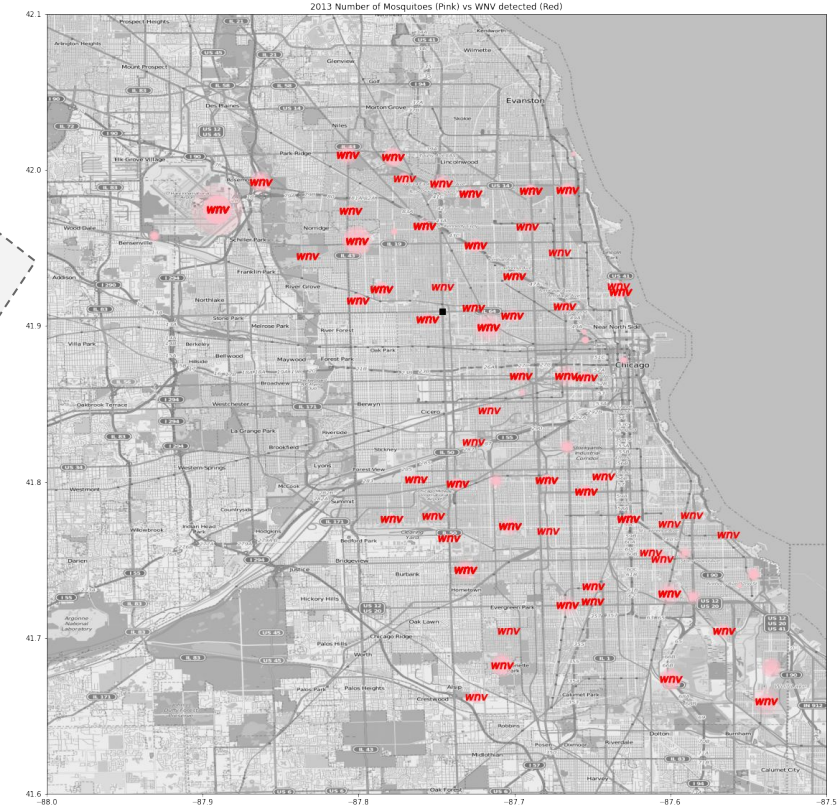




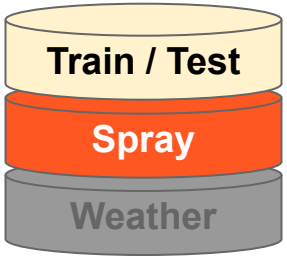
2013 WNV (Red), Moz (Pink)

- Most mosquito clusters have WNV now

2013 WNV (Red), Sprays (Blue)

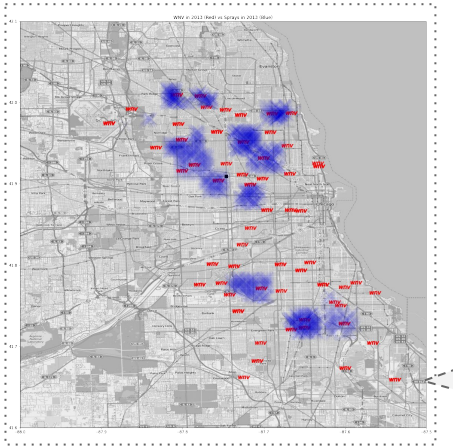


Exploratory Data Analysis: Maps 2013 (GIF)

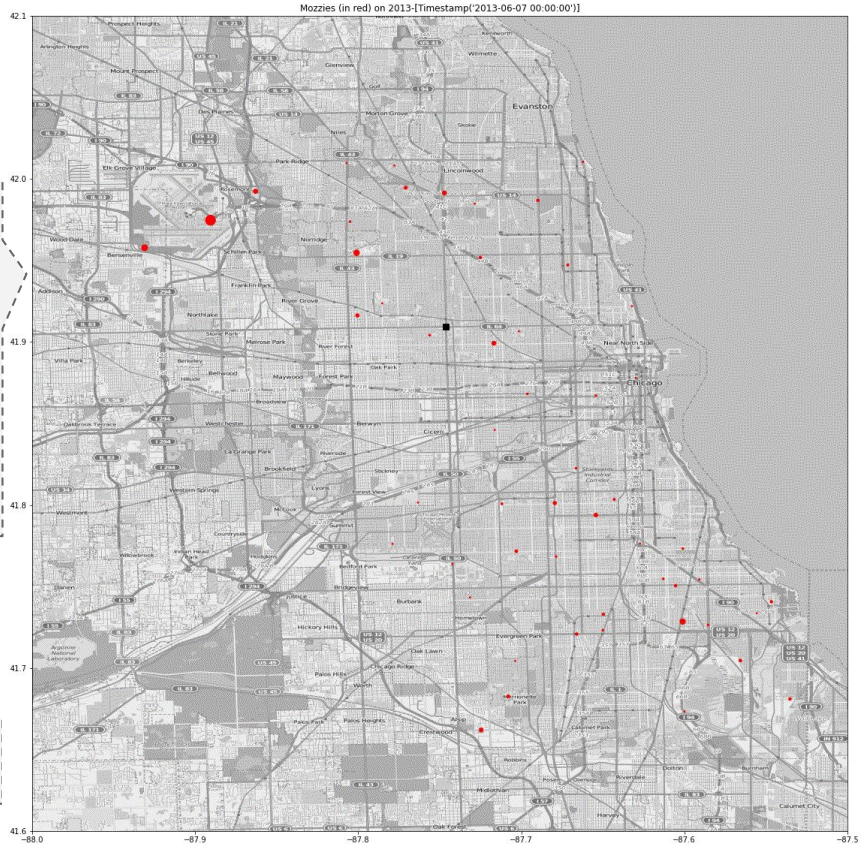


2013 Mozzies Over Time (Red)

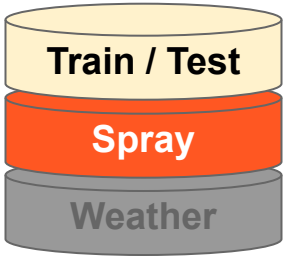
- A few large inland clusters develop over time
- Cluster sizes generally peak in July to August
- Population wanes towards end September (fall / winter)



2013 WNV (Red), Sprays (Blue)

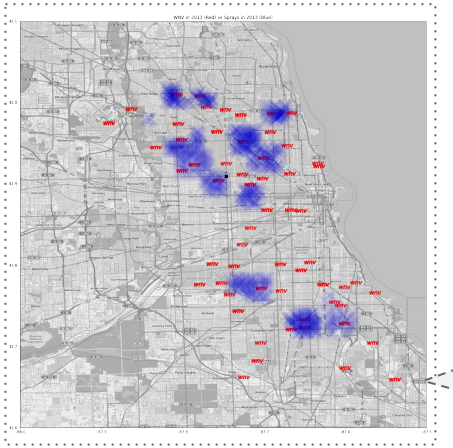


Exploratory Data Analysis: Maps 2013 (GIF)

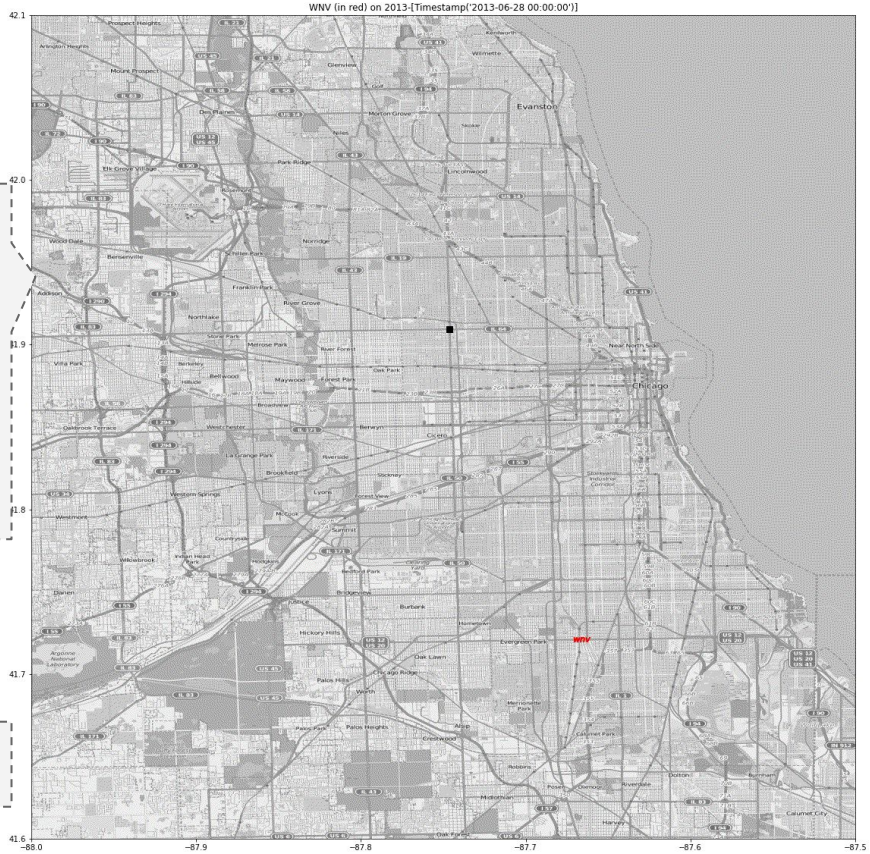


2013 WNV Over Time (Red)

- Incidence of WNV in mosquito clusters peaks in August and starts to wane towards late September



2013 WNV (Red), Sprays (Blue)



Building Models: From baseline to production

1. Baseline

Predict 0 for every row in the test Kaggle set. It gets us accuracy of 0.95, AUC-ROC of 0.50 and recall of 0. This is what our model has to beat.

2. Modelling with base features.

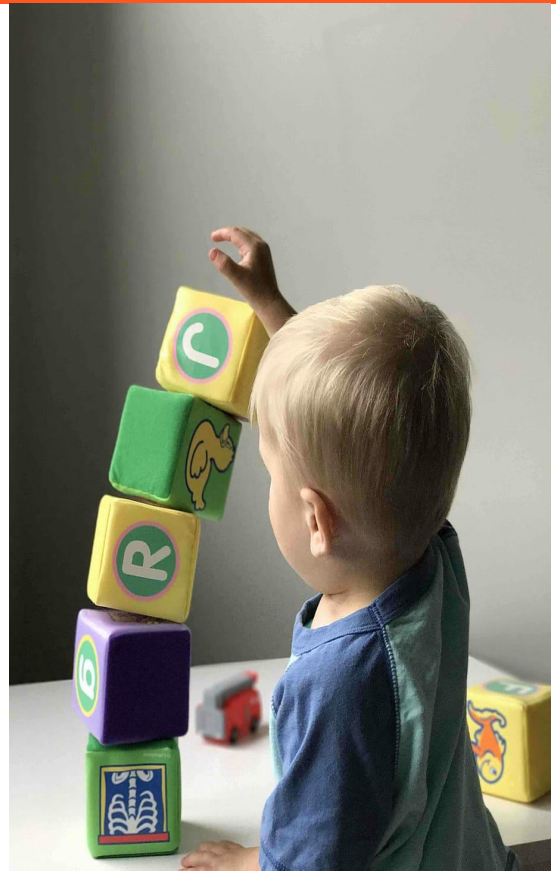
We call it 'reg' models in the notebook. It consists of applying simple models to engineered data.

3. Polynomial features and PCA

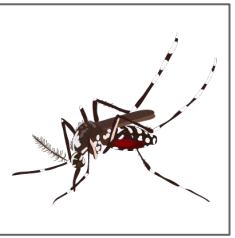
Expand the number of columns with polynomial features, then used PCA to reduce them to only essential ones (98%+ explained variance)

4. Add SMOTE and SMOTETOMEK.

Addresses the class imbalance and tries to improve recall scores which, without SMOTEing, are abysmally low.



Kaggle scoring: opportunistic submissions



VS

- Train datasets has only 8k rows, while the test one has 116k
- Avail ourselves of the test dataset for scoring purposes as much as possible

More models below!

Overview Data Code Discussion Leaderboard Rules Team — My Submissions — Late Submission

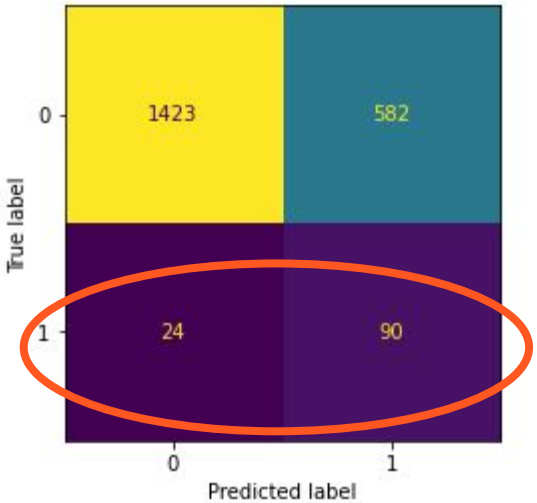
All Successful Selected

Submission and Description	Private Score	Public Score	Use for Final Score
submit_gbsmotetomek2.csv a day ago by harry dzeba gb, smotetomek, 3rd try, no stnpressure	0.66830	0.69123	☐
submit_rfpca2.csv 4 days ago by harry dzeba random forest, smotetomek, pca50	0.69289	0.69793	☐
submit_lrpca2.csv 4 days ago by harry dzeba lin reg, pca, smote	0.66091	0.67600	☐
submit_gb_pca2.csv 5 days ago by harry dzeba gr boost, pca, 2nd try	0.64079	0.64769	☐
submit_xgsmotetomek2.csv 5 days ago by harry dzeba xg boost, smotetomek	0.70069	0.70682	☐
rf_smote2.csv 5 days ago by harry dzeba random forest, smote, 2nd try	0.68911	0.71579	☐
submit_gbsmotetomek2.csv 5 days ago by harry dzeba gradient boost, smotetomek, 2nd try	0.67969	0.69984	☐
submit_gb_smote2.csv 5 days ago by harry dzeba gb, smote	0.68243	0.69743	☐
submit_lr_smote2.csv 5 days ago by harry dzeba linreg, smote, 2nd try	0.66465	0.68562	☐
submit_rf2.csv 5 days ago by harry dzeba	0.69493	0.71241	☐

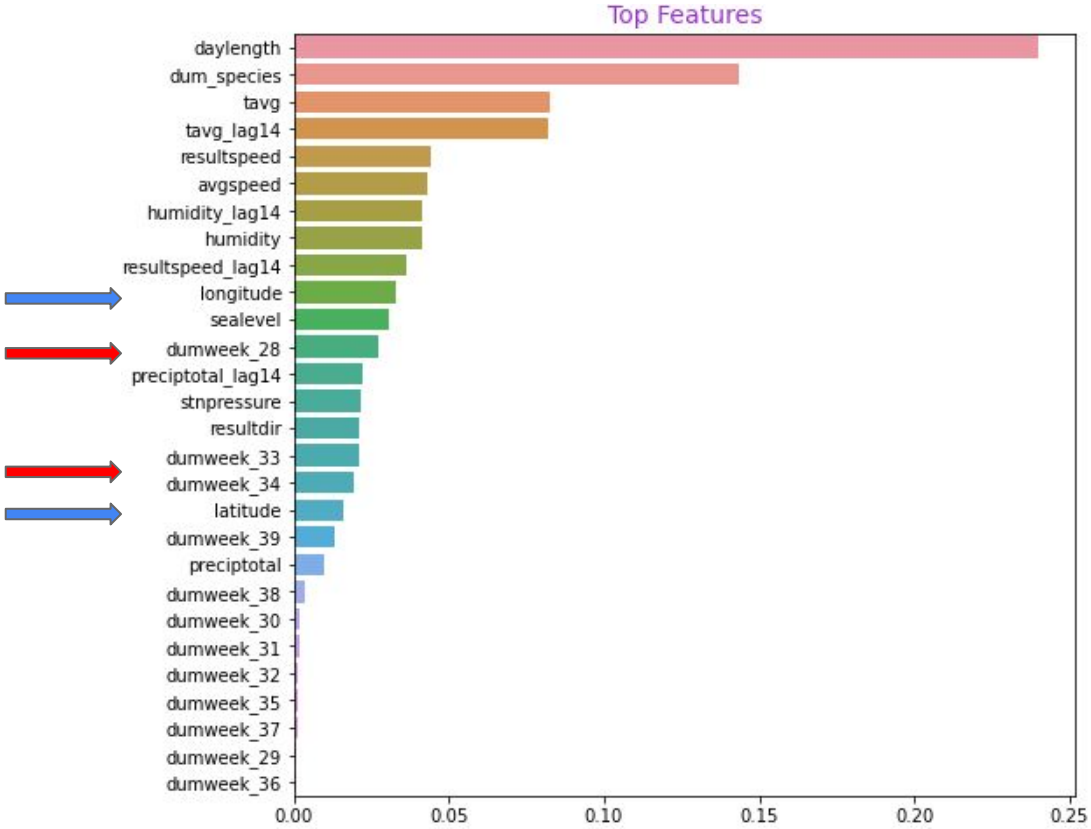
Modelling: The production model

	Model	Type	ROC-AUC	Precision	Recall	F1	Kaggle-AUC
1	RandomForestClassifier(n_jobs=4, random_state=42)	reg	0.855000	0.667000	0.053000	0.098000	0.712000
2	GradientBoostingClassifier(random_state=42)	reg	0.824000	0.280000	0.184000	0.222000	0.718000
3	GradientBoostingClassifier(random_state=42)	pca	0.836000	0.400000	0.123000	0.188000	0.648000
4	LogisticRegression(random_state=42, solver='liblinear')	pcasmote	0.810000	0.133000	0.798000	0.228000	0.676000
5	RandomForestClassifier(random_state=42)	pcasmote	0.829000	0.155000	0.684000	0.252000	0.698000
6	LogisticRegression(random_state=42, solver='liblinear')	smote	0.811000	0.123000	0.842000	0.215000	0.686000
7	GradientBoostingClassifier(random_state=42)	smote	0.831000	0.201000	0.518000	0.290000	0.697000
8	GradientBoostingClassifier(random_state=42)	smotetomek	0.832000	0.180000	0.632000	0.280000	0.700000
9	RandomForestClassifier(n_jobs=4, random_state=42)	smote	0.839000	0.134000	0.789000	0.229000	0.716000
10	XGBClassifier(random_state=42)	smotetomek	0.842000	0.205000	0.535000	0.297000	0.707000

Production model:
Confusion Matrix



Modelling: Feature Importances



- Daylength is the big winner
- Weeks 28, 33, 34 have the highest predictive value
- Longitude/latitude are near the middle/bottom

Cost-Benefit Analysis: Guesstimate Model for Budgeting

Model and Assumptions

1 serious case per annum
50 mild cases per annum

=

110,000 USD
10 weeks productivity loss + hospitalization fees for 1 pax

+

50,000 USD
1 week productivity loss for 50 pax

160,000 USD

÷

ZENIVEX
200 USD / km / week

=

Spray 8 weeks
(~100km^2 per week)

Sources: Centers for Disease Control and Prevention (USA), Various news sources (NBC Chicago, Chicago Tribune, www.chicago.org/mosquitoes)

Conclusions and Next Steps

Our final recommendation is that limited spraying should take place.

- While wasteful to allocate large budget to spraying, limited funds should be allocated to high-risk areas as predicted by our model
- It is important for a city government to show that it cares about its people
- It is important to have a functioning spraying program in place now, which can then easily be ramped up in the future if a sudden need were to arise.

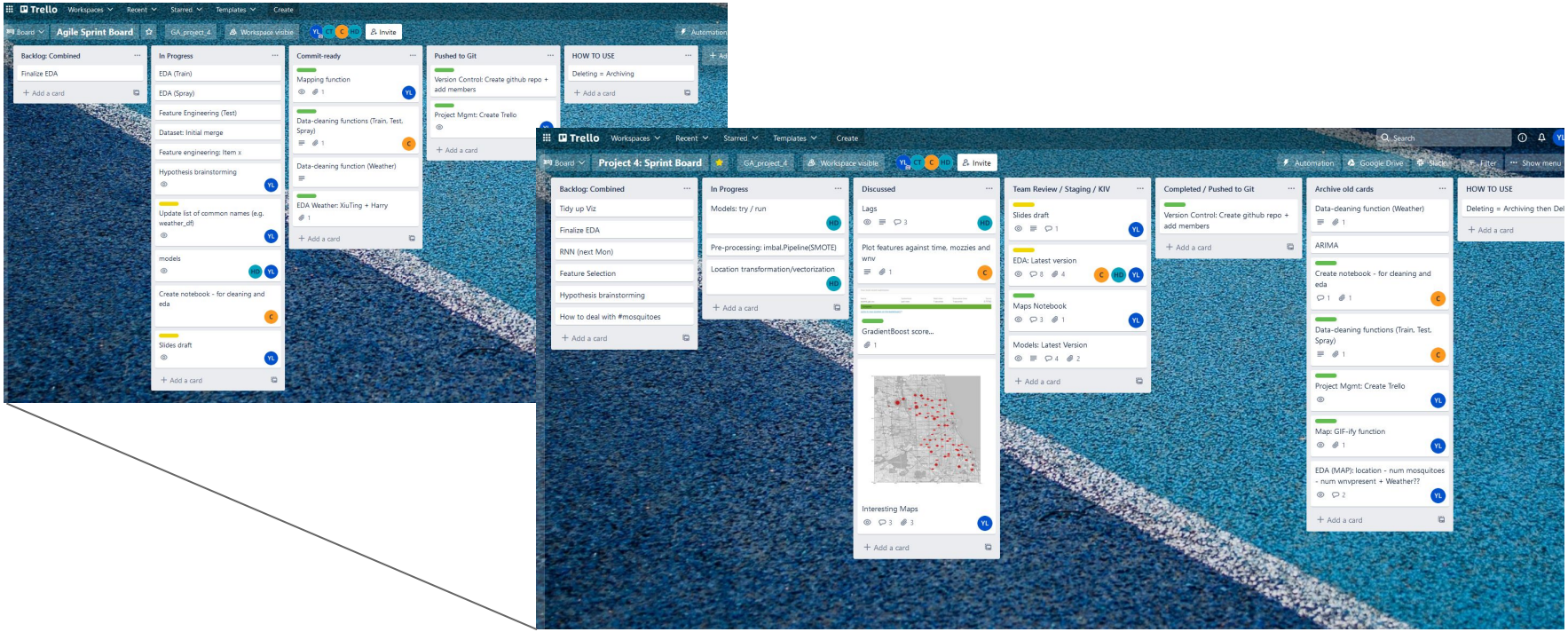
Next steps

- More advanced use of GPS data, as well as teaming up with weather experts and entomologists to model the ways in which mosquitoes move around the city based on atmospheric data and their biological needs.

Thanks for Listening.

Project Takeaways : Project Management and Version Control

Trello for Project Management

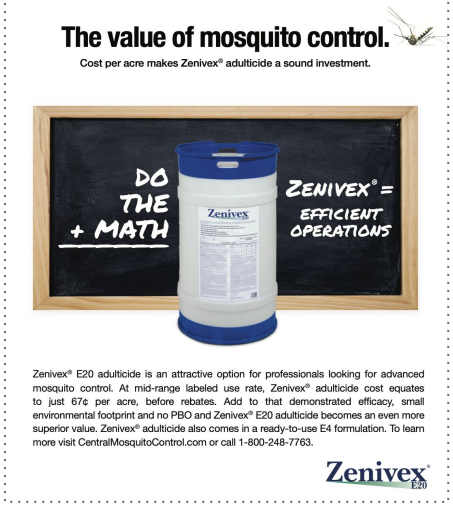


Cost-Benefit Analysis: Guesstimate Model for Budgeting

Model and Assumptions

- **Spray costs (Zenivex):** 67 cents per acre; ~200 USD per km^2 per week
- **Incidence rate:** 225 WNV cases in 2002, 57 WNV cases in 2018
  - Assume ***1 serious*** case and ***50 mild*** cases per year assuming non-interventionist healthcare policy taken
- **Benefit** to society of avoiding infection related costs (via "spray" intervention): 110,000 + 50,000 = 160,000 USD
  - Assume productivity loss of 1,000 USD GDP per capita per week (loosely based on GDP per capita)
  - ***Serious*** case: 10 weeks of productivity loss and hospitalization fees of 100,000 USD
  - ***Mild*** case: 1 week of productivity loss
  - 1 in 150 have serious side effects → 110,000 USD per serious case,
  - 1 in 5 mild → 1,000 USD per mild case
- **Planned Budget = Cost-benefit Breakeven**: 160,000 / 200 = 800 km^2-weeks

Intervention Scope: Spray 8 weeks ~100km^2 per week



Sources: Centers for Disease Control and Prevention (USA), Various news sources (NBC Chicago, Chicago Tribune, www.chicago.org/mosquitoes)

Exploratory Data Analysis: Heatmap for All Features