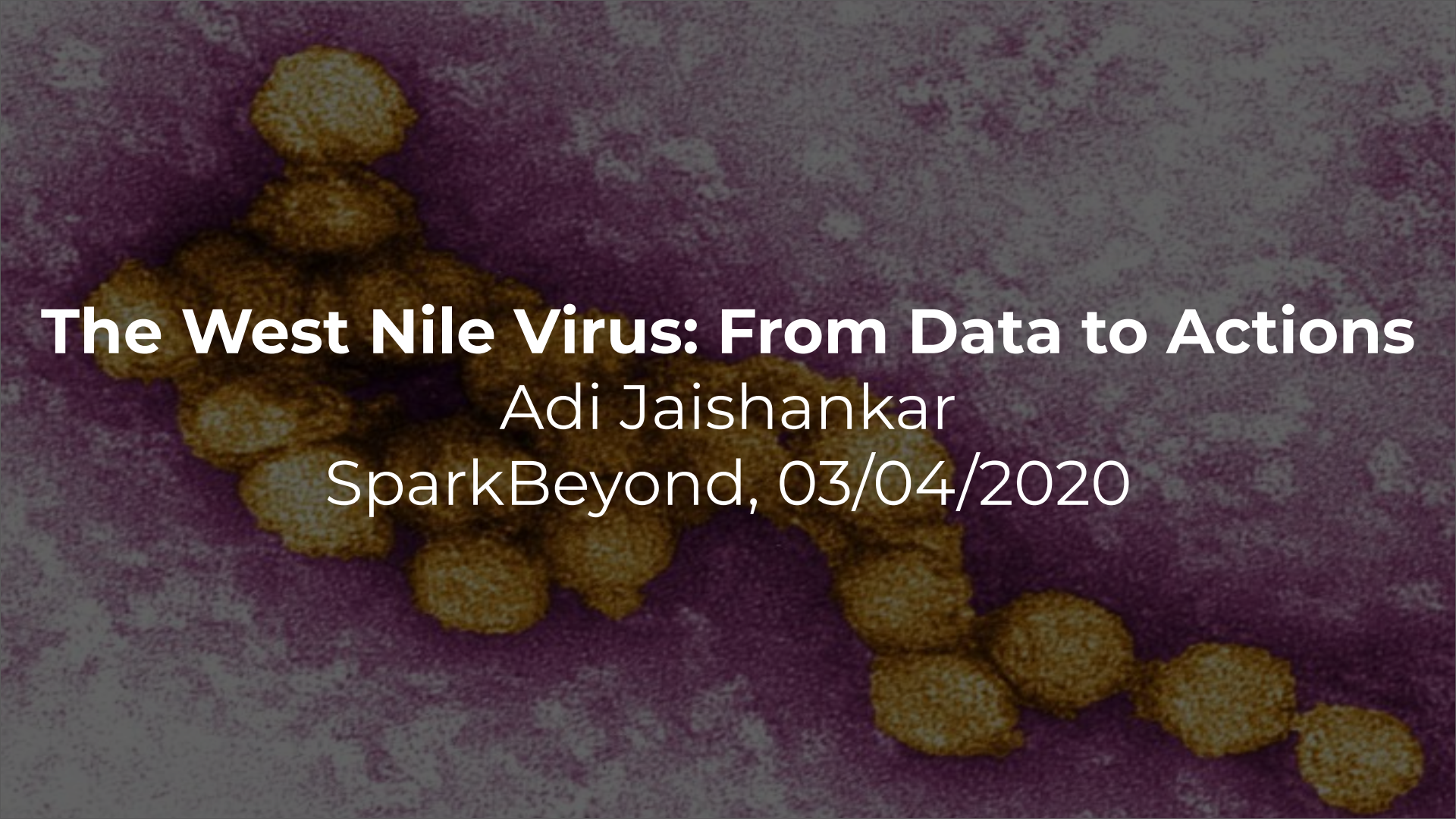


The West Nile Virus: From Data to Actions

Adi Jaishankar
SparkBeyond, 03/04/2020

WNV: leading cause of mosquito-borne disease in the US

Percentage of infected people

reporting mild symptoms

20

PERCENT

presenting serious, sometimes fatal, conditions

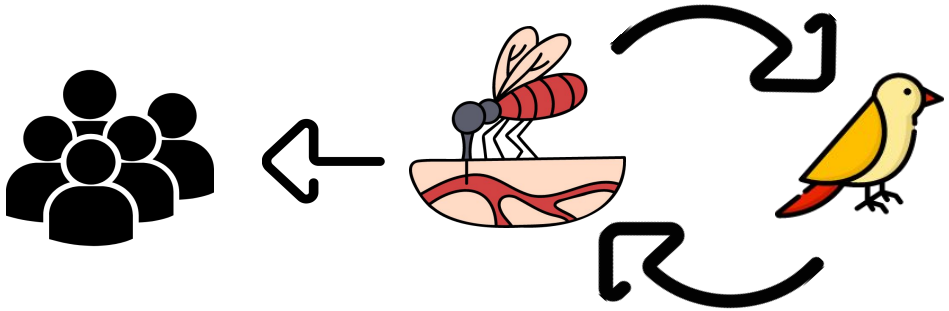
0.67

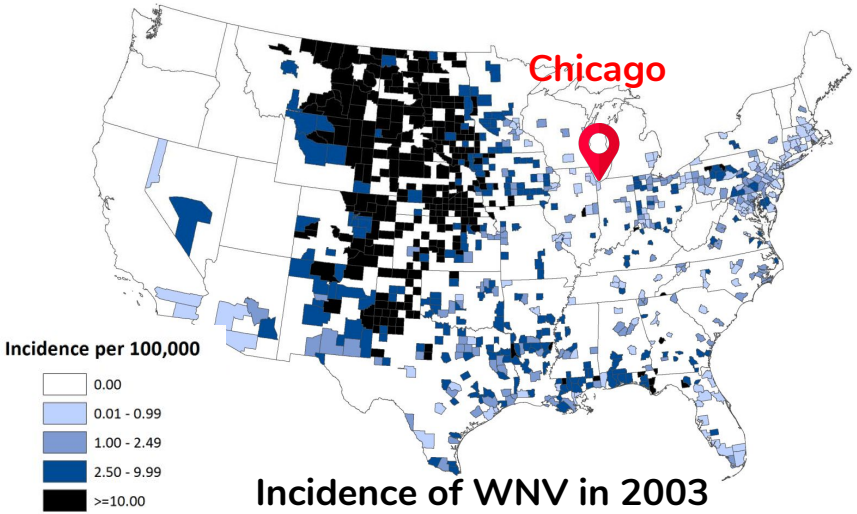
PERCENT

Cost of WNV related hospitalizations since 1999

800

MILLION





Incidence of WNV in 2003

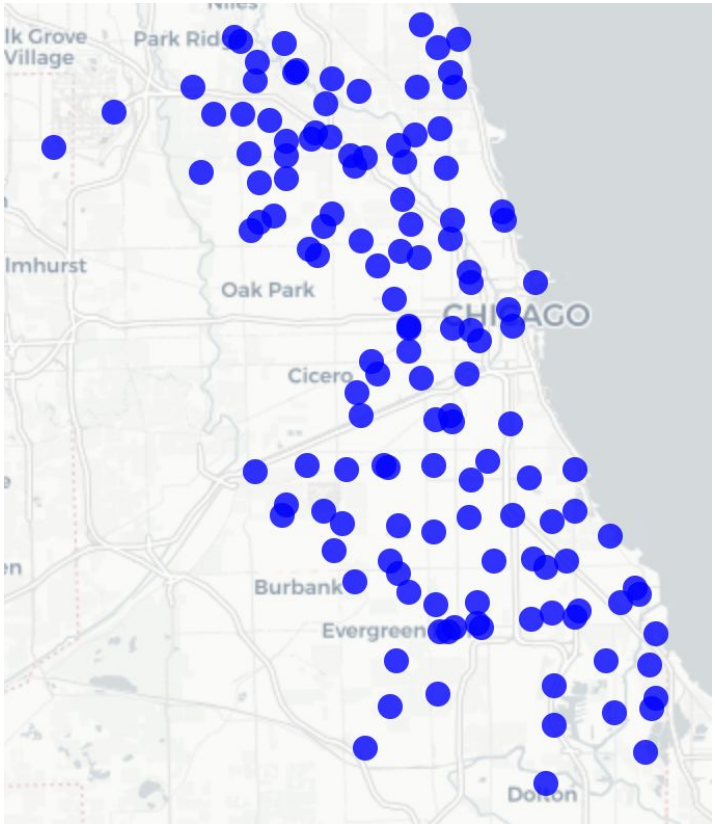
The City of Chicago has set up a comprehensive surveillance program to detect WNV

HEALTHY CHICAGO
CHICAGO DEPARTMENT OF PUBLIC HEALTH

Mosquito Data: Date of testing, location of trap, number of mosquitos in trap, WNV present Boolean

NOAA
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
U.S. DEPARTMENT OF COMMERCE

Detailed Weather Data: Date of testing, location of trap, number of mosquitos in trap, WNV present Boolean

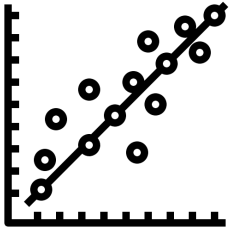


Predictive models are required for effective disease management

Predictive models are required for effective disease management

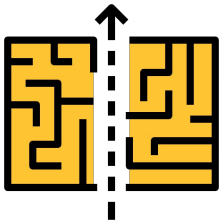
Raw Data

Feature Engineering



Modeling

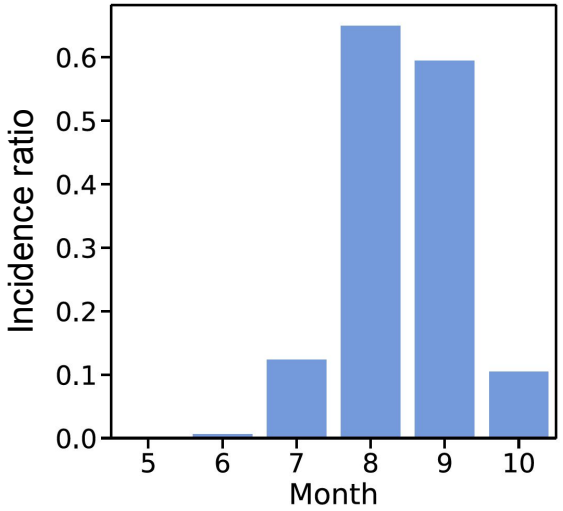
Analysis of Model
(Driving factors)



Recommendations,
Actionable insights

Pitfalls
Additional data

What is the raw WNV testing data telling us: preliminary cleaning and feature engineering



Extracted month and encoded in dummy columns

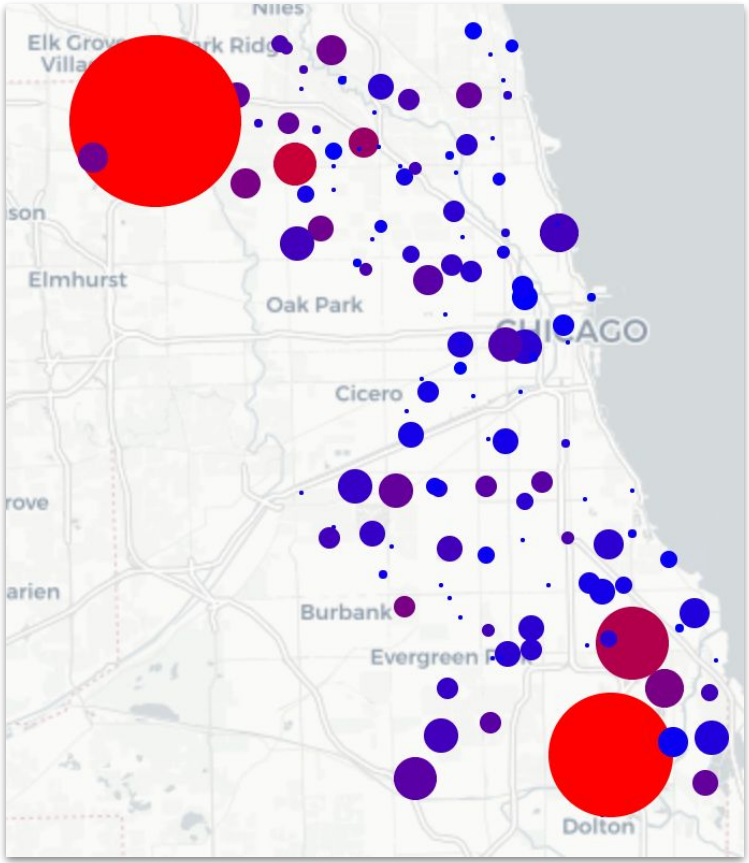
Extracted species and encoded in dummy columns

Replaced multiple lat, lon for same trap with mean

Used only Latitude and Longitude for location

Date	May	June	□	October	Restuans	□	Pipiens	Trap	Lat.	Lon.	NumMosquitos	WnvPresent
2007-10-04	0	0	□	1	1	□	1	T002	41.94	-87.80	29	0

Number of Positives correlated with number of measurements



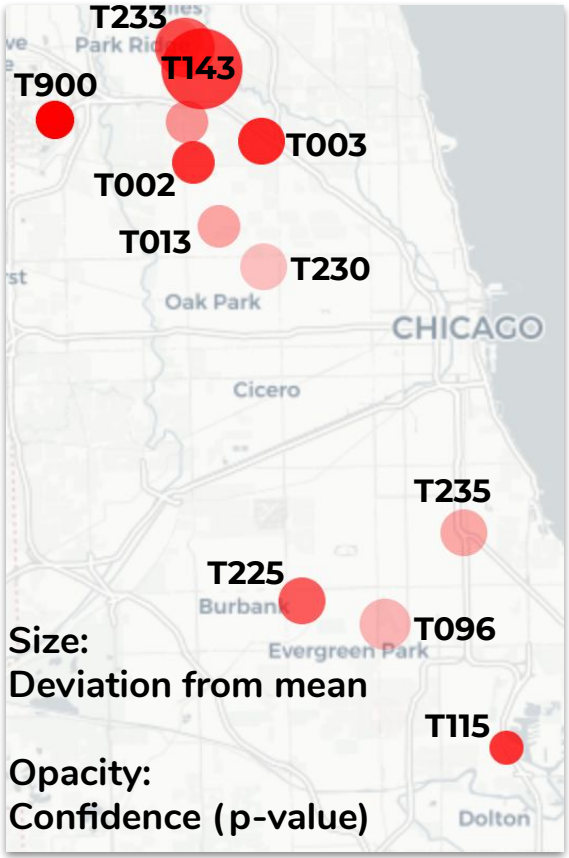
Circle size:
Number of times a trap was sampled

Circle color:
Number of positive tests at a trap location

Correlation coefficient: 0.92

Several traps show statistically significant levels of increased WNV incidence

Data grouped by trap



$$\text{Ratio of Positives} = \frac{\text{Total number of positives}}{\text{Total number of measurements}}$$

$$H_0 : \mu_t \leq \mu_0 \qquad H_1 : \mu_t > \mu_0$$

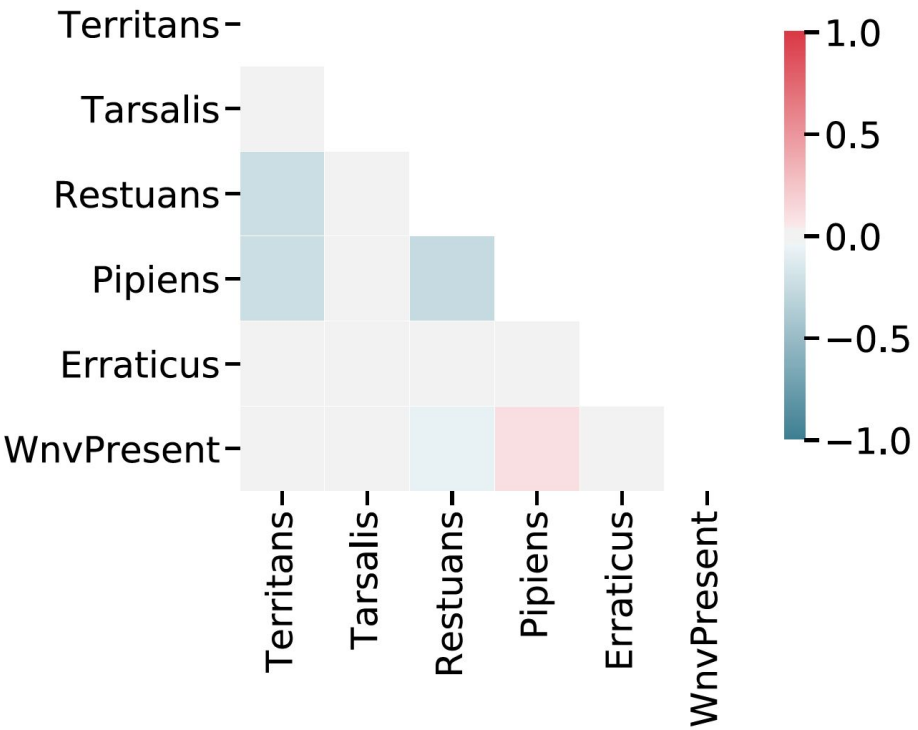
Mean incidence: 5%

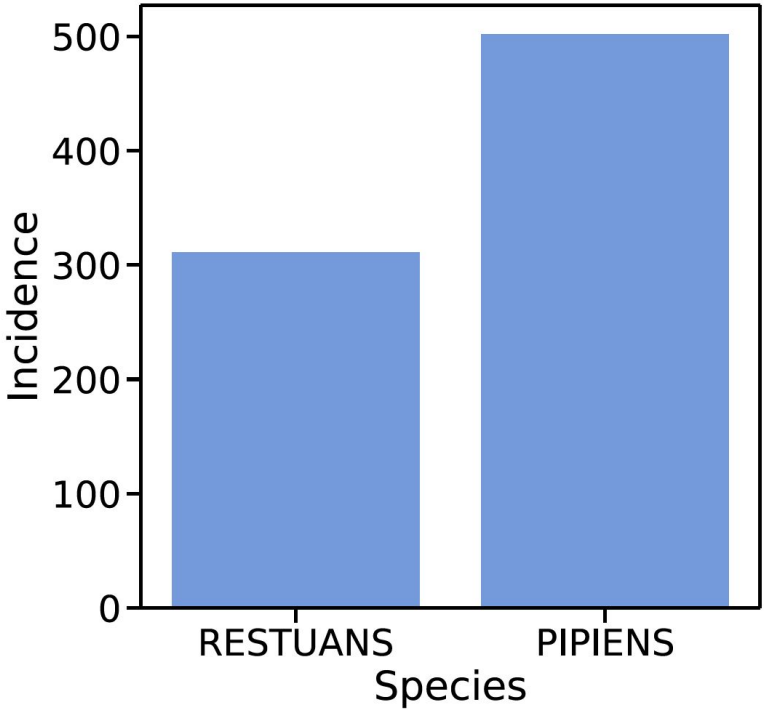
Trap Number	Latitude	Longitude	Incidence
T143	41.999	-87.796	19.4%
T223	42.010	-87.807	14.0%
T096	41.732	-87.678	13.3%
T003	41.964	-87.758	12.0%
T235	41.776	-87.627	11.5%

Action Items:

- Monitor and control high risk traps more effectively
- Collect more data on traps T006, T005, T015, T054C and T070

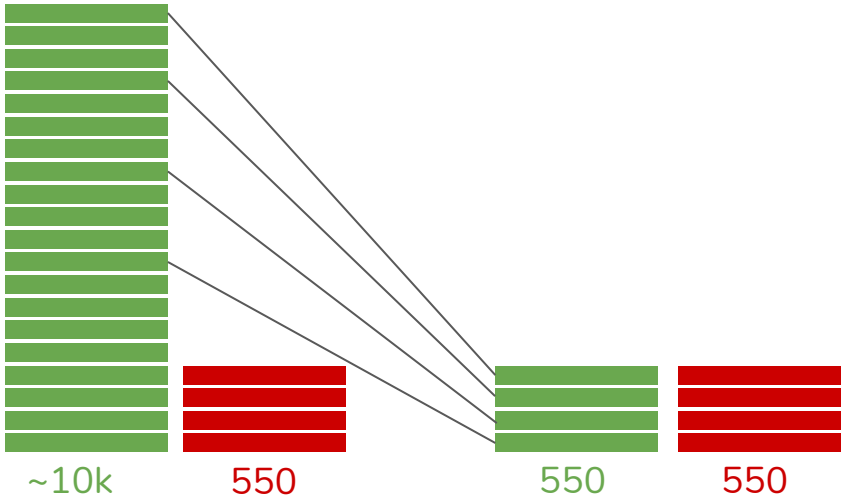
Culex Pipiens drives the presence of WNV almost single-handedly

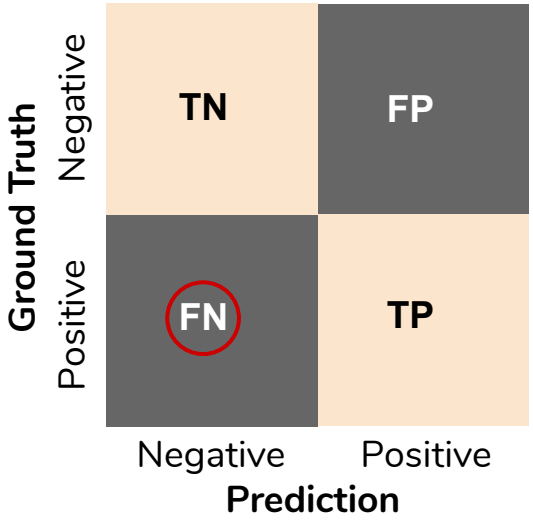




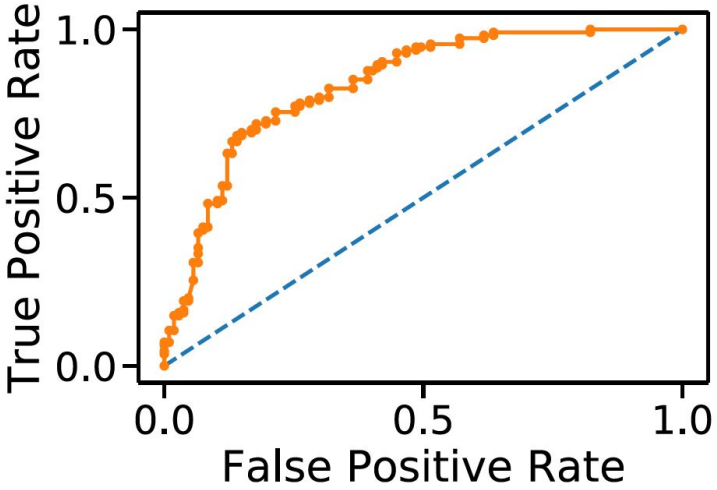
Action Item: Look into specific control measures for culex pipiens (targeted insecticides, nematodes, etc.)

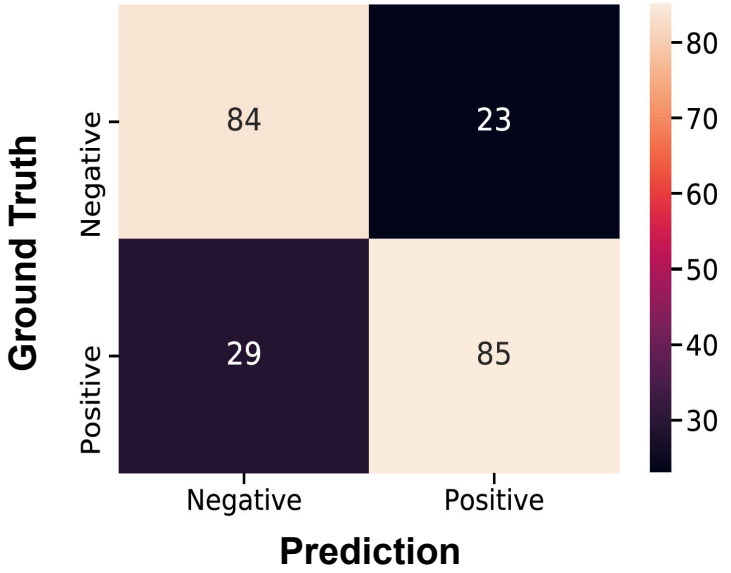
Sub-sampling performed on heavily imbalanced dataset





Baseline Logistic Regression Model performs reasonably well



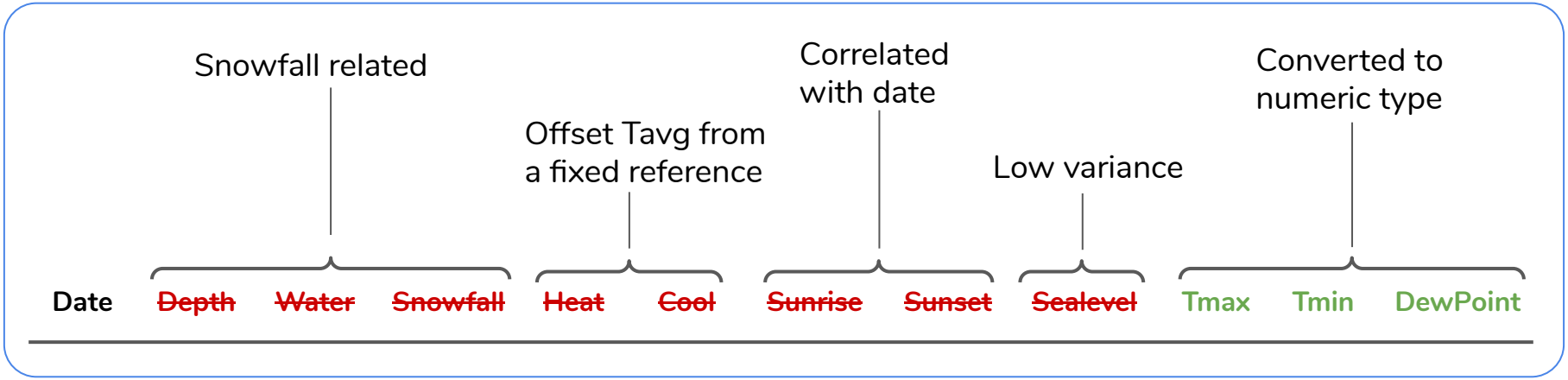


Model Metrics

Accuracy: 0.76
ROC-AUC: 0.84
F1-Score: 0.76


Can we do better?
YES! With weather data

Several data cleaning steps were applied to the weather data

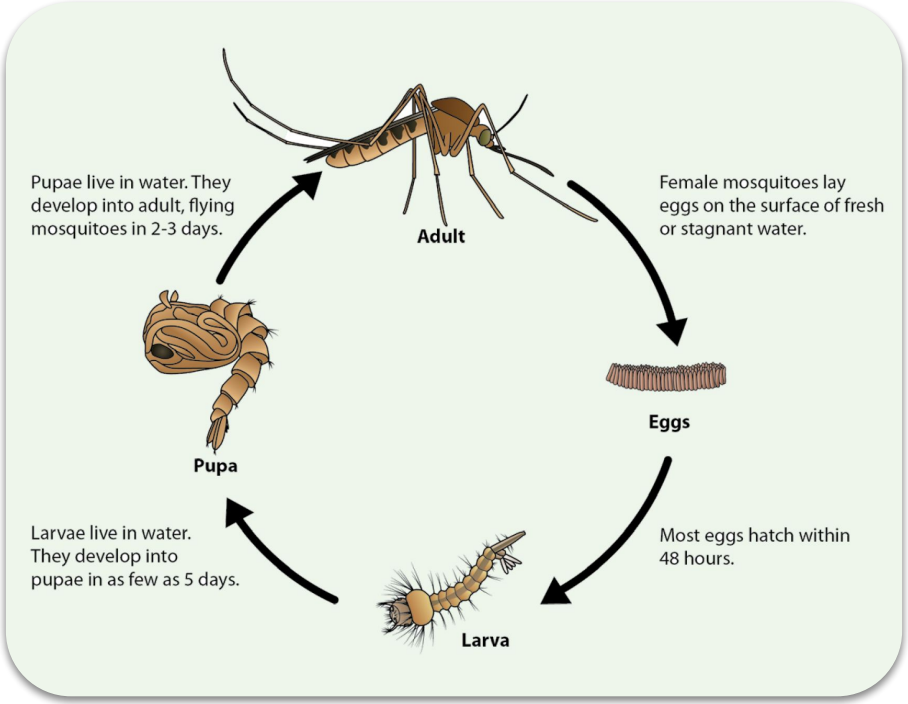


- Missing filled with the persistence model (previous non-empty entry)
- Trace rainfall values replaced with 0.005

CodeSum encodes for significant weather events

Feature engineering of weather data performed based on domain knowledge

Mosquito counts and activity are potentially a lagging indicator of weather



Engineered feature	Description
FG_count FU_count TS_count	The number of days having FG/FU/TS events in the preceding 10 days
Tavg_median DewPoint_median WetBulb_median	The median Tavg/DewPoint/WetBulb temperatures in the preceding 10 days
Total_rainfall_10days	The total precipitation in the preceding 10 days

Calculated absolute humidity from first principles
absolute humidity = f(Tavg, WetBulb, StnPressure)

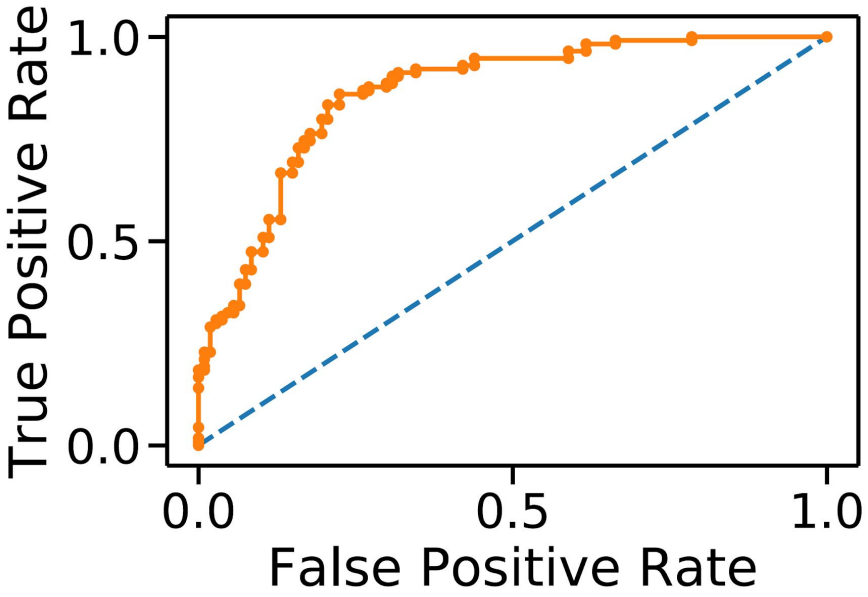
Including engineered features improves model performance

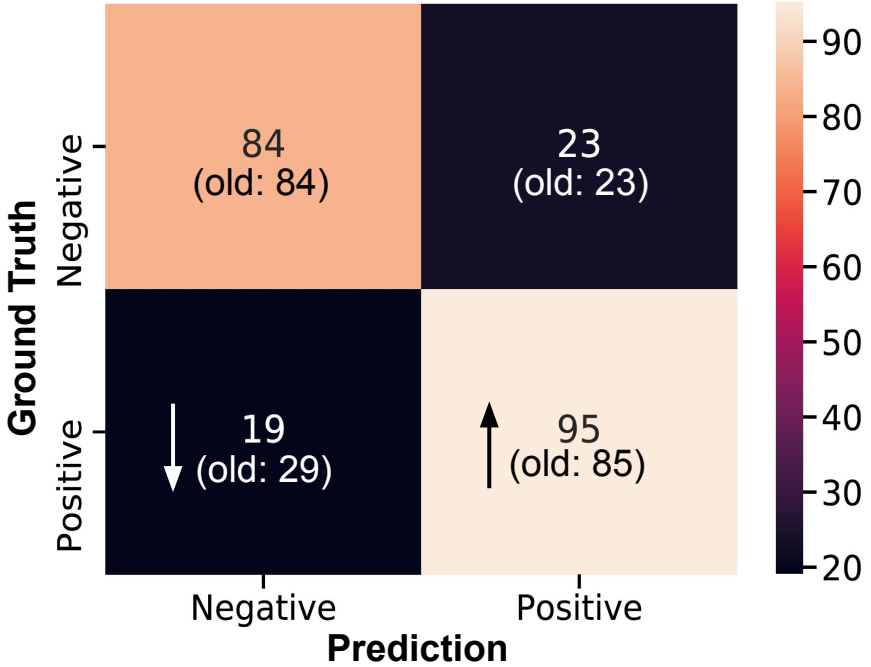
Model Parameters

Model: Logistic regression
Penalty: L1
Validation: 5-fold Cross Validation

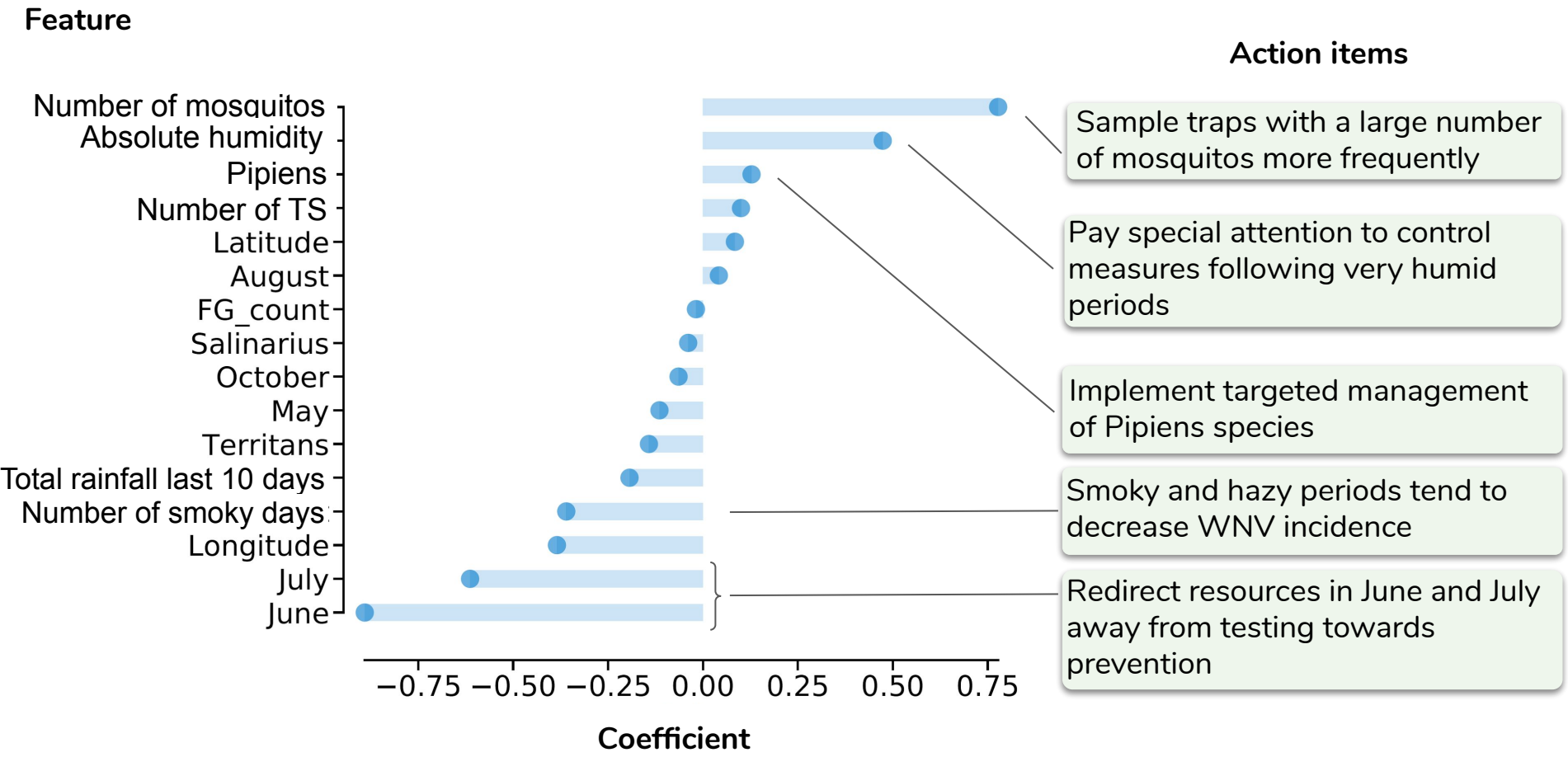
Model Metrics

Accuracy: 0.81 (old: 0.76)
ROC-AUC: 0.86 (old: 0.84)
F1-Score: 0.82 (old: 0.76)

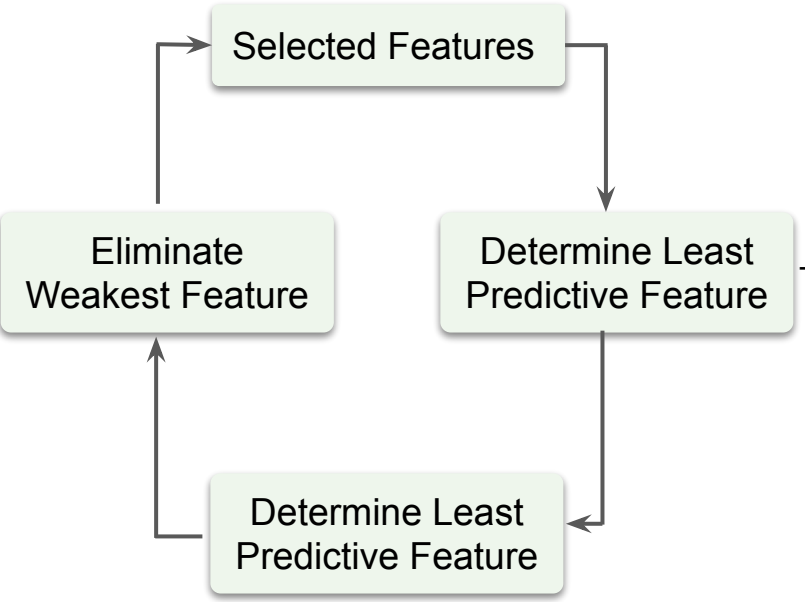




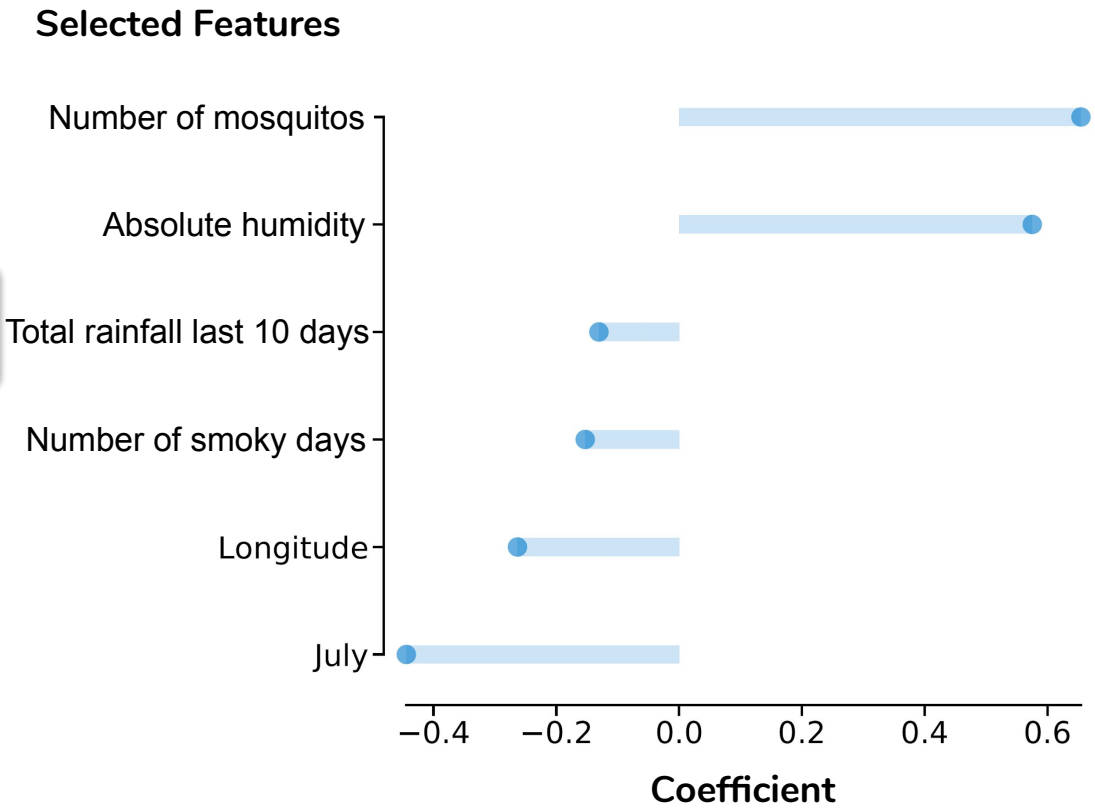
Feature importances help develop action items



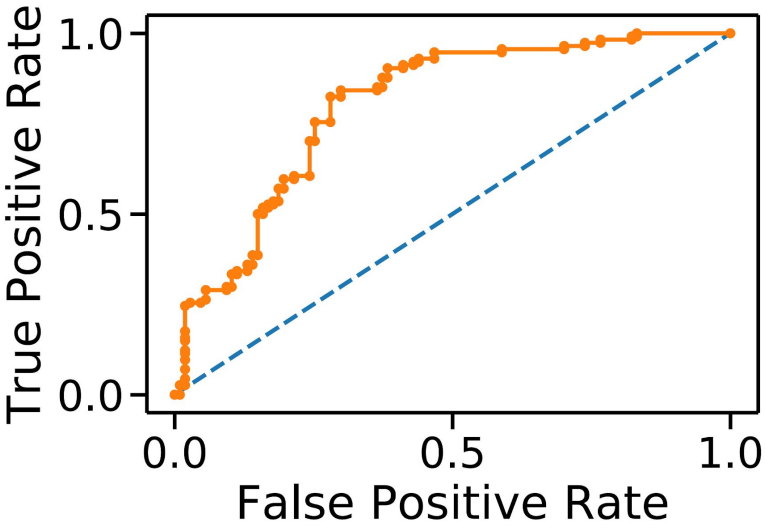
Recursive Feature Elimination helps identify key drivers for improved model interpretability

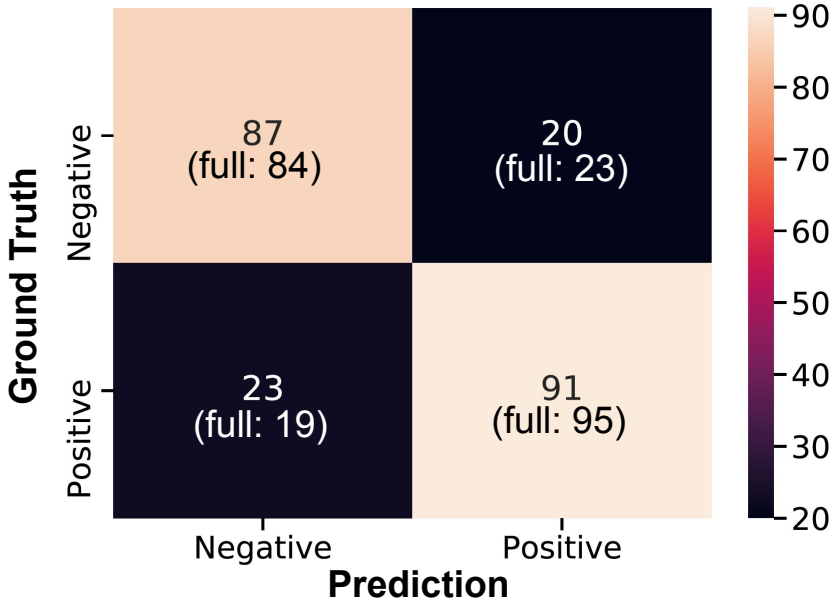


Select best model among M_0, M_1, ..., M_k with k-fold cross validation + penalty



Simplified model performs nearly as well as full model





Model	Accuracy	ROC-AUC	F1-Score
Baseline	0.76	0.84	0.76
Feature Engineered	0.81	0.86	0.82
Reduced feature set	0.80	0.84	0.81

Final recommendations and Action Items

Finding: Traps T143, T223, T096, T003 and T235 show statistically significant increased incidence rate
Actionable: Divert resources to these traps for more effective control (in a balanced manner)

Finding: Traps T006, T005, T015, T054C and T070 show borderline significance of increased risk
Actionable: Collect more data at these traps to reject/fail to reject hypothesis

Finding: Culex Pipiens single-handedly drives WNV incidence
Actionable: Research/implement effective control measures for this species

Finding: Weeks with high humidity, high number of T'showers are followed by a rise in WNV incidence
Actionable: Ramp up mosquito control measures following such periods

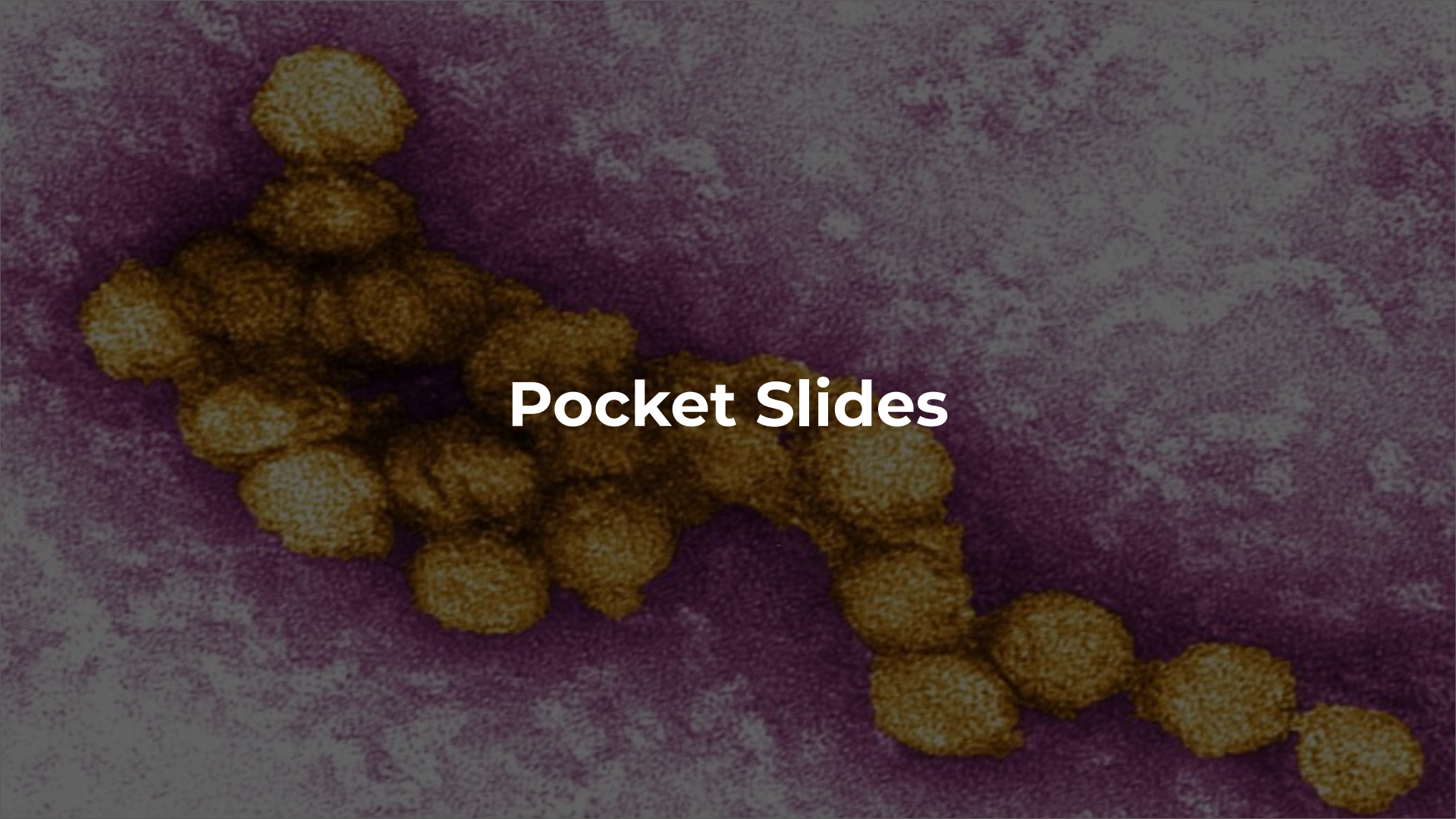
Finding: Weeks of May, June and July show low incidence of WNV
Actionable: If resource strapped, redirect from testing to prevention and control measures

Pitfalls

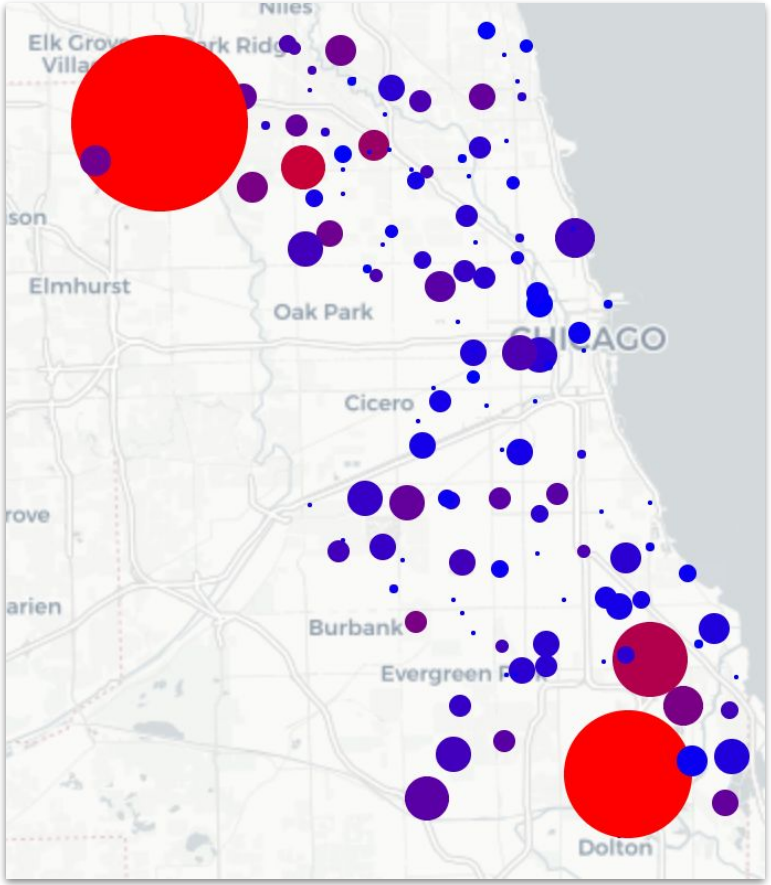
- Did not check for yearly/cyclical variations in mosquito numbers and WNV incidence
- Conclusions on species might be due to limited data
- Bootstrapping techniques might yield better results compared to sub-sampling
- Logistic Regression does not account for highly non-linear effects and feedback loops
- Viruses evolve with time, sometimes very rapidly

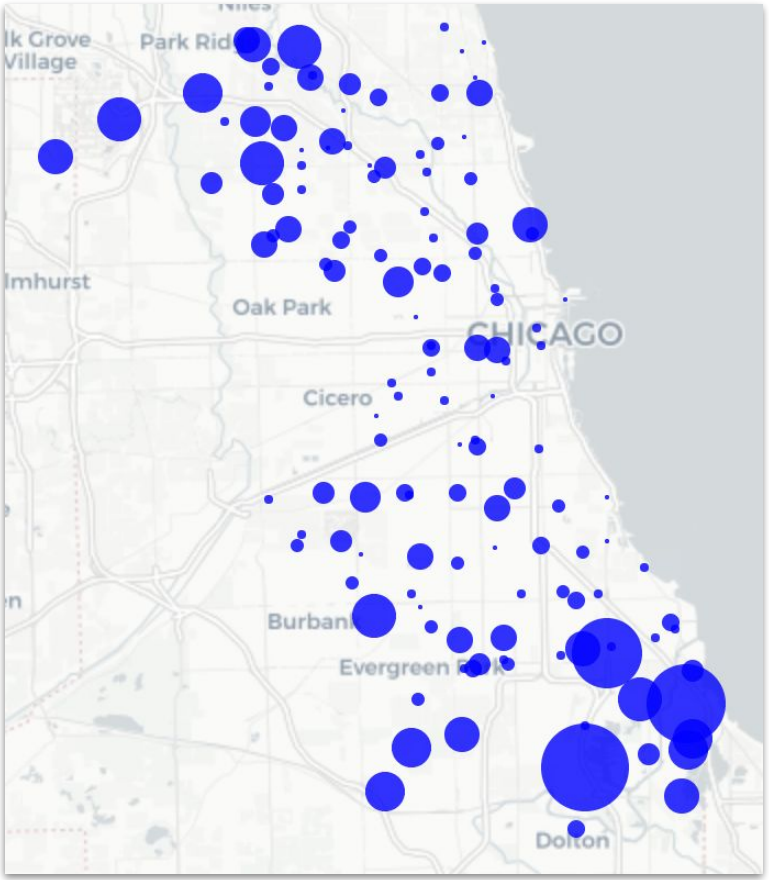
Additional Data

- Bird population, roosting and migration data
- Demographics and health data
- Data on how trap locations were selected

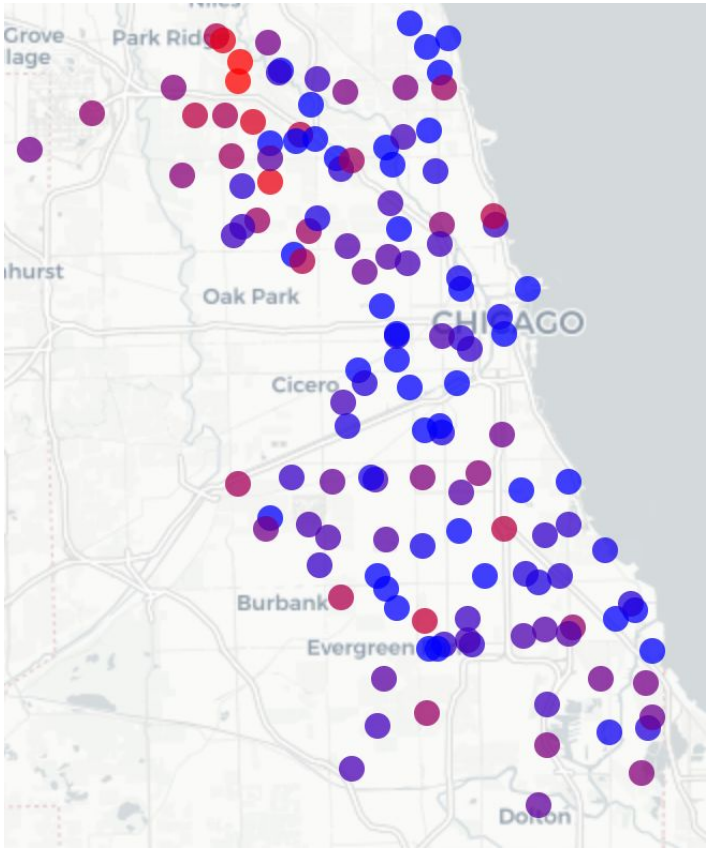

Pocket Slides

Some locations have many more tests than others

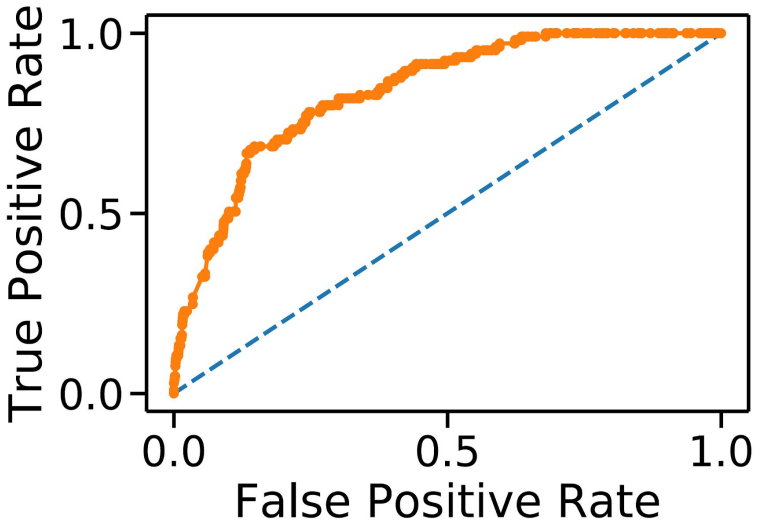


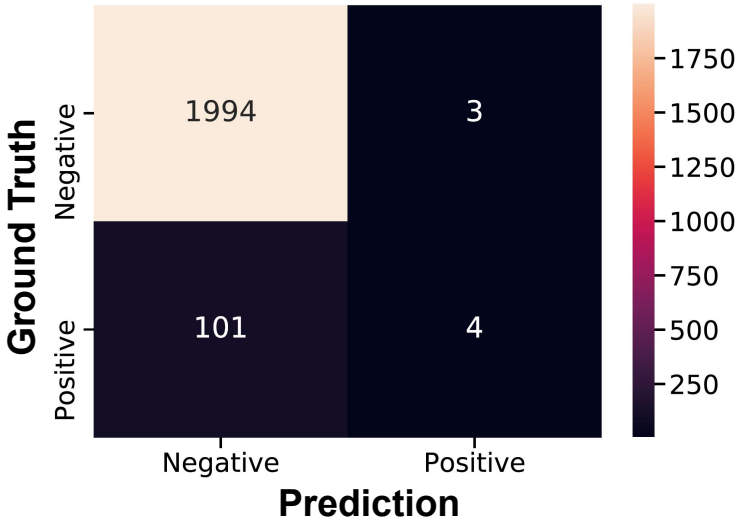


Normalizing by number of measurements reveals higher risk locations



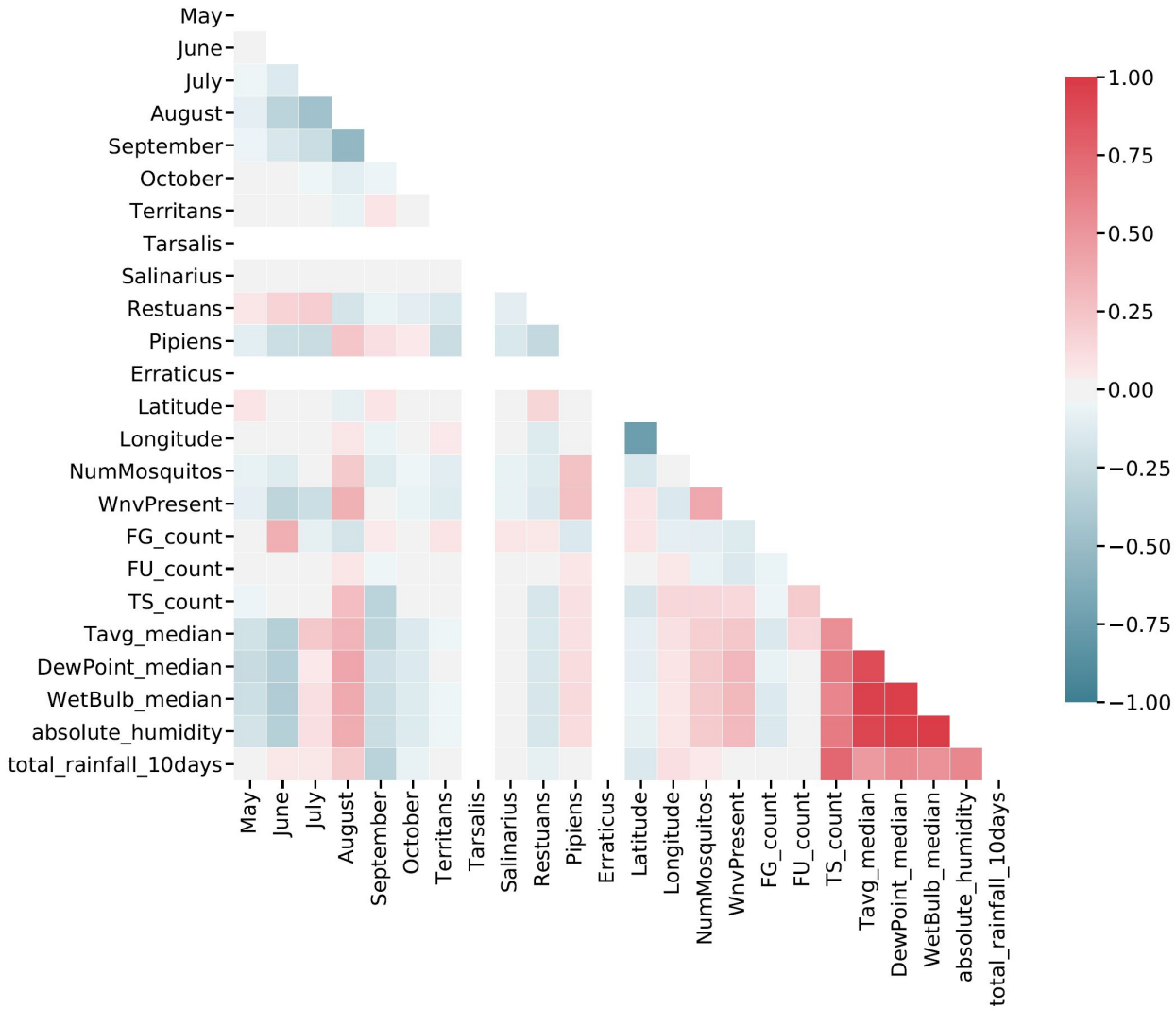
Modeling on imbalanced dataset gives poor results



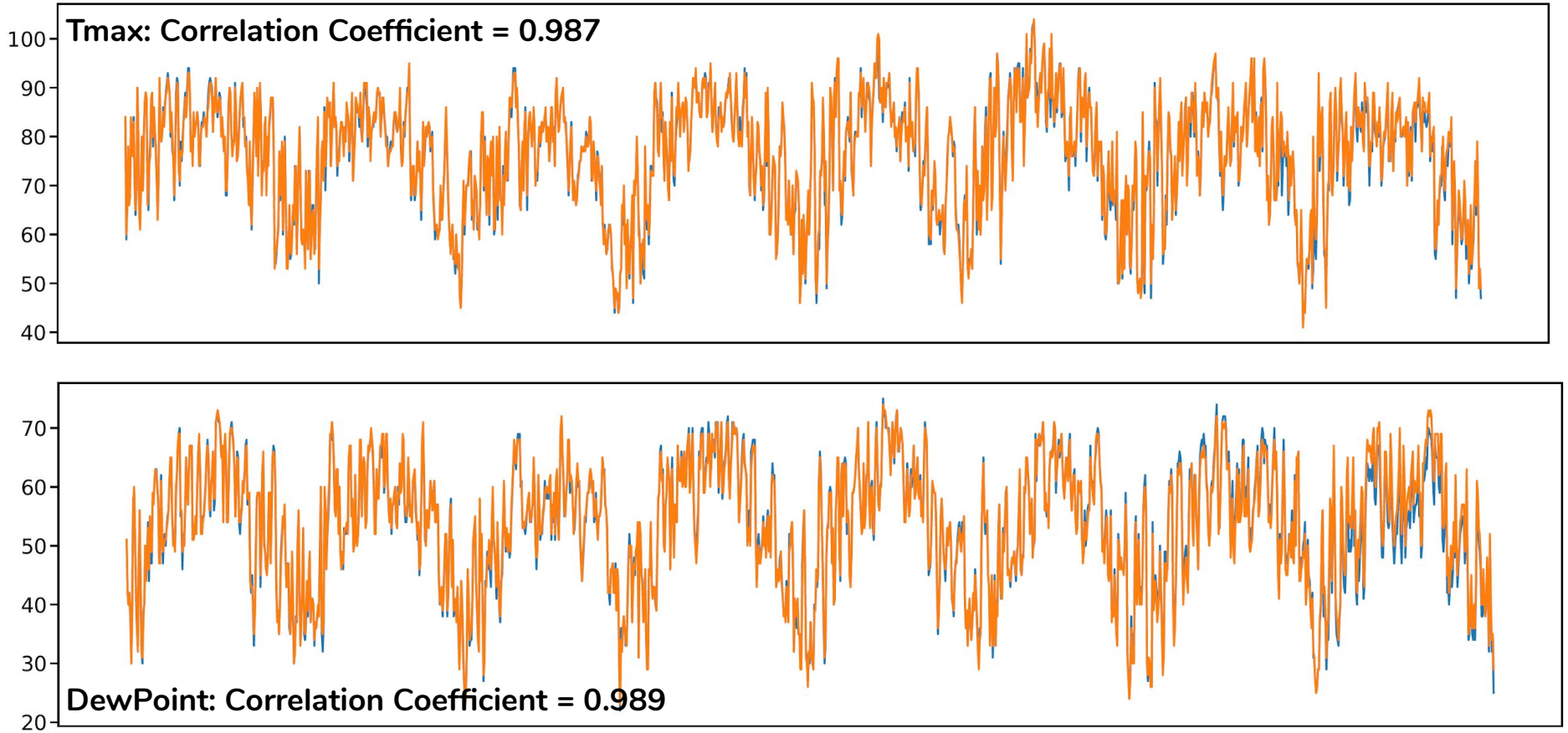


Model Metrics

Accuracy: 0.95
ROC-AUC: 0.84
F1-Score: 0.07

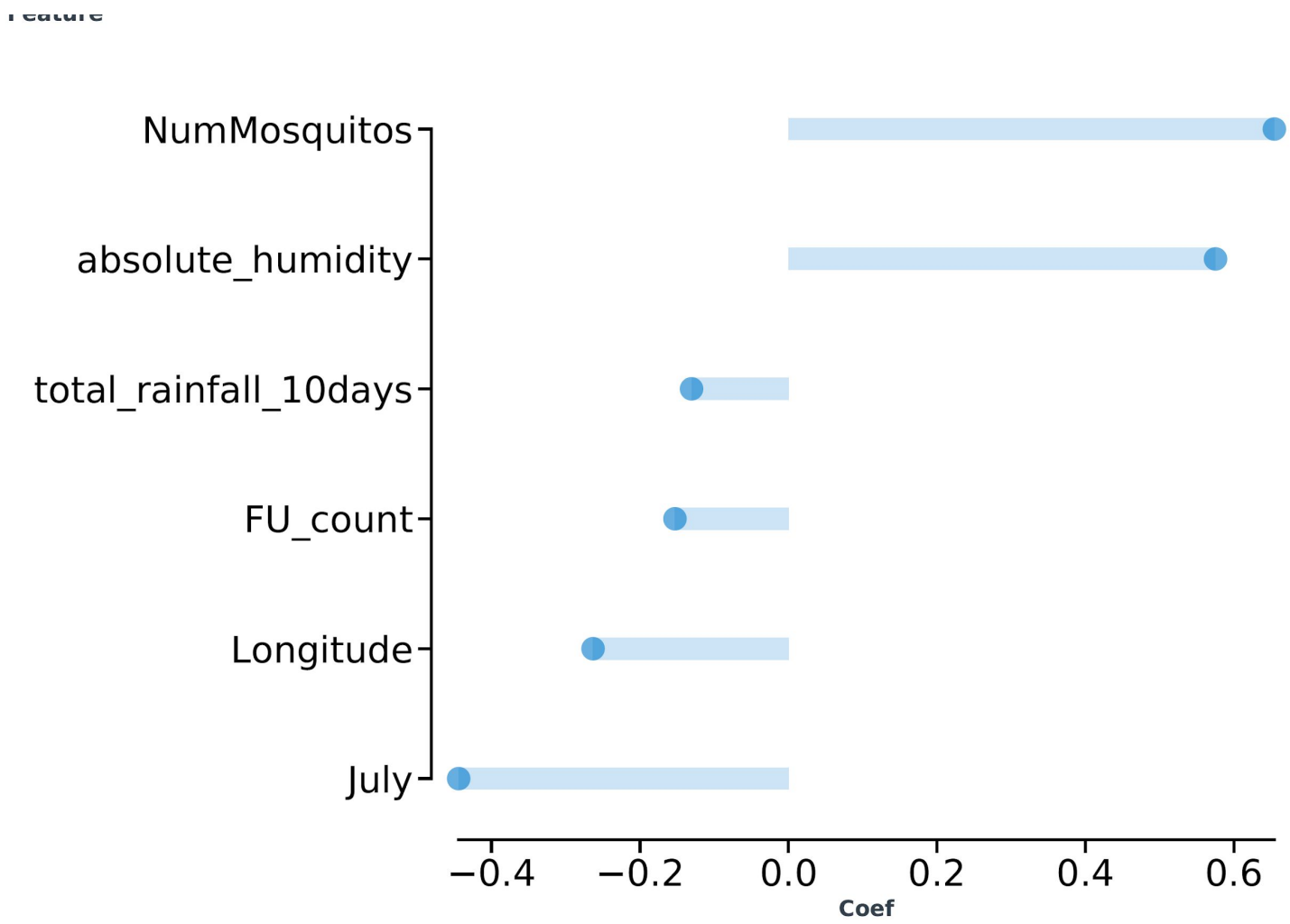

May
June
July
August
September
October
Territans
Tarsalis
Salinarius
Restuans
Pipiens
Erraticus
Latitude
Longitude
NumMosquitos
WnvPresent
FG_count
FU_count
TS_count
Tavg_median
DewPoint_median
WetBulb_median
absolute_humidity
total_rainfall_10days
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00

Station 1 and Station 2 are strongly correlated



Absolute humidity calculation

http://biomet.ucdavis.edu/conversions/HumCon.pdf


Feature
NumMosquitos
absolute_humidity
total_rainfall_10days
FU_count
Longitude
July
−0.4
−0.2
0.0
0.2
0.4
0.6
Coef