Alexis | Alyse | Dylan | Wei Tian

West Nile Virus Prediction

DSI-16 Project 4

Business Problem

What are we trying to solve?

Who are we?

- Data Science Team within the **Disease and Treatment Agency**
- **Focus area:** Chicago

Key Project

West Nile Virus

- West Nile Virus (WNV) has been prevalent in Windy City, Chicago, over the last few years.

Goals

Our Mission

To create a **predictive model** for us to identify the presence of WNV at different locations, time of year and other factors in Windy City.

- Identify problem areas in Windy City ahead of time
- Suggest suitable remedies to prevent the spread of West Nile Virus in the region

Part I
Data Cleaning

Train / Test Dataset

Overview and Processing

Train Dataset

- **12** columns, **10,506** entries
- Samples taken in ***odd*** years: 2007, 2009, 2011, 2013
- Includes date, trap, address, species collected, number of mosquitos and presence of WNV
- Unique samples are split into 'bins' of 50

Test Dataset

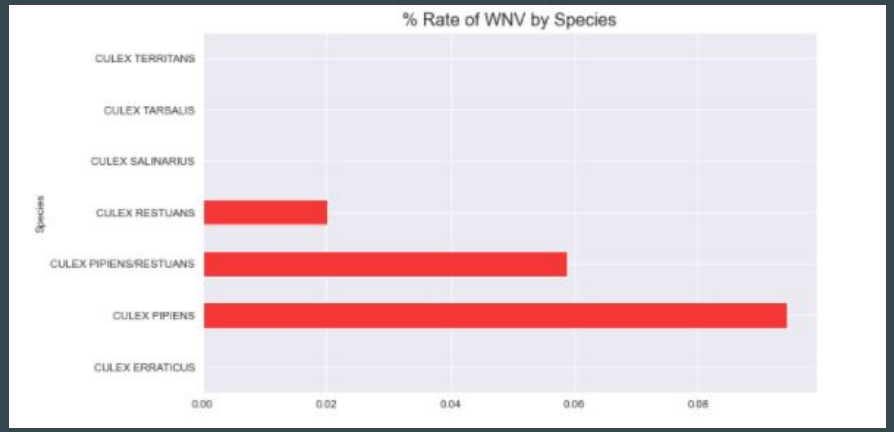
- **11** columns, **116,293** entries
- Samples taken in ***even*** years: 2008, 2010, 2012, 2014
- Same columns as train data except for NumMosquitos and WnvPresent
- Samples are also split in the same way

Processing

- **Streamlining variables:**
  - **Clustering traps** by latitude and longitude and assigning each trap a cluster ID, dropping other address columns
- **Feature engineering:**
  - ***Dates***: Day of year, month, week of year etc.
  - ***MultipleBins***: Track whether the row is part of a larger sample
  - ***One-hot encoding***: For most infectious species and clusters

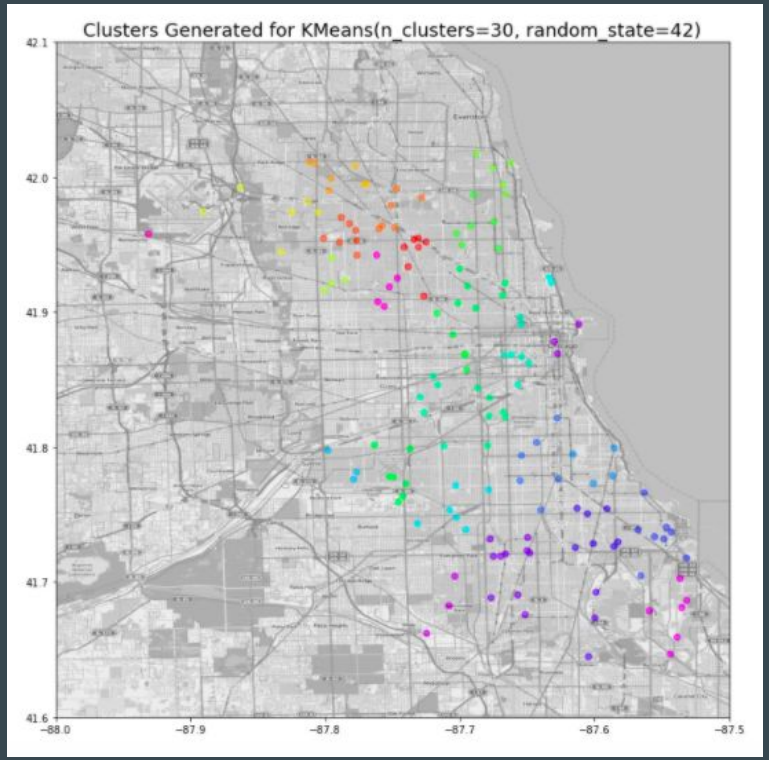
Train / Test Dataset

Feature Engineering



	CULEX PIPIENS	CULEX PIPIENS/RESTUANS	CULEX RESTUANS
0	1	0	0
1	0	1	0
2	0	1	0
3	0	1	0
4	0	1	0

One-hot encoding of **species**



Clustering traps by **latitude and longitude**

Weather Dataset

Overview and Processing

Weather Data from NOAA

- **22** columns, **2,944** entries
- Daily weather metrics from **2** weather stations in **May to October** from 2007 to 2014
- Includes date, station, temperature, sunrise and sunset time, weather conditions

Data Processing

- **Data cleaning:** Imputation of null values by monthly averages,
- **Feature engineering:**
  - Relative humidity based on dewpoint and avgtemp
  - Daylight hours based on sunrise and sunset times

Spray Dataset

Minor cleaning for the Spray dataset

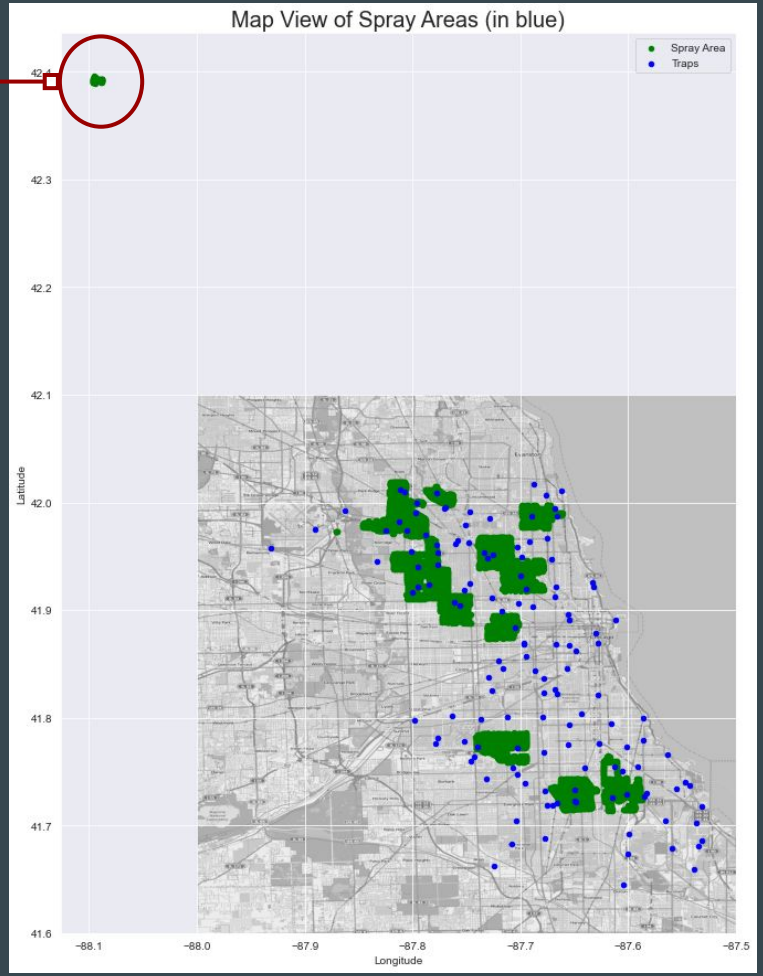
Dropped Duplicates for Spray Data

```
spray[spray.duplicated(keep=False)].shape
```

```
(543, 3)
```

```
# Drop spray duplicates
spray.drop_duplicates(keep='first',inplace=True)
```

Removed anomaly data point in Spray Data

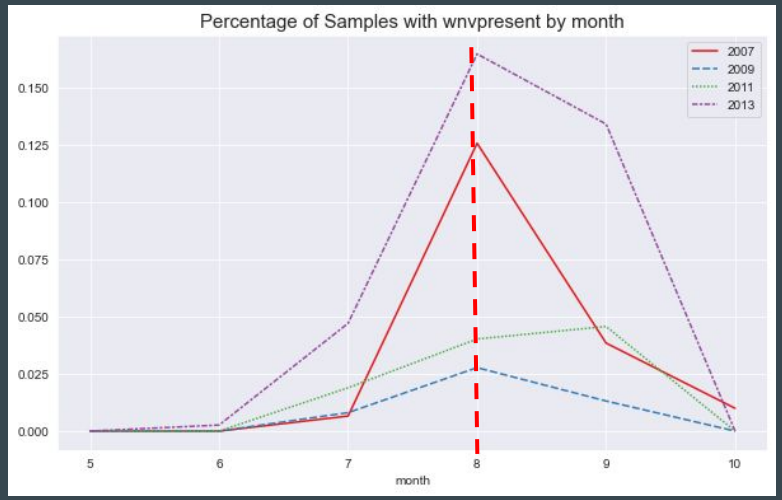


Part II
Exploration of Data

Collected Mosquito Data: Key Findings

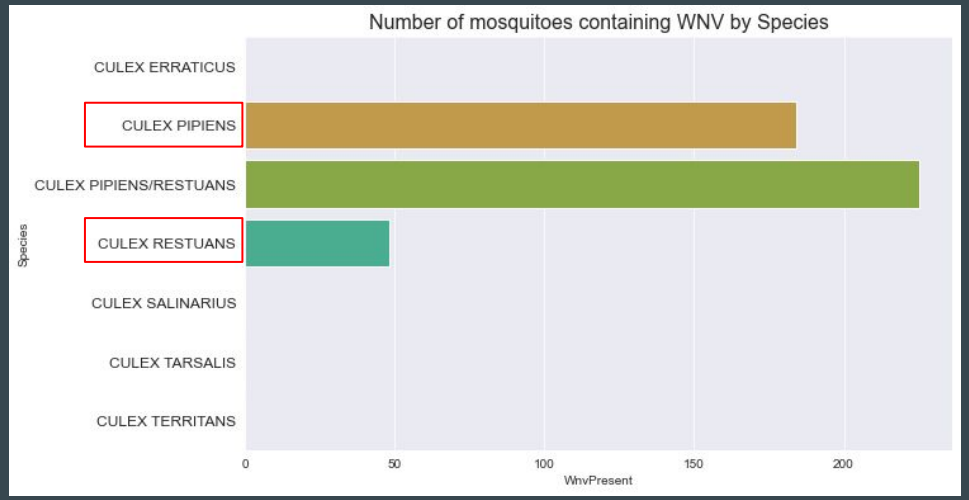
Analysis of WNV presence

Monthly WNV Analysis amongst samples



From the chart above, we observe the presence of WNV amongst mosquitos spike in **August** and subsides thereafter.

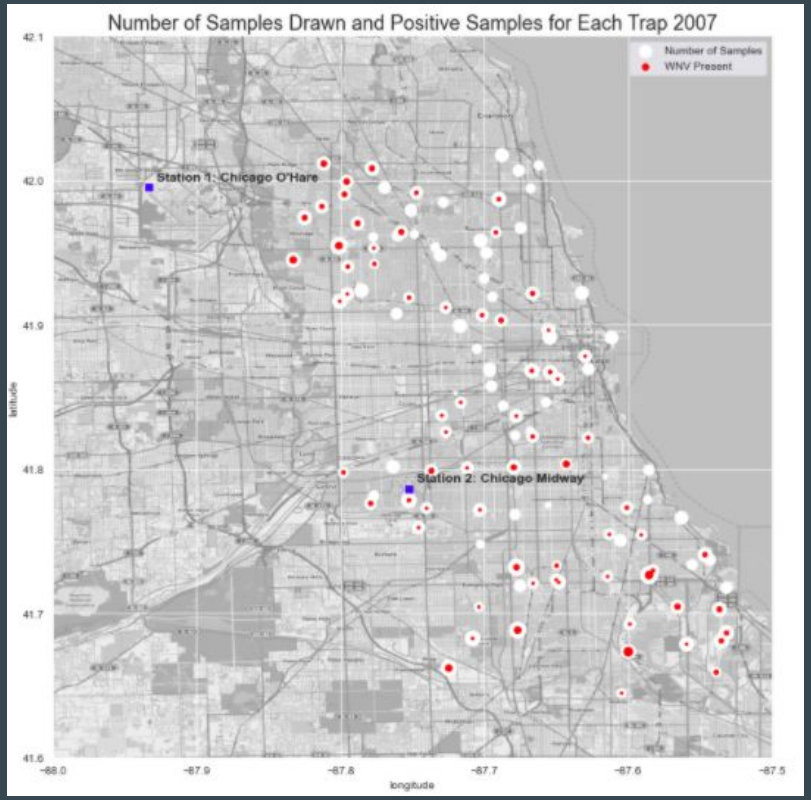
Mosquitoes Species Analysis

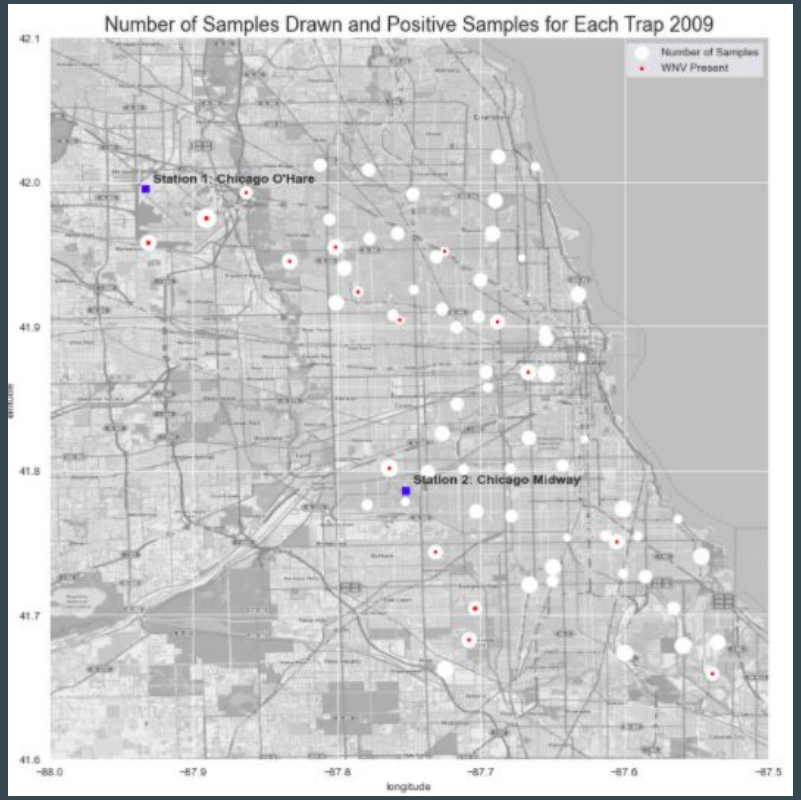


Amongst the 6 unique species of mosquitoes present in Windy City, only **2 species** carry the WNV.

Collected Mosquito Data: Key Findings

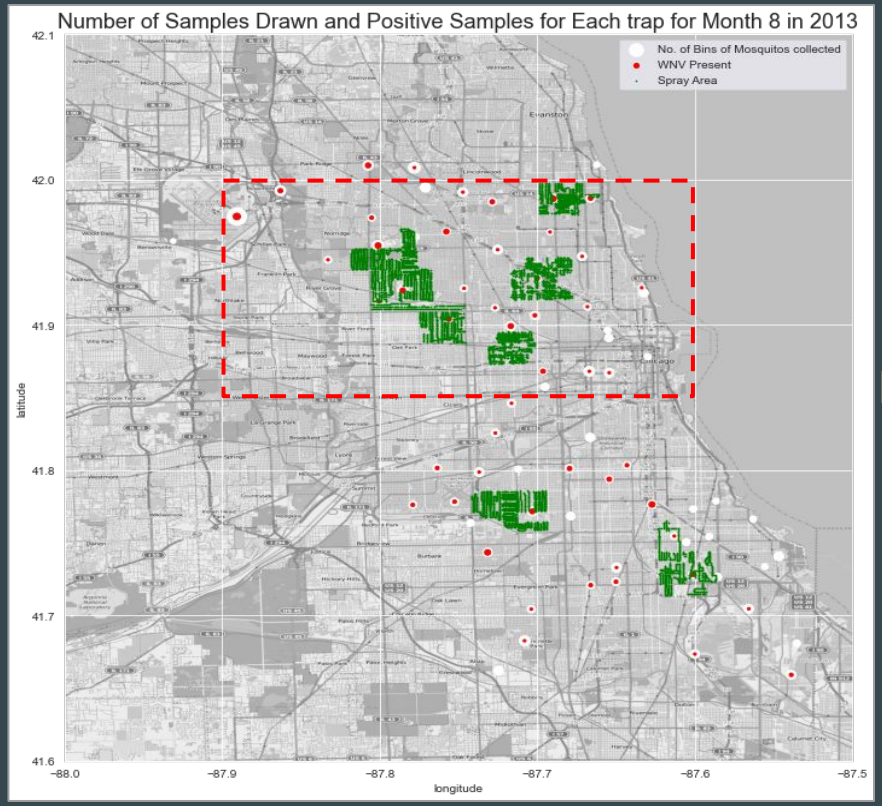
Analysis of WNV presence

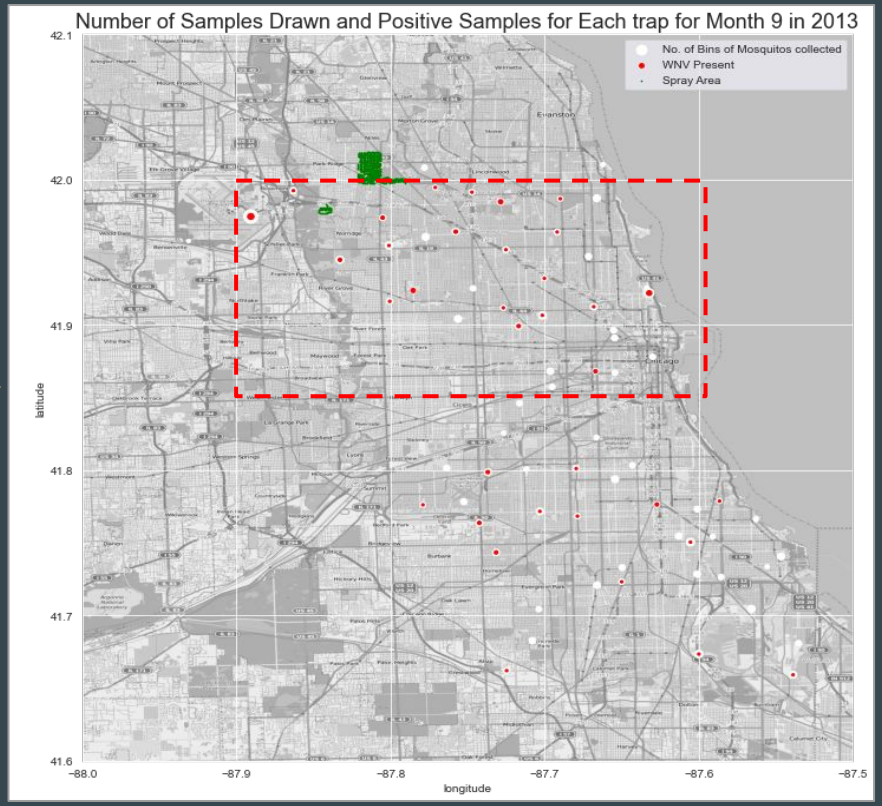




Collected Mosquito Data: Key Findings

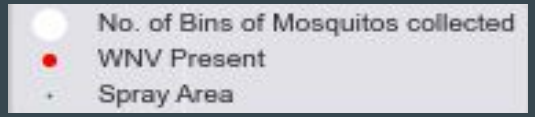
Spray Effectiveness: Reduction in Number of Mosquitos & Presence of WNV





Collected Mosquito Data: Key Findings

Spray Effectiveness: Reduction in Number of Mosquitos & Presence of WNV (Zooming in!)



Sprayed areas (in green) in Aug 2013

ONE MONTH

LATER

Mosquito count & WNV presence in Sep 2013

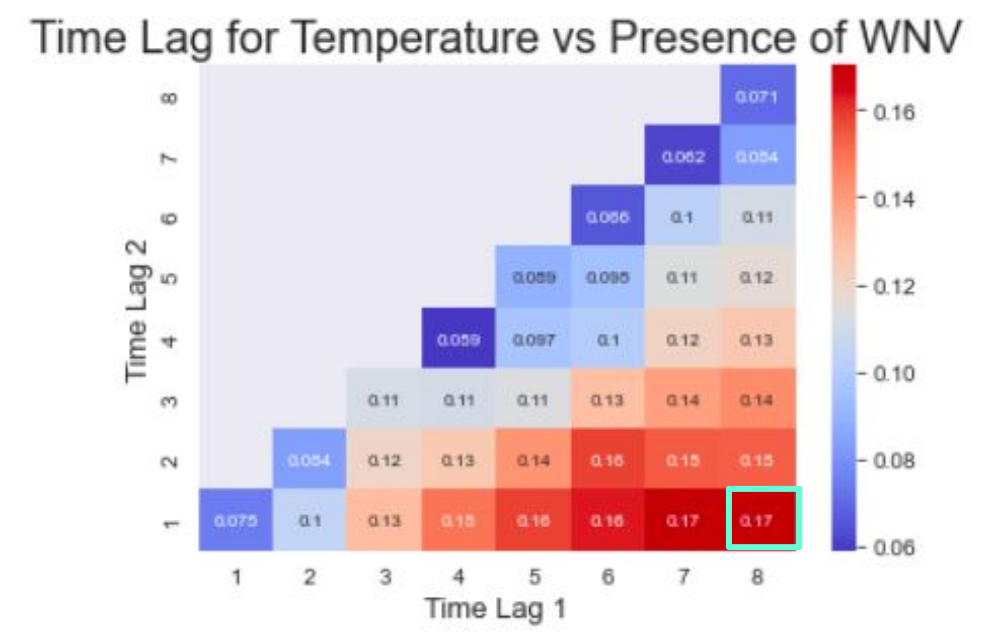
Results:

- Less 'bins' of mosquitoes collected
- Less samples with WNV presence
- Have to take into account rates of infection in each month

Weather Cross-correlation w/ WNV Presence

Which weather features are the most predictive?

Cross-correlation maps (CCMs) depict Spearman rank-order correlations (Spearman rank order correlation coefficients) between various weather factors and the presence of the WNV virus.



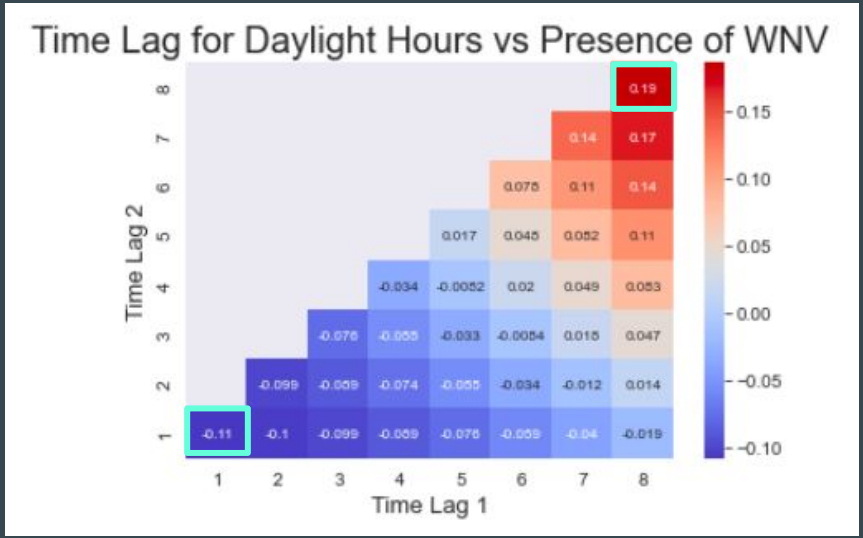
Time Lag 1: Start week for averaged data

Time Lag 2: End week for averaged data

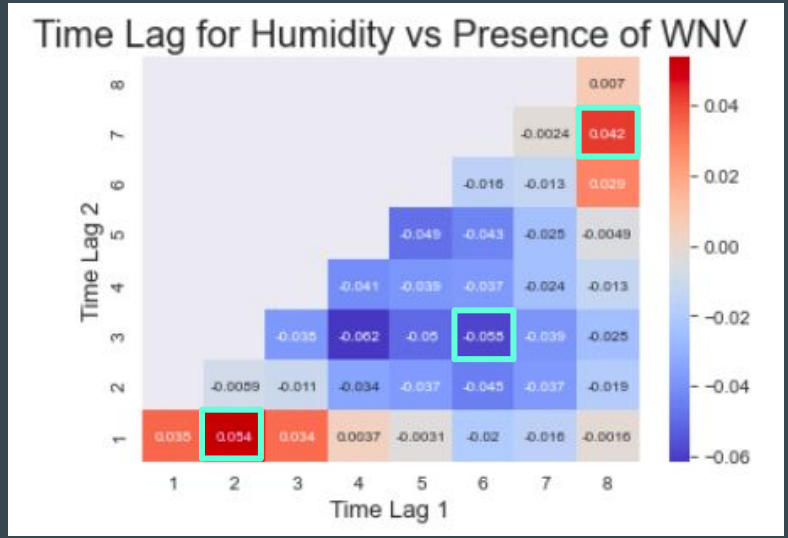
This graph illustrates that the temperature averaged across **the past eight weeks** has the highest correlation to the WNV presence.

Weather Cross-correlation w/ WNV Presence

What features did we shortlist?



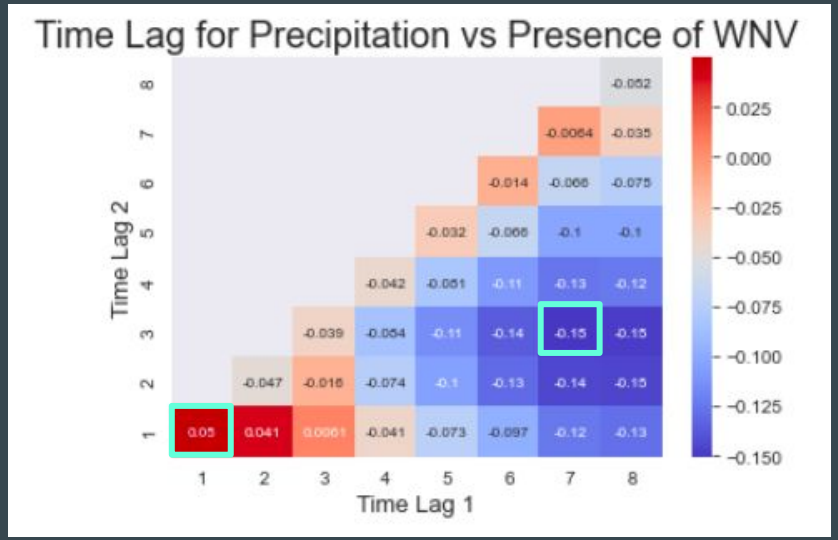
Daylight Hours: avglight_week8_8, avglight_week1_1



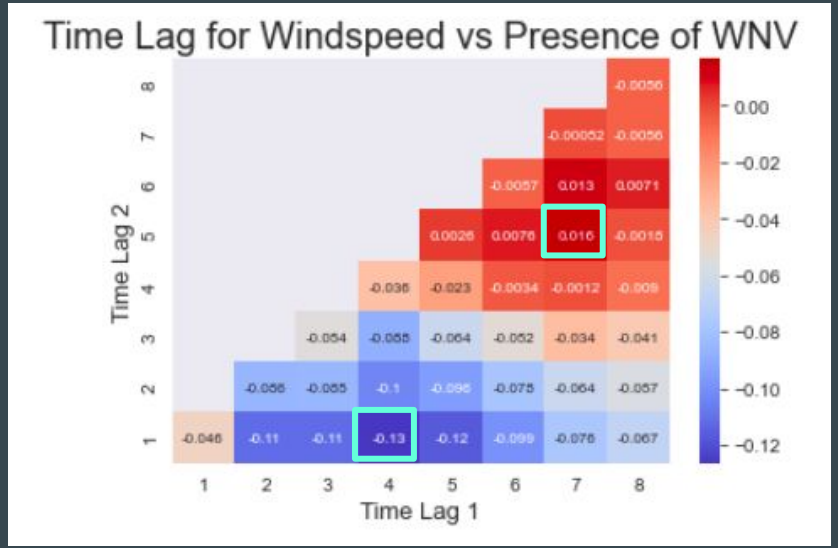
Humidity: avghumid_week7_8, avghumid_week1_2, avghumid_week3_6

Weather Cross-correlation w/ WNV Presence

What features did we shortlist?



Precipitation: avgrain_week1_1, avgrain_week3_7



Humidity: avgwind_week5_7, avgwind_week1_4

Part III
Data Modelling

Data modelling

What is the baseline model?

- The no recurrence model would validate on the train data with a 95% accuracy rate. **Is this constructive?**
- Given that the aim is to predict where the west nile virus might occur, in this situation, **we are looking for a model that picks out as many positive predictions as possible.**
- Sensitivity (SN) is calculated as the number of correct positive predictions divided by the total number of positives. It is also called recall (REC) or true positive rate (TPR).
- **The best metric to maximise might then be the recall, or sensitivity rate.**
- The models used were logistic regression, K nearest neighbours, decision tree, bagging classifier, random forest classifier, adaboost classifier, SVC, extratrees and XGboost

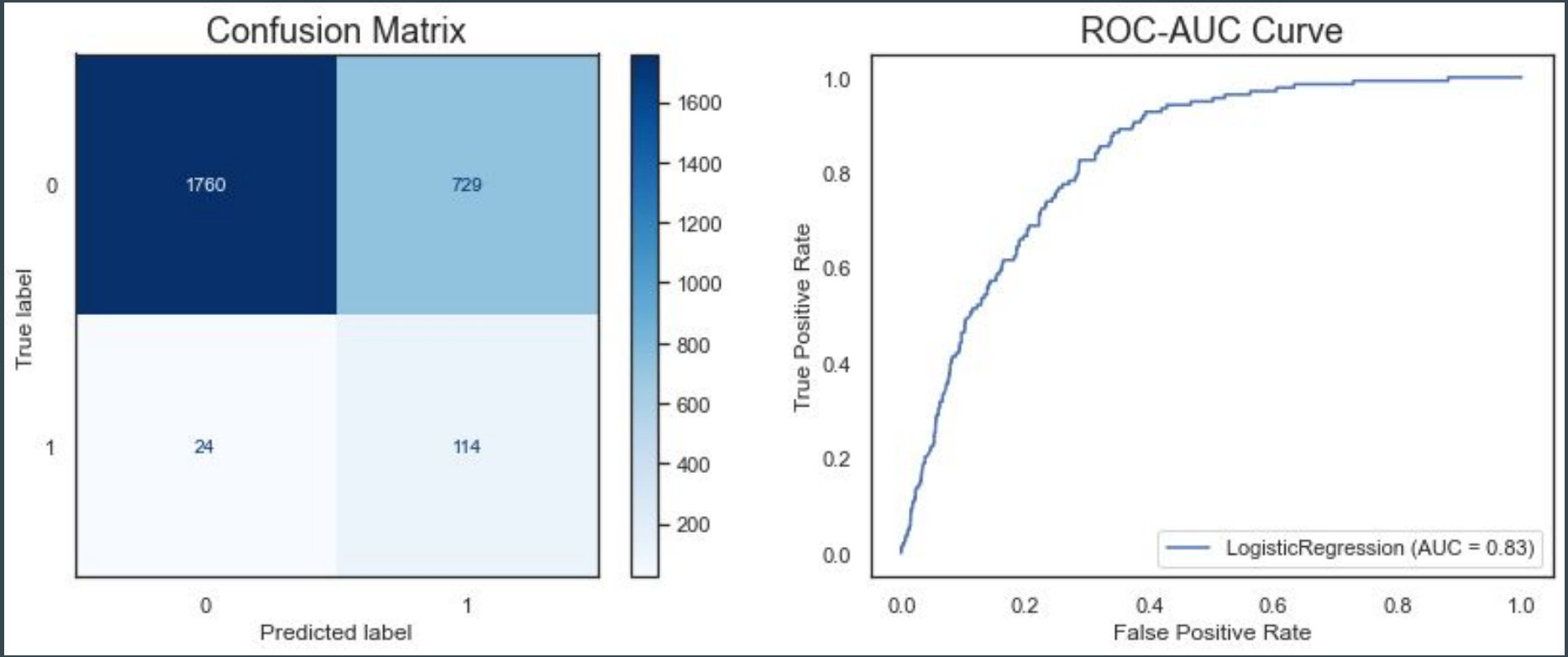
Which is the best performing model?

	Model	Train ROC AUC	Test ROC AUC	Accuracy	F1 Score	Precision	Sensitivity	Specificity
0	lr_clf	0.847	0.832	0.713	0.232	0.135	0.826	0.707
1	knn_clf	0.981	0.714	0.839	0.224	0.150	0.442	0.861
2	dt_clf	0.999	0.601	0.912	0.173	0.171	0.174	0.953
3	bagged_clf	0.998	0.719	0.925	0.154	0.188	0.130	0.969
4	rf_clf	0.999	0.790	0.933	0.201	0.272	0.159	0.976
5	ada_clf	0.961	0.826	0.851	0.282	0.189	0.558	0.867
6	svc	0.965	0.813	0.809	0.241	0.152	0.580	0.821
7	et_clf	0.999	0.732	0.923	0.186	0.211	0.167	0.965
8	xgb_clf	0.997	0.834	0.934	0.224	0.294	0.181	0.976
9	lr_clf_rfe	0.847	0.832	0.713	0.232	0.135	0.826	0.707
10	rf_clf_rfe	0.999	0.789	0.934	0.179	0.257	0.138	0.978
11	xgb_clf_rfe	0.997	0.834	0.934	0.224	0.294	0.181	0.976
12	lr_clf_gs	0.847	0.832	0.713	0.232	0.135	0.826	0.707

Which is the best performing model?

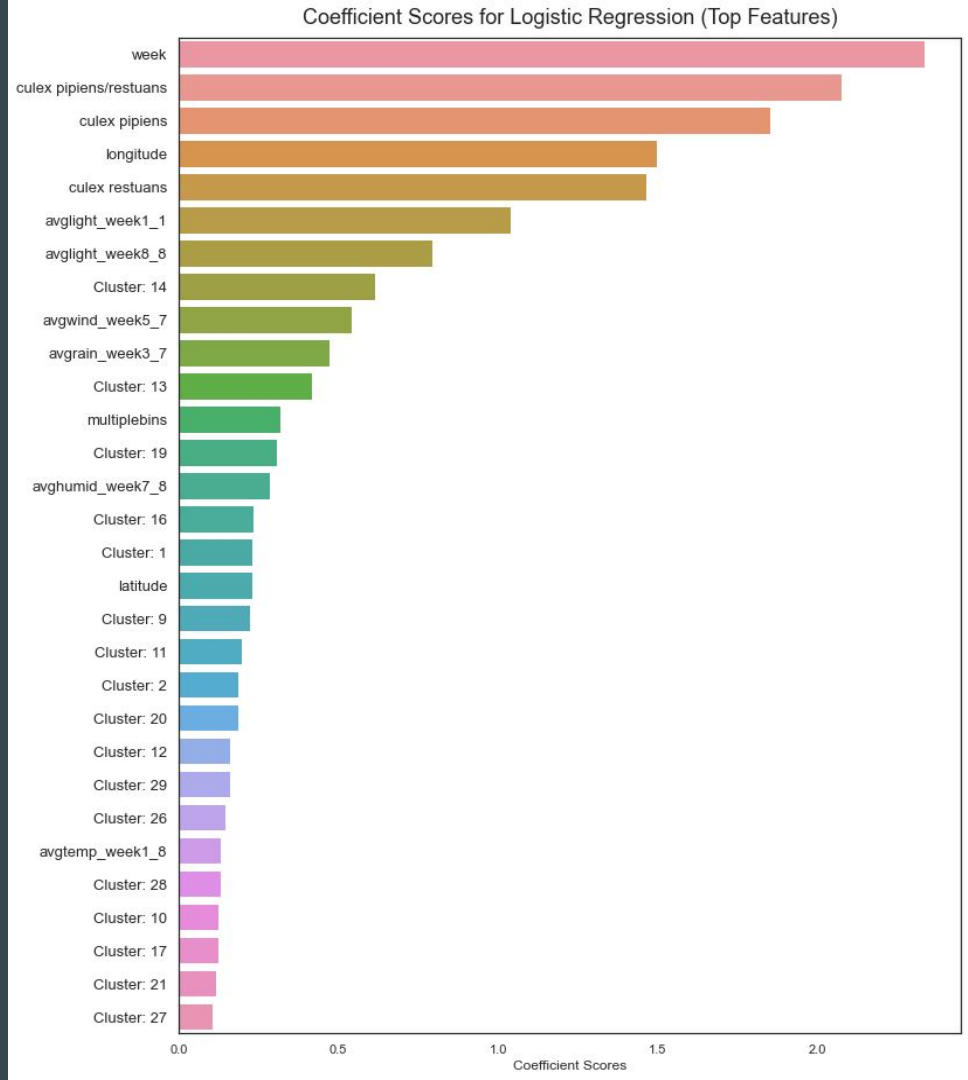
	Model	Train ROC AUC	Test ROC AUC	Accuracy	F1 Score	Precision	Sensitivity	Specificity
0	lr_clf	0.847	0.832	0.713	0.232	0.135	0.826	0.707
1	knn_clf	0.981	0.714	0.839	0.224	0.150	0.442	0.861
2	dt_clf	0.999	0.601	0.912	0.173	0.171	0.174	0.953
3	bagged_clf	0.998	0.719	0.925	0.154	0.188	0.130	0.969
4	rf_clf	0.999	0.790	0.933	0.201	0.272	0.159	0.976
5	ada_clf	0.961	0.826	0.851	0.282	0.189	0.558	0.867
6	svc	0.965	0.813	0.809	0.241	0.152	0.580	0.821
7	et_clf	0.999	0.732	0.923	0.186	0.211	0.167	0.965
8	xgb_clf	0.997	0.834	0.934	0.224	0.294	0.181	0.976
9	lr_clf_rfe	0.847	0.832	0.713	0.232	0.135	0.826	0.707
10	rf_clf_rfe	0.999	0.789	0.934	0.179	0.257	0.138	0.978
11	xgb_clf_rfe	0.997	0.834	0.934	0.224	0.294	0.181	0.976
12	lr_clf_gs	0.847	0.832	0.713	0.232	0.135	0.826	0.707

A closer look at the best performing model



Best features

- **Week** was a suitable proxy for season of the year
- Day of the year and month of the year were also tested but did not improve the model
- The presence of the **3 primary mosquito vectors** strongly predicted the presence of WNV
- **Weather features were also effective** in predicting WNV presence
- Finally, **geographical features** such as cluster, latitude and longitude had the remaining explanatory power



Part IV
Key Findings & Recommendations

Key Findings

What best predicts the presence of the West Nile Virus?

Seasonality

- Our most predictive feature was the **week of the year**
- As we saw from our EDA, the month of August and the weeks within it **provide a timeframe** where mosquitoes are most likely to carry the virus

Weather

- Lagged and averaged weather data from **different time periods in the past eight weeks** proved to be predictive of the virus
- Certain weather conditions in the past, including **temperature, length of day and wind speed**, affect the likelihood of whether the virus is present

Location

- Location features, including **longitude and cluster categories**, were also predictive
- The mosquitos tend to breed in certain 'hotspots' over the years, which may be due to the **physical conditions** of the area or **living habits** of the population in that area.

Key Findings

Cost-Benefit Analysis of Spraying Insecticide

- Cost:
  - It cost approximately USD$25,000 to spray an area of 28 squared km by Vector Disease Control International, the provider for Chicago this year.
  - Scaling appropriately, the cost would be about **USD$540,000**.
- Benefit
  - **1 in 5 people** infected with WNV develop West Nile fever and **1 in 150 people** develop more severe symptoms
  - In 2017, there were 90 WNV cases, including 8 deaths. Assuming the median household income in Chicago of $55,295 and an average hospital cost of USD$25,000 per patient, the cost was approximately **USD$490,000.**
- Analysis
  - It will **not be economic** to conduct spraying insecticide, given the lack of impact.

Recommendations

What course of action should we take?

- Insecticide spraying **has not proven to be significant** in reducing infection rates and requires more data and more campaigns for us to optimise its impact. In the meantime, we should also focus on other courses of action based on our findings.
- **Prepare ahead of time** to combat the West Nile Virus during the period it is most likely to occur and invest more resources in:
    - **Public information campaigns** to prevent behaviour that would lead to mosquitos breeding
    - Increased testing, especially of traps that have a **high prevalence of infection** and traps that are **close to previously infected traps**
- Focus on developing strategies specific to the two species Culex Pipiens and Culex Restuans e.g. releasing **sterilised mosquitoes** to breed with the mosquitos in the area.
- Leverage a **predictive model** based on weather data and past data, to predict the most infectious periods and prepare for outbreaks

Part V
Conclusion

Conclusion

Key takeaways for this project

- **Project Limitations**
  - We would have more insight into the course of action and performance of each model if **we knew how our test data scores on different metrics**: sensitivity, f1 score, accuracy etc., as currently, we are only able to score these by splitting our train data.
  - With more data on **weather for the whole year**, we would be able to better create our averaged weather metrics as well as see overarching trends for each year.
  - The **setup of the dataset** (samples split into bins of 50) is not conducive to creating a model that predicts for WNV at each location as the datasets should be grouped by samples taken.
- Overall, our model helped us to find the most predictive features and we can continue to use it to help the city to **identify potential outbreaks**. It will need to be combined with **vigilant monitoring** on the ground, particularly during high risk periods.