1. Data file examination - Train and Test files:
   1. WnvPresent is what is being predicted.
   2. Test file doesn’t have information on # of mosquitos present like the train file does, so that information is of limited use for analysis.
   3. About 8 types of mosquitos involved. Some are more prevalent. We did t tests and chi squared tests to show ourselves that West Nile Virus (WNV) increases as the number of mosquitos at a location increases, and that some types of mosquitos have a higher prevalence of WNV than others. Although some types of mosquitos carry the virus significantly more than others, data and outside research indicate can’t rule out that any type of mosquito would have the file.
   4. The columns with location information (address, block, trap Id, longitude / latitude) are largely redundant with each other. We decided to take location as a categorical variable since it is unlikely that there is a linear relationship between longitude and latitude and WNV being present. That is, WNV present likely vary up and down as longitude or latitude increases. We used TrapId as our categorical variable. 148 traps around the city.
   5. The observations (rows) in the files have multiple rows for the same trap on the same day. We combined the rows to reflect one observation per trap per day per mosquito type that we had information.
   6. We do not have trap observations for every day. They are roughly twice per week.
   7. The train and test files are for alternate years. This allows us to train a model with all the available data for a set of years and see how well it performs in different years.
2. Data file examination – Weather file:
   1. Information was provided for each day during the warmer months of the year. All days included in the train and test file were included in the weather file.
   2. Information was provided from 2 weather stations (O’Hare at Midway airports.) There were a lot of null values from weather station 2. We determined that the best way to fill them in would be to replace with the information from the other weather station on that day. We then decided to combine the data from the 2 stations into one record for each observation.
   3. Any rainfall observations of ‘T’ (trace) were set to a very small number in order to differentiate it from zero.
3. Data file examination – Spray file:
   1. All attempts to incorporated the spray data into our models was unsuccessful. We believe the issue is with the lack of spray data in terms of number of years. To further our analysis, incorporating spray data into our models would be necessary. But the caveat would be to have complete spray data for all years sprayed.
4. Data file examination – combination of weather file with train and test files:
   1. Did various scatter plots, graphs, Tableau visualizations to determine relationships between variables, particularly between weather factors and the number of mosquitos, the number of mosquitos and WNV, and weather factors and WNV. These, along with research, helped to form our hypotheses.
5. Research we did into the context of the data:
   1. City of Chicago policies on WNV and spraying.
   2. NIH and other literature on WNV.
   3. American Mosquito Control Organization and other web sites on requirements for mosquito breeding.
   4. Various sources on the meaning of weather measurements, like wet bulb
   5. Overall conclusions were that temperature and moisture patterns over time were important
   6. We considered bringing in elevation data for each latitude and longitude based on the hypothesis that moisture might collect in lower elevations. We prioritized this effort as lower priority after we made the most of the provided data.
6. Hypotheses and Feature Engineering
   1. Based on our EDA and on online research on the mosquito breeding, and spread of the WNV, we hypothesized that mosquito breeding, and thus WNV, would be affected by warmth and moisture, both currently, and over the preceding period.
   2. Based on this we added features for precipitation (rate of moisture added), wet bulb (rate that moisture removed), and cooling degree days (heat) over the last 7, 30, and 90 days. Note, that the 30 day features had the last 7 days removed, and the 90 day features had the last 30 days removed to make them more independent of each other.
   3. Variable selection: We looked at feature importances in the Random Forest models, and the confidence intervals in the Logistic Regression models. We removed features with lower feature importance.
7. Choice of Models and reasoning
   1. Decision Trees – We began with a decision tree to predict the probability of West Nile Virus being present at trap for specific days. We found the decision tree did a poor job of predicting the presence of West Nile so we decided to bag and boost.
   2. Random forest – After the failure of the decision tree to get a high ROC-AUC score, we moved on to a random forest model. This model incorporated a wide variety of variables including trap level data, weather data based on both stations, and the features we previously engineered. Overall, the random forest performed well when compared to the single decision tree and we expected this result. We checked feature importances and found many variables were not contributing much to the model and most likely increased noise and bias.
   3. Boosting Methods – After we found which features had the most importance with predicting the presence of West Nile, we moved on to boosting methods. We began with AdaBoost and Gradient Boost but both of these performed worse than our Random Forest model. Ultimately, we moved on to XGBoost in order to optimize our model. The initial XGBoost model contained all of the variables from the random forest model and showed a marked improvement in ROC-AUC. Using subject matter expertise we determined that the most important factor in the spread of West Nile Virus infection rates was cumulative precipitation and cooling days. Thus we incorporated our engineered cumulative features for cooling days and precipitation totals. We decided to include wetbulb in a cumulative fashion but ultimately it did not improve the ROC-AUC score so we removed it from the analysis. Additionally, we cross-referenced the features in the XGBoost model with the feature importances from the random forest model. The most important features were trap location, species of mosquito, day of the year, our engineered cooling degree-days, and our engineered precipitation features. This bolstered our claim that cumulative precipitation and temperature have a profound effect on West Nile infection rates. Next, we attempted to gridsearch across a variety of XGBoost parameters such as learning rate and max depth, but ultimately the default setting from the XGBClassifier wrapper on Scikit-learn performed the best. We also saw a brief improvement in our ROC-AUC score when we used the DMatrix from the XGB package, but ultimately we used Scikit-Learn. We decided to drop unimportant features thus reducing noise and bias ultimately resulting in our highest ROC-AUC score. This final model resulted in a score of 0.78550 on Kaggle. A full breakdown of our scores can be found below. A full breakdown of our code and model can be found in our public Github repository.
8. Cost Benefit Analysis
   1. We investigated monetary cost of spraying (credit to Tina Schendt and team for sharing information she obtained by contacting City of Chicago.)
   2. We investigated non-monetary cost of spraying (environmental and health risks, etc.)
   3. We investigated current spraying practices
   4. We’ve concluded that there are significant monetary and non-monetary spraying costs as detailed here:
   * COST OF SPRAYING: The city of Chicago covers 234 square miles. Based on a mosquito spraying costs of $75/square mile in 2000, after considering inflation we have $105 per square mile. To spray all of Chicago, we have 105\*234 or 24,570. This is roughly ~$25k to spray all of Chicago once.
   * ENVIRONMENTAL COSTS: Environmental costs include mosquitoes becoming resistant to pesticide agents, the death of other insects such as bees, and environmental pollution resulting from the pesticide agents.
   * RATE OF ILLNESS: Approximately 20% of people who are infected with WNV will become ill. Of those who are infected, less than 1% will develop a serious, potentially fatal, neurologic illness. Given that Chicago recorded 9 cases of WNV in 2016, we can assume that those 9 people were among the 20% of WNV-infected who became ill. Since an official recording of WNV requires professional medical diagnosis and care, we can assume that the 9 cases only capture 2/3 of the actual cases since we can estimate that least 30% of individuals do not or cannot seek medical care. This results in an estimate of 15 individuals becoming ill from WNV.
   * COST OF CARE: Based on research, care for a WNV patient who experiences paralysis is approximately $50k. We would want every individual in Chicago who becomes seriously ill to receive care, so we can estimate that the desired cost of care is 15 \* $50k = $750,000.
   * QUALITY OF LIFE COSTS: The city advises individuals to limit time outdoors during times of increased mosquito activity. This results in decreased quality of life to all citizens of Chicago.
   * ADDITIONAL ECONOMIC COSTS: Negative publicity about WNV makes Chicago less attractive to tourists and business investment.
   * TOTAL COSTS OF WNV: The quality of life costs and additional economic costs are more difficult to estimate than the medical costs. However, they are at least as great in volume as the monetary costs. Thus, we can double the cost of spraying to capture the less easily quantifiable costs, resulting in a total estimated cost of $1.5 million.
   * To offset this $1.5 million cost to the citizens of Chicago as a result of potential and actual instances of WNV, it would be rational for the city of Chicago to allocate roughly the same amount to mosquito abatement costs.
9. Spraying recommendations
   1. We recommend using our model to determine areas with higher likelihood to develop WNV based on location and weather conditions.
   2. We’ve combined the likelihoods over time by location to provide a visualization of locations that are historically more likely to develop WNV.
   3. We considered recommending investigating use of drones to spray hard to reach areas. However, they are prohibited within 5 miles of airports. Since that covers a lot of Chicago, we discarded the idea.

**Best Kaggle Scores**

Score 1:

* 0.76924 on public leaderboard
* 0.78445 on private leaderboard

Score 2:

* 0.76730 on public leaderboard
* 0.78550 on private leaderboard  
  We've elected to go with the first one since it has a lower variance across both scores even though it isn't the highest individual score.

**Responsibilities:**

Git Hub and Trello setup: Aakash

EDA: All

Data Munging: Matt (train, test), Chaim (weather)

Data set merge: Matt

Research beyond data provided: Nellie

Feature engineering: Matt

Modelling and feature selection: Aakash, Chaim, Nellie

Kaggle submissions: Aakash, Chaim

Cost Benefit Analysis: Nellie  
Presentation lead: Nellie

Visualizations: Nellie

Log lead: Matt

Code merge: Chaim, Aakash

Trello lead: Matt