Scientific Method

Problem Statement

Mosquito-borne diseases have persistent epidemiological relevance, and West Nile Virus is a recent area of concern. West Nile's currently has few options for effective treatment and 1 in 5 persons showing symptoms, with 1% showing neurological effects. The Chicago Department of Public Health has set up a surveillance and tracking program to see where mosquitoes carrying West Nile Virus are concentrated in the city. Given information from this program, times and locations of pesticide spray to control the mosquito population, and related weather information covering the surveillance period, they request information about efficacy of the program and guidance for future efforts.

What factors contribute to whether or not a mosquito surveillance trap will contain a mosquito positive for West Nile Virus?

Learn from the Data

Train.csv and *test.csv* provide dates and related location data for the mosquito traps being tested, and inform where to pull related observations on weather and spray data. Of note, it also includes species information in the trap. *Weather.csv* contains 2 weather observations per date, with several weather aspects potentially related in model building. *Spray.csv* contains locations and dates, so in theory they too could have a measurable impact that can be considered in model building.

Hypothesis

We hypothesize that a mosquito trap will contain West-Nile-infected mosquitoes based on a number of factors including its species, weather observations, and proximity and time to a recent pesticide spray.

Project Planning

Deliverable

The ultimate goal of this project is to use training data in *train.csv* to generate a predictive model as to whether or not a mosquito trap contains a mosquito positive for West Nile Virus. We can and likely should incorporate included data on weather conditions and pesticide sprays to control the mosquito population. After deciding on features to add per mosquito trap observation, and which features overall to utilize, take data from the test set in *test.csv* to generate probabilities that the trap contains a mosquito positive for West Nile Virus. Ultimately, this list of probabilities should be submitted to Kaggle, who will generate a ROC-AUC score of the prediction list. Accompanying the work is a presentation with Chicago Public Department of Health looking for an effective plan to deploy pesticides in the city, including the ROC-AUC score to reflect the model's predictive ability.

Components

1) GitHub work. Need a repo.
2) Comprehensive EDA. There are 4 files present to explore and need to understand thoroughly before any model building can commence. Understanding what columns of information are present and common to both are crucial to generating good predictions. Likewise, exploring *weather.csv* and its accompanying guide can provide valuable insight for related

weather features for an observation in the training and testing sets. Spray data in *spray.csv* is likely also important to consider for the trap observations in the training and testing sets, so seeing how often and where sprays occur can provide insight for a more effective model. Accompanying all of this is the actual exploration of the data in all 4 files - data types, missing values, reasonable strategies for imputing any missing data are all factors that have to be resolved.

3) Background information. Tied into exploration of the data, but a separate component from evaluating data integrity. Specific questions regarding weather impact on mosquito populations, and effectiveness of mosquito pesticide spraying and related durations need to be addressed.

4) Model building. After data is thoroughly explored from both an integrity standpoint, and understanding what it all represents, careful selection of features that best influence model creation is the next step. After agreeing upon these, selection of one or more modeling strategies follows. Ray will lead the actual coding for modeling.

5) Generating predictions. While a less time-consuming and tedious process than data preparation, this step deserves special attention because the final output of the model generates the crucial data for the deliverable.

6) Presenting results. Need a slide deck and presentation plan to show the work completed and suggestions for the audience.

<u>Priorities and Assignment of Components</u>

1) GitHub. Sparks will create the repo, and Ray's guidance will make conflicts minimal. High priority but low difficulty (ideally).
2) EDA. Sparks will lead in evaluating the data integrity itself, mostly irrespective to what the data represents; will also impute missing but easily calculated values. Miranda will take this work to proofread and do further imputation, and correct any mistakes Sparks made. If subjective imputations are required, team will decide. Understanding the data is high priority, easy imputation is medium, and subjective imputations low priority.
3) Background information. All will separately look up information on weather and spray data, with Ray finding specific related information on mosquitoes. Medium priority as it is important but should not be as time consuming.
4) Team will work together to decide on what makes the most sense for model - both in features, and which models to fit. High priority in all steps.
5) Once team has decided on the final model specifications, generate predictions and upload to Kaggle. Specific person TBD. Medium priority as generating predictions is required, but should be quicker compared to model generation.
6) Presenting results. Sparks will lead creating the slide deck itself, Miranda with Tableau, and Ray in overseeing presentation of information about the model, and next steps. Low priority. While very important to show results, and creation of the materials will take some time, adequate addressing of steps 1-5 should make all team members well versed and prepared for this step.