Week-1 of West Nile Project

* Step one is down loading the data and setting up our work flow management. Trello, Slack, Git repo.
  + Tasks and benchmark goals were set for the following two weeks
    - We touched base on some of our initial observations in the data sets.
* Data cleaning and EDA
  + Nulls and Noaa coding in the weather were dealt with.
  + We decided to keep duplicates, a max observation per trap was 50 mosquitos after that a identical row was populated for each following set of 50 mosquitos.
  + Loaded train, weather and spray into seaborn and tableau plots to examine variables relations and distributions. Such as rainfall and temp, sunrise vs temp, wind v departure. And in Tableau sprays and weather, sprays and west nile, and distribution of traps and WN.
  + Decided most of the location data we wouldn’t use it and would use the Trap variable as our location data.
  + Researched mosquito life cycles and other relevant facts of mosquitos.
  + Reached out to state Dep of Health to understand quest and effects or spraying.
  + Converted our data data and parsed out month and year to use as features.
* Feature Engineering
  + We first worked on engineering features for weather. We settled on averaging temps and rain fall over a 7 and 14 day period based on life cycle of mosquito.
  + Made dummies for traps and species
    - We also included num mosquitos

Week-2 of West Nile Project

* Modeling
  + Now that we had our data and merged all our features we began testing some models
  + We ran 3 different random forest models.
    - One with grid search
    - One with gird search and sratKfolds
    - One without grid search and with sratKfold
      * We then examined our feature importance
  + Using our top 5 importance features from our forest we built a logistic regression to get odds ratios for their impact on WN
  + We also ran a SVM and XGboost on our data and found our best ROC score with XGboost.
* Tuning and Feature selection
  + XGboost performed well on our train data but not well enough against our test. We approached solving this problem from to angles. First we set out to do some more feature engineering. Then we would learn as much about XGBoost tuning as possible.
* Feature Selection
  + We expanded pour weather features to include a few other measurements provided by the weather data.
  + We also decided to expand our time range to 30 days. We set weather averages at 7 days before observation, then 14 days before and finally 30 days before with out overlap.
  + After some trial and error we also settled on factorizing our trap and species data instead of using dummies.
  + Furthermore we engineered a feature to try and capture the # of mosquito data. Because it isn’t part of the test set we need to count it. More mosquitos = more West Nile.
  + We created a binary output that checked if the trap being observed had been checked more than once on the same day 7 days before the current observation. Indicating that a week before the current observation there was more than 50 mosquitos in the trap.
* Model tuning
  + The first few models we used stratified kfolds and gridsearch in the XGBoost Sklearn wrapper. After some research we deiced to switch and start using the native XGBoost syntax.
  + We found that eta or learning rate had a positive effect. We used a high value which helps protect over fitting by making it a less aggressive learner.
  + At the same time we wanted to use a higher tree depth, which allows the model to learn more relations in the train data.
  + We also tried to use eval metric and early stop but they didn’t have a effect.
  + At this time we were also grading our models performance by examining our AUC score.
  + In the end we were getting a AUC of .85 and against the Kaggle test our probabilities were toping out at approx. 75%
* Spray areas
  + We decided the best way to direct spray was using visuals and a csv. We used our probabilities by trap to recommend radial areas around traps with a high probability of WN.
* Cost Benefit Analysis
  + In working towards the creation of our Cost-Benefit Analysis of Spraying Adulticide in Chicago in to prevent the further spread of West Nile Virus we began researching the different reactions to the disease, its infection rates in humans, and the overall health costs.
  + Craig from the Illinois Department of Health retrieved data on the costs of spraying. We then received the great information that it costs at least $60 per linear mile to spray.
  + With this knowledge we looked into the effects, costs, and eventual benefits surrounding West Nile virus. We gathered information private organizations, public organizations, government sources and academic articles and compiled the most compelling information into our Cost-Benefit Analysis.