West Nile Virus

Avinash & John
January 13, 2017

Problem Statement

- West Nile Virus (WNV) is spread to humans through infected mosquitoes
- In 2002, WNV was reported in Chicago
- Chicago sets up mosquito traps across the city and the trapped mosquitoes are tested for the virus
- **Given weather, location, testing and spraying data, can we predict when and where in Chicago mosquitos will test positive for WNV?**
- Goal is to allocate resources more efficiently to prevent spread

Setup

- Train on 2007, 2009, 2011, 2013
    - We know when/where WNV was found in these years
- Test on 2008, 2010, 2012, 2014
    - We want to know when/where WNV was found in these years
- Each row has:
    - Date (May - October for each year)
    - Location (Address, Block Number, Latitude, Longitude)
    - Mosquito species
    - Trap Number
    - Number of mosquitoes caught (training only, capped at 50)

Preliminary Analysis

- 551 of 10,506 entries test positive for WNV
- Dates
  - WNV was never present in May
  - WNV was present once in June (June, 28, 2013)
  - WNV was present twice in October (Oct 4, 2007)
    - 3500 W 51st St & 6600 S Kilpatrick Ave
- Species
  - CULEX PIPIENS & CULEX RESTUANS are the only 2 species (of 7 different species) to test positive for WNV
- Location
  - 97 of 135 traps caught mosquitoes that tested positive

Tableau Visualization

https://public.tableau.com/profile/avinash8553#!/vizhome/WestNile/Sheet1

Methodology

- Assumptions:
  - WNV only occurs in July, August, September
  - Only CULEX PIPIENS and CULEX RESTUANS may potentially carry WNV
  - Only traps that caught WNV in odd years will catch WNV in even years
  - Test data drops from 116,293 observations to 27,226 observations
- For every trap in a given month, we calculated the probability of that trap catching a mosquito that test positive for WNV / Number of mosquitoes the trap caught
  - Named this feature Prob

Features & Model

- Trap
- Week Number
- Month
- Year
- Latitude
- Longitude
- Prob

- Built an eXtreme Gradient Boosting Classifier (with XGBoost)
    - Optimized the model by searching through different parameter combinations

Results

- Goal is to maximize True Positives (instances when we correctly predict WNV) and minimize False Positives (instances when we incorrectly predict WNV)

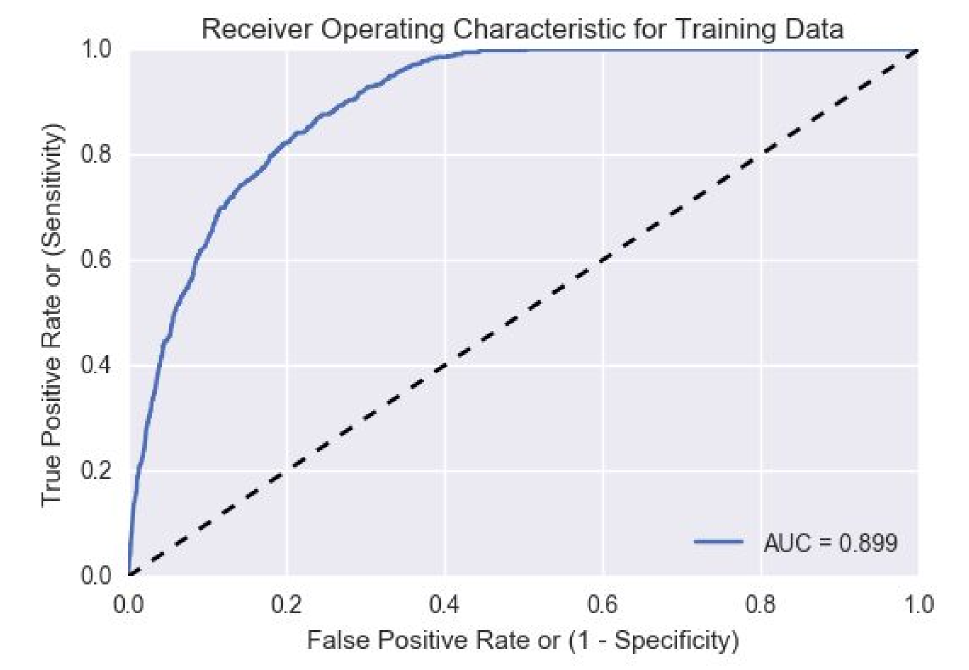
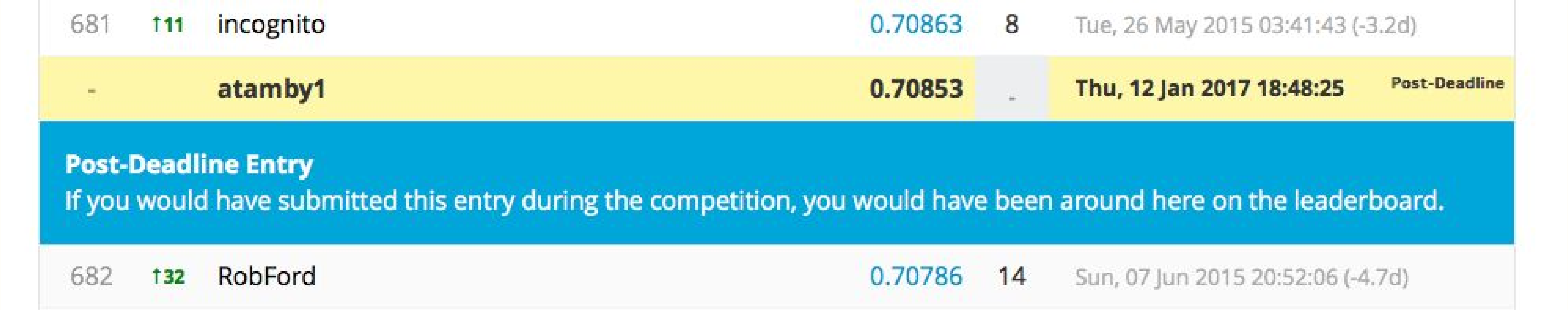


Tableau Model Probability Visualization

https://public.tableau.com/shared/3BY4WRK5S?:display_count=yes



681	↑11	incognito	0.70863	8	Tue, 26 May 2015 03:41:43 (-3.2d)
-		**atamby1**	**0.70853**	-	**Thu, 12 Jan 2017 18:48:25** **Post-Deadline**

Post-Deadline Entry
If you would have submitted this entry during the competition, you would have been around here on the leaderboard.

682	↑32	RobFord	0.70786	14	Sun, 07 Jun 2015 20:52:06 (-4.7d)

Next Steps

- Our model was overfit to the training data (0.899 training AUC vs. 0.708 testing AUC)
    - Training data could have been split by year (e.g. fit on 2007, 2009, 2011, evaluate on 2013) rather than randomly.
- Incorporate weather data
    - Our preliminary research indicated that weather data as we had incorporated it would likely not make a significant difference on results.
    - Initially we compared average temperatures and precipitation amounts within months to different years - it would likely have been more effective to look at average temperature and precipitation during a time frame leading up to the date a trap was tested.