May 3-May 7:

1, Based on the starter code where three factors (block,dmonth,species) are used, the submission performance is 0.65.

2, By lasso+ridge regression via parameter tuning, the performance is not improved at all. CV auc is 0.77 while the submission auc is 0.58. [R\_May7\_SHAO\_glmnet.R]

3,Based on starter code again, add in Longitude and Latitude, the submission auc is 0.69. [R\_May7\_SHAO\_simple\_logistic.R]

May 8:

1, Notice that the NumMosquitos is not the same for different observations, we add in weights=x$NumMosquitos and the submission auc is 0.7. And the summary of this fitting indicates that this is a better model. [R\_May8\_SHAO\_weighted\_logistic\_ERRATICUS.R]

Weight:

W.N. Venables, B.D. Ripley-Modern Applied Statistics With S-Springer (2003), page 190

<http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_logistic_sect013.htm>

This Freq in SAS is what we need first.

Adjusting for oversampling the event level in a binary logistic model

<http://support.sas.com/kb/22/601.html>

We see that there’re many terms are not significant so next step—model selection.

## prep the species column by moving the test-only UNSPECIFIED CULEX to CULEX ERRATICUS, and re-doing the levels

According to the above, the test data file has one species which does not contained in train data file. So essentially, we need try to specify all observations as “UNSPECIFIED CULEX” to one known category BY HAND while the default one in starter code is “CULEX ERRATICUS”. Or, we should put each observation to one of the known category in random.

A better idea is try to infer this “UNSPECIFIED CULEX” from other features.

OUTLOOK:

The nonlinear effects of longitude, latitude and the spray information should be considered later. Final expectation, our model should have auc>0.77.

May\_12:

Using GAM, the cv auc can go to 0.89~0.9 and the submission score is 0.759. Add in s(NumMosquitos) improve the cv auc from 0.81->0.89 and submitted one from 0.71->0.73 and add in s(dYear) improve the submitted one from 0.73->0.759. If we further add in s(dMonth), the submission score is slightly lower (almost same). The test file is revised and variable “NumMosquitos” is created. NumMosquitos equals to the number of observations with all conditions same except for Id. Surprisingly, if we simply let all NumMosquitos=1 in test data set, the submission score decreases to 0.7 which is a significant drop!

This indicates that the test data is possibly synthetic by resampling.

File: starter\_GAM\_num.R

Meanwhile, we also tried the mgcv package which also contains gam function with more flexible adjustments and implemented better in computation than above gam pakage. There’s some improvement in cv auc but no improvement in submission score. There’re so many things can be tuned in mgcv gam function.

File: starter\_mgcv.R

It seems that TrapNumber has the same information as Latitude, Longitude these predictors results in similar auc.

Next try, tensor product of splines in mgcv package.

Feature: Address number may be useful as well.

Considering leaderboard:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 11 | **↑230** | [Rafael](https://www.kaggle.com/users/84462/rafael) | 0.79143 | 30 | Tue, 12 May 2015 01:12:44 (-2.2h) |
| 19 | **↑275** | [Victor](https://www.kaggle.com/users/27805/victor) | 0.78179 | 31 | Mon, 11 May 2015 10:13:02 (-0.1h) |
| 39 | **↑201** | [Jianmin Sun](https://www.kaggle.com/users/180151/jianmin-sun) | 0.76476 | 12 | Sat, 09 May 2015 04:47:36 (-24.1h) |
| 46 | **↑164** | [maglab.physicists](https://www.kaggle.com/c/predict-west-nile-virus/leaderboard) | 0.75979 | 29 | Tue, 12 May 2015 02:37:25 (-1.2h) |

These are significant improvement ahead of us, such improvement basically indicates that there’s some qualitative change of method or features. The improvement does not come from any fine tuning.

It seems that we still miss something significant or we should try complete different method and a possible candidate is regression time series model.

May\_15:

The GAM,GAMboost,GB methods have been tried and varies of parameters are tuned.

File: starter\_mgcv.R starter\_gbm.R starter\_mboost.R shao\_mgcv\_gbm\_mboost.rmd/pdf

Regarding the weights, it’s illurstated in simple example,

/useful materials /weight\_clarification.R

<http://stackoverflow.com/questions/9111628/logistic-regression-cbind-command-in-glm>

“When doing the binomial or quasibinomial glm, you either supply a probability of success, a two-column matrix with the columns giving the numbers of successes and failures or a factor where the first level denotes failure and the others success on the left hand side of the equation. See details in ?glm.”

Above are correct in normal case but this way makes our results worse in gam,gamboost or gb method. Meanwhile, the train\_expand data also does not provide good performance. Our data corresponds to unequal sampling which needs adjustment:

<http://support.sas.com/kb/22/601.html>

Useful website!!!

<https://public.tableau.com/profile/jaysha101#!/vizhome/WestNileinChicago/WestNileDashboard>

<https://www.kaggle.io/svf/11469/c395c2e01c2247738f16f0b4f3078fba/output.html>