4th place (brief) tips

By Μαριος Μιχαηλιδης KazAnova

Posted in talkingdata-adtracking-fraud-detection 3 years ago

135

I would like to start via praising my teamates for an excellent effort and congratulate the winners for an intense last-days' race.

Also thank you to all the people who shared code and ideas - they made it a great competition (with the exception of the latest high-scoring kernels) . Special thanks to anttip for his overall contribution with wordbatch and kernels in general .

My favourite positions in a kaggle competition are 1st and 4th. 1st you get most points/money. 4th, you dont get as many points, but you dont have to reproduce your solution :)

Our validation schema was as follows:

We used days 7,8 for training and we were making predictions for 9th day (hours [4,14])
For test predictions, we were training on all 7,8,9 and making predictions for the test day (10th)

Things that work for us apart from what it is already in forums/public kernels:

1) Restacking - When we first started doing Stacking, we could barely get 1,2 points our of it (like from 0.9818 to 0.9820). After adding ALL the features used in our standard modelling to the predictions of the ninth day - we got another +3 boost (to 0.9823). We ended up having around 50 models - mostly lightgbms, but also nns , FMs and some linear models. NNs were on par with LGB models.

2) We got another +4 from creating WoE (Weights of Evidence) features for many combinations of all variables (ip,app,device,os) . This is very strange , because we tried standard target encoding and **it did not work**. This is very strange, because the ordering of likelihood features and woe should be the same/similar (only the range in woe is more condensed) . We are not sure why this happens, maybe it has to do with the binning of Lightgbm?

3) We got a boost of +4 via sorting ties of time in the same groups of ip,app,device,os, making certain the is_attributed==1 **is always last** as it was pointed out in the forums.

Each one of the team members had different features - mine were more related with time-series. Like counts of previous (and next) days, hours, minutes and seconds of ips,apps,device (and their combinations) .