Featured Prediction Competition

Instacart Market Basket Analysis

Which products will an Instacart consumer purchase again?

$25,000
Prize Money

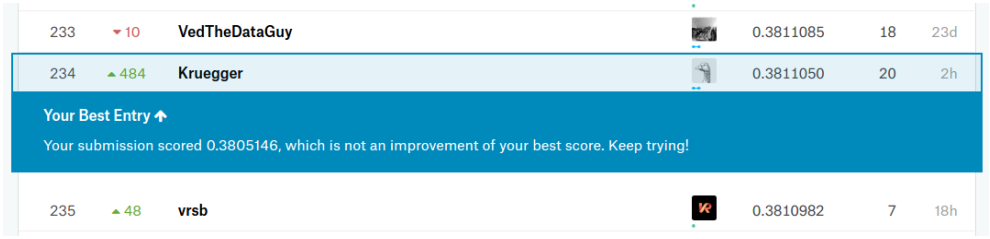
Instacart · 2,025 teams · 18 days to go

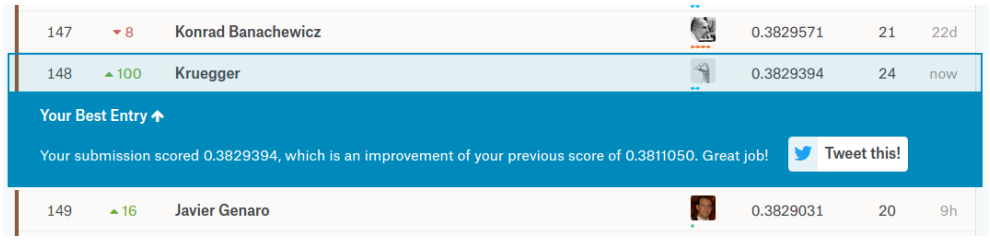
Kruegger
46th place

How to lift up to Bronze area in a week

posted in Instacart Market Basket Analysis 5 days ago

66





Hello!

I am novice kaggler and writing this post to help another novices not to give up competition even if you think that you are out of ideas and knowledge.

At first I want to say 'Thank you' to all people who shared their thoughts and expertize - your information and approaches very helpful and useful. Small list of these guys (but not compelte) here:

@AlphaMeow @SVJ24 @raddar @Fabienvs @paulantoine @Li Li ... and many many others ...

It's my second serious competition here and second post on kaggle) So I am not pretend to be experienced people, but want to share some thoughts and information compiled from discussions and kernel analysis.

Main topic - you can do it! Really, as mentioned early in some discussion, there is no

Magic feature or leaking, just EDA/Feature engineering and some luck) If you are sticking on 0.38 LB by some of public kernels - don't give up, just try understand what you can do and do it!)

Some basic ideas about this competition compiled from discussions:

- basic model (presented in high voted kernels) is binary classification with logloss over (order, product) pair and 1/0 as target from 'reordered' column on train dataset. You can also try bayesian/rnn/..., but this model is simple and well done.
- Correct CV!!! Mercedes show us that we can't underestimate importance of correct CV scheme. In this competition it is easy - just create folds based on disjointed user_id.
- Treshold to convert probability to 1/0 is the key! Don't use default 0.5 - start with 0.2 and tune it on CV. The more advanced idea is to use different treshold for different orders. I give a link to some topics later.
- Feature engineering. In this competition it is the most important part. You have to read the book from post @Rodolfo Lomascolo ("Repeat Buyer Prediction for E-Commerce") https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/36411

Thats all to jump over 0.38. Really, just to try it)

And some advanced to improve your result:

- try to predict None as separate product in the order or try to predict basket size of the user - to implement F1 expectation scheme.
- add bayesian aproaches to your model
- ... here the place for your imagination ...)

And some links to the discussion i mention interested:

Data understanding

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/33128
https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/33448

CV

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/36493

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/35048

What does 'Reordered' mean

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/33211

NONE handling:

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/36134
https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/35716

F1Score:

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/36544
https://www.kaggle.com/aikinogard/python-f1-score-function

Features:

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/35468

Model selection

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/33131

Kernels:

https://www.kaggle.com/fabienvs/instacart-xgboost-starter-lb-0-3791/code
https://www.kaggle.com/paulantoine/light-gbm-benchmark-0-3692
https://www.kaggle.com/nickycan/lb-0-3805009-python-edition
http://kelsh.tech/blog/2017/06/21/analytical-approach-to-kaggle-instacart-competetion/ http://blog.stylemyimage.com/post/kaggle-instacart-market-basket-analysis-competition-solution-with-c-and-vowpal-wabbit

Advanced)

https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/36312

Thats all. Thank you for reading and happy kaggling!!!

@kruegger.

P.S. you can upvote if you want)

Options

Comments (38) Sort by Hotness

Click here to enter a comment...

xentropist • 4 days ago • Options • Reply ^ 2 v

waiting for the guide to jump over 0.40 :-) nice work.

AlphaMeow • (15th in this Competition) • 5 days ago • Options • Reply ^ 0 v

It's my first competition actually, so I am sure you will do better than me. ;)

At current stage, I lost direction and don't understand how to gain more without having more features/stacking... I wonder how people get more than 0.402 and even higher. They should have some excellent ideas that I am not sure whether they will give me a hint :)

Learning to think and understanding why it works are more important than using other people's ideas.

Kruegger • (46th in this Competition) • 5 days ago • Options • Reply ^ 1 v

You did the best work for the first competition! Grats!)

About ideas, from my point of view there are three important part of solution any participant have to use to achieve good results:

- Feature engineering (as I understood even one more nice feature can lift performance of the model better than parameters (eta/nrounds) tuning)
- Handling of None (build separate classifier or use statistical way)
- Handling tresholds (use ExpF1 technics or try to predict BasketSize)

Additional possible way to improve

- Ensemble ??? I don't think that ensemble in classic way helps, but who knows..
- Bayeaian approach. My unexpert intuition tell me that it may be the next best step...
- RNN. we don't have long period of series to implement rnn, but may be we have to look at this problem in the out-of-the-box manner

Can you add something that I miss?)

P.S. BTW, another two questions about methods:

1. feature selection. my current approach is to build model using block of new features, then look at 'gain' of f-score, normalize it to [0,1] and select only features above some treshold (0.5 / 1.0 / ...) Is it normal? Or there is more stat of the art scheme to select important features?
2. find best nround. In the same time when I add new block, I try to estimate new best nrounds for boosting. Using fixed lr=0.1, then CV/Holdout with early stopping=20 until logloss doesn't improve. Then multiply best nrounds value to 1.2 (I use 5-folds) and voila - next best nrounds is on the scene)

Is it correct? I ask you as I get some strange results last night: - old features - best nrounds=400, score (for example) 0.24433, EstF1 = 0.3840 - with new block best nrounds=200!!!, score 0.24431, EstF1 = 0.3800 !!! (worse) - if I relearn model for the same 400 rounds I get EstF1 = 0.3850 (better) So it seems that CV stop learning early than best point. I'll try to use different early stopping values and so on, but my question - did you see same behavior? And how you choose best nrounds for new block of features?

AlphaMeow • (15th in this Competition) • 5 days ago • Options • Reply ^ 1 v

Thanks.

I don't know any others ways to improve score currently. Certainly there must be something I haven't discovered, e.g., basketsize patterns. Maybe it is a good idea to ask more experienced people around.

1. I also wonder. After organzing my codes, I have too many features(approx 88) now, and it seems dropping out the least important ones doesn't work, which is the worst of all.

This will be a disaster for me to shift to xgboost, which uses much more RAM(I can only have 60 features for xgboost) but produce better result. ・゜・(ノД`)・゜・ Would be very grateful if someone can help, or give some clues on reducing features.

1. For this one, I had the same problem, with CV gave me a much earlier stopping time. Thus in the end I not only did CV but also run lightGBM manually on each CV to see what is approximately the best round(give best log-loss and AUC). And yes, I still feel unsafe that I am overfitting/underfitting.

raddar • (13th in this Competition) • 4 days ago • Options • Reply ^ 3 v

why would you want to switch to xgboost. It underperforms to lightgbm here :)

AlphaMeow • (15th in this Competition) • 4 days ago • Options • Reply ^ 1 v

@raddar I remembered someone saying xgboost has better None handling and produce more accurate results(+ 0.0006 LB) in another post?

jcleon • (140th in this Competition) • 4 days ago • Options • Reply ^ 0 v

@AlphaMeow, I follow you long time, and saw you are very hardworking and smart person. May I ask a question? My None handling model CV's F1 is above 0.97, which is based on order tables as you said. However, When I combine this None handling with my main model, the score decreased to 0.3654. I use the fixed threshold for main model right now. Is it the reason I lose in my None handling?

raddar • (13th in this Competition) • 4 days ago • Options • Reply ^ 1 v

@AlphaMeow, I have different conclusion - xgboost takes longer to train (even approx or hist methods) and do not yield better results.

AlphaMeow • (15th in this Competition) • 4 days ago • Options • Reply ^ 0 v

@raddar Really? Thank you very much for this information(that really saves a lot of time)! I was just about to run xgboost with a fraction of my features.

Do you think stacking xgboost and lightGBM models will help? In fact I am not even sure whether stacking will help in this particular competition...

raddar • (13th in this Competition) • 4 days ago • Options • Reply ^ 7 v

Well you can try, but for me it is a waste of time. I'm still missing some important features, and i want faster feedback with faster models now ;)

stacking does work - single best: 0.4023414 -> stack: 0.4028114

AlphaMeow • (15th in this Competition) • 4 days ago • Options • Reply ^ 1 v

No I am neither smart nor hardworking. :) Truly, spent hours on simple mistakes.

I don't know why but did you simply assigning Nones to orders that are predicted as None? This will not work, since you can either add None to products or make the order None, and the latter way can be an overshoot.

AlphaMeow • (15th in this Competition) • 4 days ago • Options • Reply ^ 0 v

@raddar I see the point here. Thank you!

jcleon • (140th in this Competition) • 4 days ago • Options • Reply ^ 0 v

Thanks a lot. I am wrong because I simply assign Nones to orders that are predicted as None

Kruegger • (46th in this Competition) • 4 days ago • Options • Reply ^ 0 v

@AlphaMeow - about None prediction. Do you mean that when I get the results from my None classificator (orderNone = 1/0), instead of just put 'None' to all (order, product) pair where orderNone==1, i should ADD 'NONE' as product to the predicted product list for this order?

AlphaMeow • (15th in this Competition) • 4 days ago • Options • Reply ^ 2 v

@Kruegger

You can predict None from two sources: 1) build your own none classifier on order basis, this outputs the probability of order A being a None; 2) From your user-product probability model, P(order is None) = (1 - P(prodA reordered)) * (1- P(prodB reordered)) * (1 - P(prodC reordered)) ...

You should figure out when to make an order None and when to add None to products. They are two different cases that should be distinguished.

mezoganet • (227th in this Competition) • 4 days ago • Options • Reply ^ 0 v

@AlphaMeow

I think you did more than I did in my first competition.

You will get a silver medal at first attempt ;) For sure.

GeorgeGui • (2nd in this Competition) • 4 days ago • Options • Reply 0

@raddar it's interesting that your lightgbm is better than xgboost both in speed and accuracy. Probably due to my poor tuning ability in lightgbm, it givers 0.001 lower LB than xgboost.

neverdies • (252nd in this Competition) • 4 days ago • Options • Reply 0

@jcleon Hi, I'm also working on None handling but cannot get a high f1 score. My none classifier is based on order_id, I use similar features as the main classifier: order related features, user related features and mean() of product related features. But I could only get .65 f1 score with xgboost. Not sure if I'm doing something wrong with it....

AlphaMeow • (15th in this Competition) • 4 days ago • Options • Reply 0

@neverdies

Probably not the best way, but your binary classifier outputs probability of an order being a none order. You can use that probability to determine whether to add none to products, or make an order complete none.

Later I found out this can also be done more easily with the user-product classifier, by using P(order is None) = (1 - P(prodA reordered)) * (1- P(prodB reordered)) * (1 - P(prodC reordered)) ... Strangely the result of this approach is not the same as previous approach, thus staking the two gives me better result.

Yet you can also treat none as a product. I am still trying and testing thus can't tell you more.

Hope it helps.

happycube • (560th in this Competition) • 4 days ago • Options • Reply 0

I find that lightgbm needs to be properly tuned to match/beat xgboost - then again the air up there at #2 might be a bit thin for lightgbm - I've never flown that high myself. ;)

jcleon • (140th in this Competition) • 3 days ago • Options • Reply 0

You should CV on prior and train tables as the basis, just split by index will be fine. However, though you can get very high F1 in this model, assigning the value to orders that

are predicted as None in your submit does not work. Now, I am trying treat None as a product and add it to my main model. When my main model cannot determine whether the order is None, the second model based on prior and train tables try to predict for the main model.

jcleon • (140th in this Competition) • 3 days ago • Options • Reply 0

@AlphaMeow May I ask whether I am in the correct direction?

AlphaMeow • (15th in this Competition) • 3 days ago • Options • Reply 0

Yes, but in my case, sometimes assigning complete Nones to orders gives better result than attaching them to products. The reason is unclear, probably due to number of products available to the user.

jcleon • (140th in this Competition) • 3 days ago • Options • Reply 0

Lukasz Grad • (36th in this Competition) • 3 days ago • Options • Reply 0

Hey @AlphaMeow I was trying to build separate classifiers for None orders, seen some posts about 0.97 F1 score. However after a whole day of struggling my classifier isn't better than my main one :) (I trained on train + prior set also, order based) Are you sure that this 0.97 is without any leakage? Did you test it on LB? I'm sorry for being so nitpicky, I just want to be sure it is possible :)

AlphaMeow • (15th in this Competition) • 3 days ago • Options • Reply 1

@Lukasz Grad In my case it gives me roughly +0.0005(?) on LB, extra gain from my main none handling based on user-product reorder model. **It is possible that this extra gain is due to stacking, not this binary model**, since the result I obtained from case 1) none handling by separate binary classifier+main model and 2) none handling by main model + main model are similar, but the combination gives a small increase in score. I guess they are solving the problem from different aspects.

I trained this separate classifier based on all orders, and treated very carefully with some of the orders(e.g., the first order is always None thus are bad cases that should be discarded, etc).

Lukasz Grad • (36th in this Competition) • 3 days ago • Options • Reply 0

Yes I discarded early orders from prior, I also calculated cumulative user aggregated features, and cumulative aggregated product features from previous orders, for each user, and some other things and I feel like I'm stuck.

Hmm i thought there would be bigger increase in score, in train data there are 6,5% None orders and you basically can predict them perfectly. Anyway, thanks for help, maybe I will come up with some better ideas today :)

AlphaMeow • (15th in this Competition) • 3 days ago • Options • Reply 0

@Lukasz Grad

I think SVJ24 said treating Nones as products in training can help in another post, but I haven't tested yet. It is definitely a good direction to try as proved by many others, though you need to adjust your 3 sets and avoid leakage.

Also, think about not only predicting Nones, but when to predict None and when to add None to products. I guess their model handles this better than mine.

link: https://www.kaggle.com/c/instacart-market-basket-analysis/discussion/36134

Kruegger • (46th in this Competition) • 2 days ago • Options • Reply 0

@AlphaMeow - thank you for suggestion. I'll try it.

Steven Nguy... • (261st in this Competition) • 5 days ago • Options • Reply 0

As someone that's only been playing with parameters in the kernels (and catboost), this is a procrastinators perfect resource!

kobe • (32nd in this Competition) • 5 days ago • Options • Reply 0

thank you !

CSAdu • (570th in this Competition) • 5 days ago • Options • Reply 0

This post is very helpful for the people just start the competition! Thanks!

mezoganet • (227th in this Competition) • 4 days ago • Options • Reply 0

Really nice !!!

Jegan Karun... • 4 days ago • Options • Reply 0

Thanks. Very useful info for the beginners

Emin Ozkan • (125th in this Competition) • 3 days ago • Options • Reply 0

@Kruegger Thanks for all these suggestions. Could you please tell me the importance of CV? I do understand that training and validation user_id cannot overlap. But because we have so much data does CV matter. Even if you do 10 fold CV, when you have so much, aren't all models going to be similar anyway. Or is there another purpose to CV that I am missing?

TonyChan • (497th in this Competition) • 3 days ago • Options • Reply 0

Local CV prevents overfitting the public LB.

Kruegger • (46th in this Competition) • 2 days ago • Options • Reply 0

Just look at this question on different angle:

You have huge dataset and want to split it to parts (folds) for different goals:

- use several fods as holdout data (for testing, tuning another parameters, e t.c.)
- use CV for select best nround for GBM
- ... and so on

When you do splitting you should be absolutely sure that data in all folds has identical charasteristics (attributes distribution, and so on), that there is no any kind of leakage, et.c.

When you CV you loop throw all folds (ex: 5), use combined 4 as train data and 5-th as test data. To be sure in results - folds have to meet criterion I wrote earlier.

So - proper dividing your data to folds is the key to build robust model. In this competition it is simple - just split to folds with disjoint user_id.

Even more - as all folds have the "same" characteristics as the whole data, you can use only small part of data to build your model (1 fold), and if you splitting right way - you will get correct results.

Hope it helps.

Kruegger • (46th in this Competition) • 2 days ago • Options • Reply

0