Featured Prediction Competition

Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?

$30,000
Prize Money

Corporación Favorita · 1,675 teams · 3 days ago

Hide on bush

preliminary LB 16th solution——Self-examination of a violator

4

posted in Corporación Favorita Grocery Sales Forecasting 6 hours ago

Self-examination

Sorry to tell everyone, I violate the rule. My 16th rank was removed by kaggle after preliminary LB public 12 hours later. Our team used a second account to submit for about 9 times total in the last 6 days. I am very very sad and regrettable. Suddenly fell from the mountain into the bottom.

This is our first kaggle competition after I learned machine learning myself for about 8 months, so we are not particularly clear about specific rules. We participated this competition at 11/24/2017 and only do this competition every day in the next 2 months. I actually know this behavior of our team, and I also participated in. Although I know this behavior violate the rule, I still do this because I don't read the rule even once. I just heard the second account that for submit will be cleaned and I don't know the main account will be removed too. I had been searched this question that another account for submission online in Baidu or Zhihu and nothing can be found. So I relaxed vigilance and just think this is a small trick. I am too silly and foolish. If I know the result i will never do that. I told my mother I had ranked 1% in 1709 teams the first time and my mom was proud of me.

Now I don't dare to tell the truth to her.

Sorry to everyone and sorry to myself !

There is no record in this competition about me. I just want to prove I had been here before. I wish I could feel better after i post this discussion.

Solution

I am new to kaggle and machine learning compared to many experts and masters. So my experience may inspire some new like me. Although I violate the rule, but I don't think the improvement of these extra nine submissions would be much for a new.

My solution is a update version of Bojan Tunguz's lgb model.

https://www.kaggle.com/tunguz/lgbm-one-step-ahead-c8de0f/code

The final weighted average can not improve the public LB score, so the final score is just a single model lightgbm.

I had tried some methods to handle holidays and the pubilc LB score are all decrease.

We had tried to replace the holidays' sales with the sales a week before the holiday according to specific store_item combination. Because the sales according to specific store_item combination highly related to the day of the week, so i wanted to smooth the effect of holidays.

We had also divide the sales at holidays by 1.5/0.7, it also don't work well.

So i didn't take holidays into account at last.

2. Categorical features

There are many categorical variable in this competition, like type, cluster, city, state, family, class, perishable, store, item. Cashine Lee's *LGBM Starter* only use the sales of store-item combination. According to these categorical variable, there are many kinds of combination, like city_item, city_type, family_class etc. So there are many hide information.

We use Boruta to select features from these combination. These features are (mean/std/min/max/median) (sales/transactions) over the last (3/7/14/30/60/140) days.

Although we select many features from method above and public LB score increase at the beginning , we decide adopt only 2 from these combination, city_item and country_item(the country's sales of this item) from the perspective of preventing overfitting and reducing calculation. Because there is a lot of redundancy between features above from other categorical features combinations.

These can improve the public score from 0.514 to 0.512 with Bojan Tunguz's lgb model.

3. other featues

   We counted NaN values from past 7/14/30 day sales, and NaN values from past 7/14/30 day onpromotion. These can improve the public score from 0.512 to 0.511.

   And I also dropped some redundancy features from Bojan Tunguz's lgb model according to the importance. These can reduce the calculation.

   Although some other features are not so important, like store_nbr, item_nbr, but can improve the score 0.001 too. You can download the code below.

4. Key Training Set Settings

   Public kernels are all predict future 16 days' sales use data before 7.26. We think this method cannot take full advantage of recent data from 7.26 to 8.15. So we decide to divide 16 models into 3 parts.

   The first part of model predict sales in first 6 days of next 16 days, we can set data from 5.31 to 8.2 as training set and data 8.9 as validation sets.

   The second part of model predict sales in middle 7 days of next 16 days, we can set data from 5.31 to 7.26 as training set and data 8.2 as validation sets.

   The third part of model predict sales in last 3 days of next 16 days, we can set data from 5.31 to 7.19 as training set and data 7.26 as validation sets.

   These can improve the public score from 0.511 to 0.509.

   Another discovery is we find start the date from 6.7 can be better than start the date from 5.31. It could be there is a National Holiday Mother's Day in May and there is big influence to sales.

   Finally, we canceled validation sets to use more recent data.

   These can improve the public score from 0.509 to 0.508. And the final private score of 0.508 is 0.516.

Ending

Above is almost all I want to share. You can download code for a better understanding. Perhaps these settings do not seem difficult, but I really spent a lot of effort to find out.

final_code_to_discussion.ipynb (112.84 KB)

Options

Comments (0) Sort by Hotness

Click here to enter a comment...