Featured Prediction Competition

Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?

$30,000
Prize Money

Corporación Favorita · 1,707 teams · 15 hours ago

Luck Yu
2nd place

2nd place solution overview

posted in Corporación Favorita Grocery Sales Forecasting 2 hours ago

12

First, I want to thanks @sjv. He shared model in WTF is our best model baseline. and his code is clean. I think he is good model for us.

Second, thanks to our team members to stay with me, discuss and work hard together until last minute.

Also thanks to @CPMP, his success on WTF with CNN encouraged me to peruse

I wont introduce the basic model architecture. Because sjv already posted on WTF and this competition.

Please refer to https://www.kaggle.com/c/favorita-grocery-sales-forecasting/discussion/47529 https://www.kaggle.com/c/web-traffic-time-series-forecasting/discussion/39370#220679

I just post my understanding and my experience with this model.

I think the model and the way to training model is naturally robust.

Feeding data with mini-batch with randomly sampled 128 sequence. Then randomly to choose the start of decode/target date. So, we could say somehow the model will see different data for each training iteration. because the total dataset is around 170000 (seq) x 365 days. My feeling is the wavenet trained in this way will be good to handle overfitting.

During the training, I set the learning rate as 0.0005 and with a learning decay.

In fact, the train loss is hard to get decrease after some iteration.

We take the approach for validation is step by step. I keep the last 16 days validation data. which is different with original solution.

onpromotion data, one of big improvement on original model is to encode shifted onpromotion data into decoding part. At the beginning, I only use unit_sale and onpromotion sequence as encoding and decoding features. Trigger by WTF top solution, the shifted unit_sales and onpromotion data could be in encoding part. and shaifted onpromotion data will also be in decoding part.

In general, we transferred the question from based on past sequence pattern. what is your prediction for unit sales. to Not only based on past sequence pattern, also based on in a short future onpromotion information what is your prediction for today's prediction. this boost our Wavenet model 0.002 on public leader board

the way we feeding data gives us robust. but I think there is a little drawback. which is the prediction from model will be slightly 'overfitted' the last several hundreds or thousands of mini-batch data. So the result has a little fluctuation, not very stable. This experience also shown in Top solution in WTF. Also CPMP mentioned by different cv setting he got different performcen.

The best way we thought is averaging. In WTF top solution 30 models prediction averaging. We trained the model and after 5000 mini-batch, model will generate a prediction for every 2000 mini-batch.

The first best model which score 0.508 on public leader board which is 5 model average from iteration 90000 to 98000. but if I try other 5 or more model predictions the performance will be slightly dropped on public leader board. So I tried to use exponential moving average on prediction. history_prediction(t) = history_prediction * beta + current_prediction * (1-beta). to find the best beta. We use Local CV. Finally, we start from iteration 5000 until 121000 and use beta=0.95 to get EMA prediction.

If we do not do Local CV. just blind training. and after 80000, randomly to choose 5 prediction average. the final result might still be in gold zone.

New Item processing, it might only has minor impact, but because we are hanging on the edge of gold zone, so finally I suggest to do some work on new items. which is agreed by all team member to bet on it for gold medal.

after analysis of the data, I found the on the date 2015-07-01, 2015-09-01, 2015-10-01 and 2016-11-27, there are lots of items start to have unit_sales records. but the hard problem still exists, we do not know in test period most of new items on which day start to have unit_sale. Also on which stores ? By checking onpromotion information, all onpromotion with new items only on 08-30 and 08-31. by reviewing the 'old new items' on 2015-10-01 or 2016-11-27, those some new items is shown onpromotion on that day. So, we bet the most new items will start have unit_sales on 08-30 and 08-31. Before that there might be only a few new item was sold. We do not use model to predict, we only use different value to

validation on 'old new item' on the previous date. The candidate is previous 14 days (in fact, it is -28 to -14 days) same store same class unit sale mean, same store same class onpromotion==1 unit sales mean, 0, 1, etc. by quick checking if on promotion happend, fill zero is worst, fill 1 slightly better, the best is same store same class and onpromotion==1 unit_sales mean is winner. We only predict new items which has onpromotion==1, left 0 for all others. I believe this will make better score. but I guess it can only give u rank up 1 postition.

We also used Max Observation value cap. etc.

Finally, thanks to our team member work together to get this very good result.

Options

Comments (7) Sort by Hotness

Click here to enter a comment...

鲲(China) • an hour ago • Options • Reply 1

Well deserved score!!, the solution is awesome!! thanks for asking me join your team, but I missed the email😟, I am not good at deep learning, and I think this is the most important thing I need to learn in the coming year!

Luck Yu • (2nd in this Competition) • an hour ago • Options • Reply 1

Thanks. My teammates are awesome, but i still think it is kind of pity that we missed cooperation. I notice you since summer too. You always perform good. !

鲲(China) • an hour ago • Options • Reply 0

Yeah, good cooperation between teammates is one of the keys to winning the competition, you guys are awesome! hope to team up in the next time😄!

CPMP • 2 hours ago • Options • Reply 1

Thanks for sharing, glad you found my advice to be of interest. You made the difference with new items, good work on that, and congrats on the result! Prize well deserved.

Luck Yu • (2nd in this Competition) • 2 hours ago • Options • Reply

0

Thanks a lot!.

CPMP • 2 hours ago • Options • Reply

1

You too leveraged NN predictions at various training time, will definitely try that next time!

Lesaffrea • (994th in this Competition) • an hour ago • Options • Reply

0

Very well done and thanks a lot to share :)