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

CPMP

posted in Corporación Favorita Grocery Sales Forecasting 3 hours ago

First of all, I'd like to thank Giba, my team mate, without whom I would not have entered this competition. I also want to thank Kaggle and the host for providing this challenging dataset. Congrats to all gold winning teams and individuals, you did an awesome job here. And there is lots of solution sharing already, which is great. Last but not least, final standing is a great, positive, surprise to us, esp considering we both had the flu.

Our solution is mostly a single NN model that scores 0.515 private and 0.508 public, with a secondary lgb model. Believe it or not, we missed the deadline by few seconds and could not select our second sub ourselves. It scored 0.514 on private LB, maybe we would have advanced a couple of place, but we are very happy with the result anyway. Details below.

Missing values

onpromotion is fully known after 2014-03-31. We ignored the missing values issue by only using data posterior to that. We actually did not use data prior December 1st 2015. We also assumed that any item/store that appears somewhere in train or test data has 0 sales for the dates it does not appear. This seems to be what the data description says. It means that store/item that do not appear in train have zero sales during the train period.

Seasonality

In time series problems there are two key ingredients: seasonality, and cv setting. Here weekly seasonality is extremely strong, yearly seasonality is weak, and there

is a monthly seasonality due to pay day. We dealt with weekly seasonality by using time periods that align on weekdays for everything. We did not find a good way to use the monthly seasonality. And we captured the yearly seasonality by using sales figures from 364 days ago. We used 364 and not 365 to align with weekdays.

CV setting

We selected 2 validation periods plus the test period: (2017-07-15, 2017-07-31), (2017-08-01, 2017-08-15), and (2017-08-16, 2017-08-31). For each period we construct a series of training datasets aligned on weekday, for instance for the first validation period:

```
dataset up to 2017-03-11 predict period of length 16 starting on
2017-03-12
dataset up to 2017-03-18 predict period of length 16 starting on
2017-03-19
dataset up to 2017-03-25 predict period of length 16 starting on
2017-03-26
dataset up to 2017-04-01 predict period of length 16 starting on
2017-04-02
dataset up to 2017-04-08 predict period of length 16 starting on
2017-04-09
dataset up to 2017-04-15 predict period of length 16 starting on
2017-04-16
dataset up to 2017-04-22 predict period of length 16 starting on
2017-04-23
dataset up to 2017-04-29 predict period of length 16 starting on
2017-04-30
dataset up to 2017-05-06 predict period of length 16 starting on
2017-05-07
dataset up to 2017-05-13 predict period of length 16 starting on
2017-05-14
dataset up to 2017-05-20 predict period of length 16 starting on
2017-05-21
dataset up to 2017-05-27 predict period of length 16 starting on
2017-05-28
dataset up to 2017-06-03 predict period of length 16 starting on
2017-06-04
dataset up to 2017-06-10 predict period of length 16 starting on
2017-06-11
dataset up to 2017-06-17 predict period of length 16 starting on
2017-06-18
dataset up to 2017-06-24 predict period of length 16 starting on
2017-06-25
```

```
dataset up to 2017-07-15 predict period of length 16 starting on
2017-07-16
```

We use all datasets but the last to train, and we use the last one for early stopping. We use 10 fold CV on the training periods. The only caveat is to make sure all store/item pairs for a given time series are in the same fold. I used a very similar CV setting in WTF competition.

Feature engineering

All unit sales are clipped to be non negative, then log1p transformed. For each of the above datasets, we use:

- lags over periods similar to @Ceshine Lee starter LGB kernel. We use mean, max, and proportion of zero entries, as well as proportion of promotion days.
- Sales for each of the last 7 days
- Average sales per weekday over the last 8 weeks
- For each of the test/validation days: sales from 364 days earlier, and promotion status.
- Class of the item
- Store

We did not use oil price, item number, and other info in our primary model.

NN Model

A feedforward model with 3 dense layers, relu activation. Class and store are embedded in 4 dimension vectors, and appended at the second level. Let x be the output of the dense layers. Its length is equal to the length of the test or validation period. Last level is a 1D convolution with x, the on promotion input vector, and the product of the onpromotion vector with x:

```
y = Multiply()([x, promo_test_input])
y = Reshape((-1,1))(y)
x = Reshape((-1,1))(x)
z = Reshape((-1,1))(promo_test_input)
x = Concatenate(axis=-1)([x, y, z])
x = Conv1D(1,1, activation='linear')(x)
```

This captures the influence of onpromotion day per day quite well. The target for the model is the sequence of train/validation sales, i.e. 15 or 16 days depending on the period.

Other models

We trained a lgb model using the same features and cv setting, the only difference being the target: we train one model per day, like in @Ceshine Lee starter LGB kernel. Best lgb scores 0.510 on pblic LB and 0.517 on private LB. I don't know how others got better results with LGB than with NN.

Giba created a classifier model to predict when sales are zero or not. Alone this model is useless but it adds when ensembling. I'll let him describe it if need be.

Ensembling

Turns out that what would have been our best sub is an unweighted average of 3 different runs of the NN model using a different sets of training periods, and one lgb model. We tried some limited stacking with a NN with one hidden layer. Giba also looked to how to post process submissions. By default we set to 0 all store/item pairs that do not appear in train, but he found that we should not do it if that pair has a promotion in test. He got a sub that scores 0.514 private that way, but, due to few second glitch we did not not select it... And we did not have time to combine his postprocessing with our best blend either.

Conclusion

Main takeaway from me is once again to ignore public LB feedback. Indeed, that feedback was on the first 5 days of the test period, and it gives zero useful information on what happens for days further in the future. Having some reasonable CV setting was key to estimate how well a model predicts further in the future. I could not say I was expecting to get a result as good as what we got in the end, but I knew we were not overfitting to the first few days of the test period ;)

Let me end with a special notice to Ceshine Lee without whom many would not have fared as well as they did. He shared a great kernel, and he shared very useful tips in the early days of the competition. There should be a sharing gold medal in Kaggle competitions, and he would have won that one here!

Options

Comments (14) Sort by Hotness

Click here to enter a comment...

30CrMnSiA • (28th in this Competition) • 3 hours ago • Options • Reply 2

onpromotion is not fully known after 2014-03-31 !

just look this figure , it obviously the item on promotion from 2017.8.1~2017.8.15

but because someday there is no sales for this item，we lost some information about promotion.

If you use these bias data to train, you will get a result that all items onpromotion have a sale>1.

that is not true in test data.

date	store_nbr	item_nbr	unit_sales	onpromotion
2017-08-02	1	108701	1.098612	True
2017-08-03	1	108701	0.693147	True
2017-08-04	1	108701	1.098612	True
2017-08-05	1	108701	0.693147	True
2017-08-07	1	108701	0.693147	True
2017-08-09	1	108701	1.386294	True
2017-08-10	1	108701	0.693147	True
2017-08-11	1	108701	0.693147	True
2017-08-13	1	108701	0.693147	True
2017-08-14	1	108701	1.098612	True
2017-08-15	1	108701	1.609438	True

2018-01-16 20-09-34屏幕截图.png (11.76 KB)

CPMP • 3 hours ago • Options • Reply

0

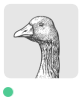
Not sure why you say that. Running this code:

```
import pandas as pd

train = pd.read_csv('../input/train.csv', usecols=['date',
'onpromotion'])

train.onpromotion.fillna(-1, inplace=True)

train.groupby('onpromotion').date.max()
```

yields

```
onpromotion
-1       2014-03-31
False    2017-08-15
True     2017-08-15
Name: date, dtype: object
```

What am I missing?



mykper • (9th in this Competition) • 3 hours ago • Options • Reply

0

Records with zero unit sales

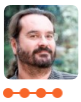
30CrMnSiA • (28th in this Competition) • 3 hours ago • Options • Reply

0

We don't know the onpromotion information in the records with zero unit sales.

If we fillna with zero unit sales records to False.

Our model will get a result that if a item on promotion , its sales will >1



CPMP • 2 hours ago • Options • Reply

0

> it obviously the item on promotion from 2017.8.1~2017.8.15

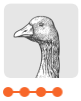
> that is not true in test data.

How do you know both?

Anyway, we did not use that assumption. The way we use onpromotion (in the NN) is to modify on average the sales level. I guess that way was good enough to yield reasonable predictions.

I did play with how to fillna for onpromotion when adding 0 sales rows. Best CV result was obtained with fillna(0). Sure, it is a bias, but it is the best bias I found.

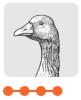
How did you validate your way of filling missing onpromotion values?



30CrMnSiA • (28th in this Competition) • 2 hours ago • Options • Reply

0

So how do you deal with this item promotion information in 2017.8.12



30CrMnSiA • (28th in this Competition) • 2 hours ago • Options • Reply

0

Just check your model's result(not merged with test data) ,all items on promotion's sales >1

I had probed leaderboard , that's not true in test data

A lot items'sales on promotion is 0.

CPMP • 2 hours ago • Options • Reply 0

As I said, when adding zero sales rows we put onpromotion to False. We tried other ways, and they led to worse CV values. How did you validate your way of filling NA? You say medals are not due to good use of CV, but I'm afraid this precisely shows the opposite: we used CV to find out how to best deal with that issue, and it worked reasonably well.

We also probed LB a bit, and found that setting sales > 0 for items that have at least on promotion day was improving LB (both public, and private actually).

30CrMnSiA • (28th in this Competition) • 2 hours ago • Options • Reply 0

If you fillna(0), that caused bias both train and valiation data,

the items'sales on promotion>1

so you will get a good cv.

but not in test data

CPMP • 2 hours ago • Options • Reply 0

We are circling: "How did you validate your way of filling NA? "

re test data:

> We also probed LB a bit, and found that setting sales > 0 for items that have at least on promotion day was improving LB (both public, and private actually).

30CrMnSiA • (28th in this Competition) • 2 hours ago • Options • Reply 0

I also probed LB, i found make some rules to deal with bias ("last 21 days no sales → onpromotopn to False") will get 0.003 booster. The bias can't be handled by local cv. So I used leadboard to validate. Unfortunely , the bias exist between 5days and 11days too.

That's the reason why a lot team drop a lot

you can see sjv‘s post . he used leadboard to validate too.

CPMP • 2 hours ago • Options • Reply 0

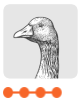
> you can see sjv's post . he used leadboard to validate too.

And he drops in performance too.

Back to this:

> If you fillna(0), that caused bias both train and valiation data, the items'sales on promotion>1 so you will get a good cv.

Now that you insist, I remember I ran lgb at point with the original data as evaluation, i.e. without any additional 0 sales rows. I wanted to make sure that adding these rows was fine. That's how I validated that filling missing onpromotion with 0 was the best I could get. And our limited LB probing confirmed later than it was a reasonable bet.



adityasinha • (54th in this Competition) • 41 minutes ago • Options • Reply

0

Congrats CPMP and Giba for gold medal again. I thought that you would be using your WTF model here, your solution there was great and here insight into data was great. After Caesar I do not have time to participate fully in this competition but giba and you have achieved quite a lot in short duration..

CPMP • 34 minutes ago • Options • Reply

0

Thanks! I did try to reuse my WTF model but made a mistake in it, and lost a week trying to get it running correctly. Then we restarted from scratch, but ended up with a model similar to that of WTF, except for the last convolution layer.

Don't mention Caesars to me ;) I see people complaining about data preparation here, but there it was an order of magnitude worse, at least... Not to mention the reset when I was fighting for top spot... Very bad memories to me. Almost made me quit Kaggle.

I feel better now of course with a gold medal :)