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 · 2 days ago

Eureka
1st place

1st place solution

posted in Corporación Favorita Grocery Sales Forecasting 20 hours ago

53

Congrats to all winner teams and new grandmaster sjv. Thanks to kaggle for hosting and Favorita for sponsoring this great competition. Special thanks to @sjv, @senkin13, @tunguz, @ceshine, we build our models based on your kernels.

- https://github.com/sjvasquez/web-traffic-forecasting/blob/master/cnn.py
- https://www.kaggle.com/senkin13/lstm-starter/code
- https://www.kaggle.com/tunguz/lgbm-one-step-ahead-lb-0-513
- https://www.kaggle.com/ceshine/lgbm-starter

Like the Rossmann competiton, the private leaderboard shaked up again this time. I think luck is on our side finally.

Sample Selection

we used only 2017 data to extract features and construct samples.

train data：20170531 - 20170719 or 20170614 - 20170719， different models are

In fact, we tried to use more data but failed. The gap between public and private leadboard is not very stable. If we train a single model for data of 16 days, the gap

will be smaller(0.002-0.003).

Preprocessing

We just filled missing or negtive promotion and target values with 0.

Feature Engineering

1. basic features
    - category features: store, item, famlily, class, cluster...
    - promotion
    - dayofweek(only for model 3)
2. statitical features: we use some methods to stat some targets for different keys in different time windows
    - time windows
        - nearest days: [1,3,5,7,14,30,60,140]
        - equal time windows: [1] * 16, [7] * 20...
    - key：store x item, item, store x class
    - target: promotion, unit_sales, zeros
    - method
        - mean, median, max, min, std
        - days since last appearance
        - difference of mean value between adjacent time windows(only for equal time windows)
3. useless features
    - holidays
    - other keys such as: cluster x item, store x family...

Single Model

- model_1 : 0.506 / 0.511 , 16 lgb models trained for each day source code
- model_2 : 0.507 / 0.513 , 16 nn models trained for each day source code
- model_3 : 0.512 / 0.515，1 lgb model for 16 days with almost same features as model_1
- model_4 : 0.517 / 0.519，1 nn model based on @sjv's code

Ensemble

Stacking doesn't work well this time, our best model is linear blend of 4 single models.

final submission = 0.42*model_1 + 0.28 * model_2 + 0.18 * model_3 + 0.12 * model_4

public = 0.504 , private = 0.509

Options

Comments (12) Sort by Hotness

Click here to enter a comment...

plantsgo • (4th in this Competition) • 12 hours ago • Options • Reply 0

Congrats.Thanks for sharing.It seems I missed "days since last appearance".

Dan Ofer • (601st in this Competition) • 15 hours ago • Options • Reply 1

That's a cray batch size (65K). Subtracting the mean is also a good trick, i'd tried it but didn't get anything better.

CPMP • 20 hours ago • Options • Reply 5

Thanks for sharing! And congrats on winning the competition. You say luck, but it is also due to this: We just filled missing or negtive promotion and target values with 0.

CPMP • 19 hours ago • Options • Reply 2

Downvoted? LOL.

I bet that we can get a very strong correlation between how team performance evolved between public and private LB, and how they dealt with onpromotion for missing 0 sales rows in train.

Again, a fact hard to cope with for my courageous and anonymous downvoting friends here ;)

FabSchreiber • (177th in this Competition) • 20 hours ago • Options • Reply

0

Congrats Eureka on winning and thanks for sharing some code. That allows newbies like me to learn a lot!

yanglu • (129th in this Competition) • 12 hours ago • Options • Reply

0

I'd tried LSTM-256, Densen512-256-64, not success.Finally , I haven't use LSTM model. Your team is very great!

THLUO • 10 hours ago • Options • Reply

0

Congrats.Thanks for sharing

yyqing • 6 hours ago • Options • Reply

0

Thanks for this sharing. I want to know how do you decide the weight(0.42 0.28 0.18 0.12) in the ensemble ?

Eureka • (1st in this Competition) • 5 hours ago • Options • Reply

1

In fact , we did a two-stage linear blending and all weights are decided based on both cv and public score.
submission = 0.7 * (0.6 model_1 + 0.4 model_2) + 0.3 * (0.6 model_3 + 0.4 model_4)

yyqing • 5 hours ago • Options • Reply

0

Thank you! Beginner learned two-stage linear blending here !

W.JM • (512th in this Competition) • 6 hours ago • Options • Reply

0

Congrats

Pepe Bawagan • an hour ago • Options • Reply

0

Congratulations!