Featured Prediction Competition

Corporación Favorita Grocery Sales Forecasting

Can you accurately predict sales for a large grocery chain?

$30,000
Prize Money

Corporación Favorita · 1,693 teams · 16 hours ago

Nicolas

posted in Corporación Favorita Grocery Sales Forecasting an hour ago

The 6th place solution was an ensemble of two solutions that score .517 privately (top 40). Here is my top 40 solution:

Basic Outline

- LGBM Framework which used only 2017 data.
- Over 500 features.
- Fairly reliable cross-validation on 07-26 --> 08/10

Features

These are the most important features. The following order reflects importance; however, by nature of having 16 models, there is no exact order. Unless otherwise noted, this was all done at the item-store level.

- ma_median * isd_avg / isd_week_avg - popularized early on in the competition.
- ma_median - Coupled with binary for whether or not it is equal to 0.
- Day-of-week averages - Different time periods including 7, 14, 28, 56, 112.
- Days since appeared - Difference between the 'start date' of the training cycle and the first date the item showed up in the original train file.
- Quantiles for several different time-spans.
- Whether the item will be onpromotion.

- Simple averages for different time-spans.
- Item-cluster means - The past 5-day mean was a good predictor.
- Future promotional 'sums' - E.g: sum of onpromotion 8/16 through 8/18.
- mean_no_zero_sales - Mean of instances where unit_sales > 0.
- Frequency encoding - calculate the frequency of appearance for items, stores, families and classes (four columns that each sum to 1).

The above predictors consistently showed up as important predictors, or made immediate score advancements. The following predictors definitely helped as a whole:

- One-hot-encoding clusters, states, types, cities and families.
- Weekend and weekday means.
- Staggered mean data (e.g. 8/07 --> 8/14 for the test-set).
- Staggered quantile data.
- Past number of zero sales for different time-spans.
- Past number of promotional days.
- Past promotional sales averages.
- Past day-of-week quantile data.

Final Submission

Mentioned earlier, this is a top-40 solution on it's own. My teammate, Eran, was scoring similarly. He also used LGBM, but we had quite different Feature Engineering. Specifically, we had different top predictors and we generally used different time-spans. We each ran our models about 5 times and created a final ensemble - the 6th place solution at .514 (.507 public LB).

Some Thoughts

- This was a supply-demand problem. In a more realistic setting, knowing what day stores get product for restocking, and potentially Retail Management Software data (i.e. data on when products are being pulled from the backroom) would make prediction a lot easier.
- I did not worry about the onpromotion bias. The data-set is very complex, and 'filling in the blanks' seemed unrealistic. If an item was on promotion, but had 0 unit_sales, then it was either very low demand or out-of-stock.
- The only way to know if a new item-store combination is 'real' is if it was on promotion at some point during the test-set days.

- Having your model data start from 8/9 and predicting 8-16 --> 8-20 can be useful to see how it will do for the private leader board.
- Examining which items were getting predicted badly in the validation set was very useful for coming up with feature ideas.

Options

Comments (0) Sort by Hotness

Click here to enter a comment...