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

Takuya Akiyama
18th place

18th Place Solution - Simple Improvement of Public Kernel

11

posted in Corporación Favorita Grocery Sales Forecasting 8 hours ago

My approach is almost based on this kernel.
https://www.kaggle.com/vrtjso/lgbm-one-step-ahead

You can reach this rank by just improving kernel.

- Private Score : 0.523 -> 0.516
- Public Score : 0.515 -> 0.508

Feature Engineering

In the kernel, most features were made from time series data of each combinations of stores and items. I added some variations of averaging features. Further more, I made similar features about stores and items separately, by

Effective Use of Data

In this competition, we have to predict sales for the next 16 days after the given dataset. When we predict the data at the farthest days (16 days) based on time series features, we can only use data up to 16 days before the predicted date. The kernel used data up to 3 weeks before the prediction target period because of considering above reason and characteristics of the day of the week. However when predicting closer date sales, such as the next day or ten days later, more

recent data can be used. So I used data up to 1 week before prediction date for predicting future 1-7days sales, and data up to 2 weeks before for predicting 8-14days sales. This improve public LB score about .002 (but maybe not affect private LB so much because LB is separated by target dates).

Extending Training Dataset for Predicting Test Set

For validation, the kernel uses a dataset from 26th July to 10th August as validation set, and data before the validation set was used for training models which predict validation set and test set. This validation process is very important to confirm whether or not a model is overfitting. However when predicting test set, it is not good to only use the data that made for predict the validation set because dates for training and predicting dates are far away. So after validation process I retrain model using data which was extended 2 weeks, and predict test set by the model. This model improve public LB score about 0.002.

Kaggle kernel always gives me useful information and stimulation. Thanks for fantastic kernel authors and all competitors!

Options

Comments (0) Sort by Hotness

Click here to enter a comment...