Model documentation

Name: Vadim Sokolov

Location: Moscow, Russia

Email: mr.sokolov.v.v@yandex.ru

Public leaderboard for LightGBM (RMSE): 0.8981

Private leaderboard for LightGBM (RMSE): 0.9120

Public leaderboard for stacking (RMSE): 0.9050

Private leaderboard for stacking (RMSE): 0.9070

Coursera grade: 10/10

1. Background

I am an junior data scientist who has taken many courses on DataCamp, Stepik and Coursera. This year my team won the All-Russian hackathon, became a medalist in the hackathon from rosatom. I like to participate in competitions on Kaggle.

I spent about 5 days in this competition.

2. Summary

- *Pandas, numpy, matplotlib, seaborn, catboost, lightgbm, sklearn and jupyter notebook*
- In *final_project.ipynb* - all code without hyperparameters tuning
- In *hyperparametr_tuning.ipynb* - code of hyperparameters tuning
- In *data folder* – all datasets with finished dataset and submission for lightgbm and stacking models
- In *model folder* – files with all training models
- The most important features is lagged month intervals
- The training method I used is ensemble model

3. Feature selection/engineering

Shop features:

Add new features – city coordinates, categorical encoding city of a shop and country part (0-4) based on the map.

Item features:

Add new features – item category, more common item category.

Basic lag features and mean encoding.

Most important features are the lagged mean encoded values from the categorical data.

4. Training method

In my project I considered models of gradient boosting (catboost, lightgbm), random forest, linear regression and their stacking.

For the future: probably need more careful selection of models for stacking and careful selection of parameters for models. However there was not enough time for this due to a large amount of data.

5. Interesting findings

*Data Leakage**:***

around 42% of training shop_id ~ item_id pairs are present in test set.

Models:

as a single model, lightgbm with tuning parameters turned out to be the best. This model was also better on the public leaderboard compared to stacking. However, the ensemble performed better on the private leaderboard.

6. Model execution time

The whole stage of model development took place on a virtual machine of the Kaggle platform.

How long does it take to train model? **40 minutes.**

How long does it take to generate predictions using model? **2 minutes.**

7. Environment

There is available a requirements.txt file, but in fact the main used tools are:

numpy 1.19.5

pandas 1.2.4

sklearn 0.23.2

matplotlib 3.4.2

seaborn 0.11.1

lightgbm 3.2.1

catboost 0.26