Group Outlying Aspects Mining

Shaoni Wang[1], Gang Li[2]

[1] Xi'an Shiyou University, China
[2] Deakin University, Australia

Introduction

Predict future sales

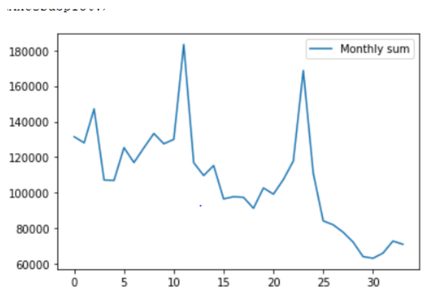
- given: a challenging time-series dataset consisting of daily sales data, kindly provided by one of the largest Russian software firms - 1C Company.
- target: predict total sales for every product and store in the next month
- evaluation: Submissions are evaluated by root mean squared error (RMSE)

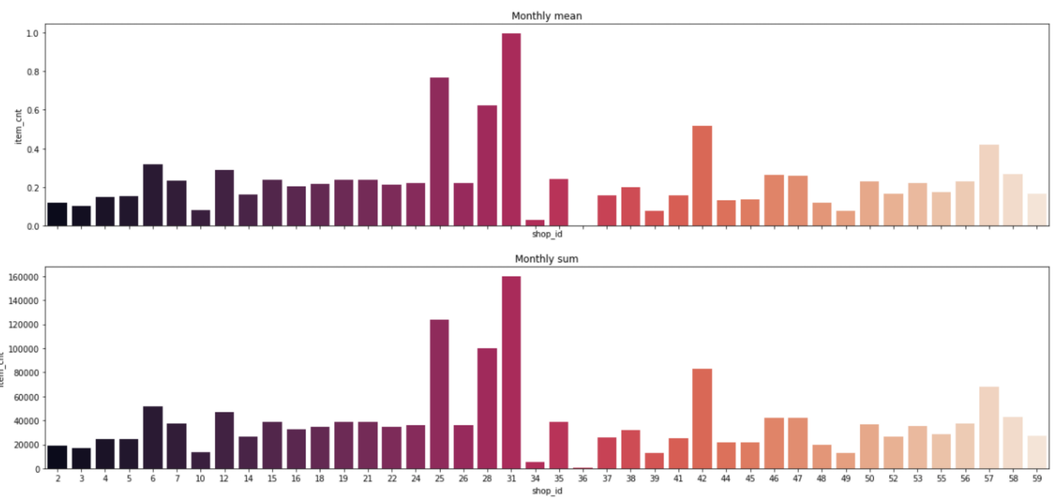
Data Cleaning

Missing Value and Non Value :Finding null and missing values
Cartesian product : Finding those in test set instead of training set
Data leakages : Finding those in training set instead of test set
Data duplication : Finding duplicative items
outdated items :Finding outdated items
Negative :Finding negative values

Data analysis

- Monthly sales of goods
- Shop sales
- Sales of different category
- Item and Shop Information





decision tree

In machine learning, decision tree is a prediction model, which represents a mapping relationship between object attributes and object values. Each node in the tree represents an object, and each branch path represents a possible attribute value, while each leaf node corresponds to the value of the object represented by the path from the root node to the leaf node. The decision tree has only a single output, if you want to have complex output, you can establish an independent decision tree to deal with different outputs. Decision tree is a frequently used technology in data mining, which can be used to analyze data, and also can be used for prediction.

Candidate algorithm

- GBDT
- Xgboost
- lightgbm
- neural network

future selection

Data feature 'dateblocknum', 'shopid', 'itemid', 'itemcategoryid', 'cattypecode', 'catsubtypecode', 'shopcitycode', 'shoptypecode'

- item_cnt_month
- date_avg_item_cnt
- date_item_avg_item_cnt
- date_shop_avg_item_cnt
- date_cat_avg_item_cnt
- date_cat_shop_avg_item_cnt
- date_type_avg_item_cnt
- date_item_type_avg_item_cnt
- date_city_avg_item_cnt

delay:1,2,3,6,12

Conclusion

$valid'_1 srmse$: 0.880256

TULIP Team for Universal Learning and Intelligent Processing

The 11^{th} International Conference on Knowledge Science, Engineering and Management (KSEM 2018), 17-19/08/2018, Changchun, China