Flip 00 Project Report

Huanhuan Ge
Qingdao University of Technology, China

Introduction

This is a time-series problem. Many information are given about daily sales data. And the goal is to to predict total sales for every product and store in the next month.The raw datasets contain six files, which attributes are shown below.

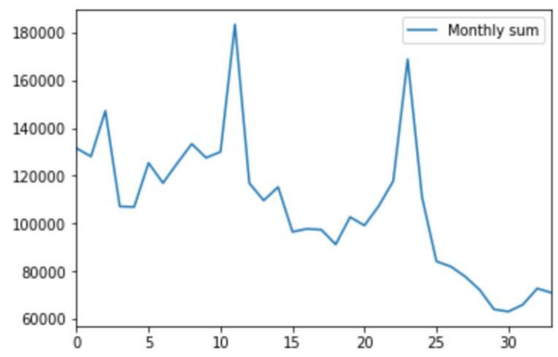
Name	Attirbute
sales_train.csv	date,date_block_num,shop_id,item_id,item_price, item_cnt_day
test.csv	ID,shop_id,item_id
items.csv	item_name,item_id,item_category_id
shops.csv	shops_name,shops_id
item_categories.csv	item_categories_name,item_categories_id
sample_submission.csv	ID,item_cnt_month

Data processing

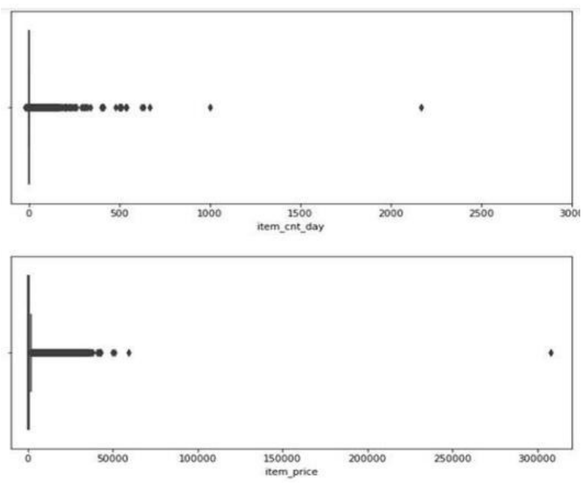
- Remove missing value and NaN value
- Filter outliers and duplicate Data
- Process shops sets(encode and modify the ID)
- Process items sets(modify the ID)
- Process categories sets(encode and modify the ID)
- Sales analysis(closed shops and discontinued items)

Data Visualization

Monthly sales with all items over time, which are able to show some information about historical feature.



The outliers is obvious by using box plot.

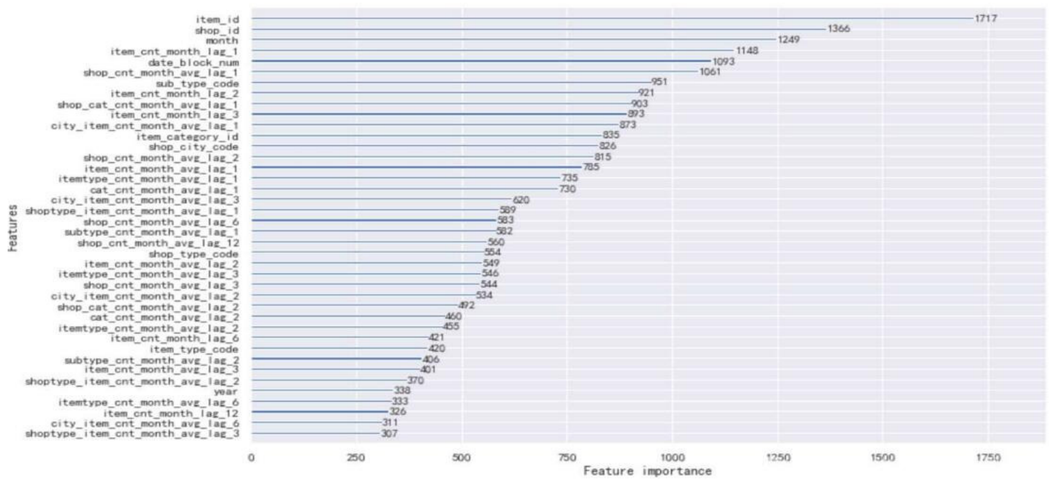


Feature Selection and Feature Importance

- Data features which are given
  shop_id, item_id, item_cnt_month, shop_type_code, shop_city_code, year, month, item_category_id, item_type_code, sub_type_code

- Monthly sales features
  average monthly sales of items
  average monthly sales of shops
  average monthly sales of categories
  average monthly sales of types and subtypes average monthly sales of shops cityitem
  average monthly sales of shops typeitem

- Historical features
  historical delay:1,2,3,6,12
- Feature importance



Modeling and Result

- Model:Lightgbm
- Score:0.93740(RMSE)
- Rank:3027/8738

Conclusion

Compare to midterm presentation,RMSE decreased from 1.04885 to 0.93740. The reason mainly is more features and different modeling. In the figure of feature importance and monthly sales, some historical feature play an important role. And lightgbm model's and xgboost model's result have a marginal difference, but processing speed of lightgbm model is faster.