Group Outlying Aspects Mining

Shaoni Wang[1], Gang Li[2]

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

Introduction

Kaggle Tabular Playground Series - Jan 2022, this topic gives two stores in three The daily sales volume of three products in different countries from 2015 to 2018 requires us to forecast them Sales volume in 2019. Data set: the title gives a 26298 line × 6-column training set, a 6570 row × 5-column test set and one Submit samples. The training set includes sales data of each date-country-store-commodity combination. date From 2015 to 2018, there are three countries, two stores and three commodities. Test set compared with training set Lack of sales volume. The training set header is as follows:

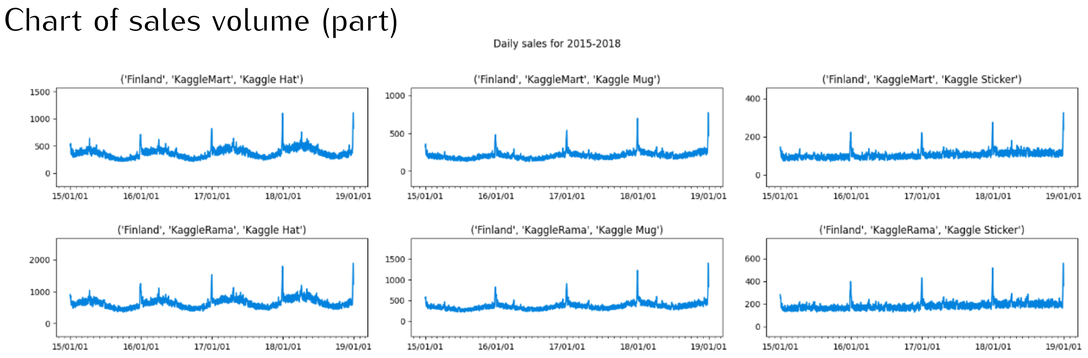
' date	country	store	product	num sold
2015/1/1	Finland	KaggleMart	Kaggle Mug	329
2015/1/1	Finland	KaggleMart	Kaggle Hat	520
2015/1/1	Finland	KaggleMart	Kaggle Sticker	146

Data Analysis

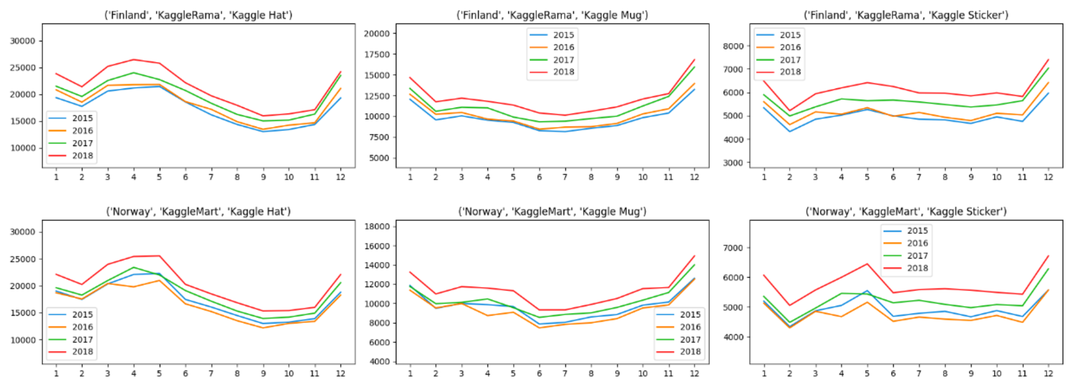
- After analyzing the data set, we can find that three countries, two stores and three products can

  The combination training data in composition 18 covers the period from 2015 to 2018. Let's first look at each combination

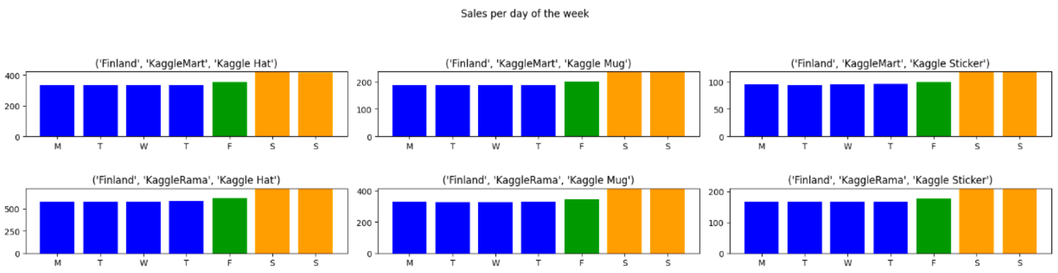
  Chart of sales volume (part)



- At the same time, view the monthly sales figure.



- View daily sales of the week.



Data Processing

- Use Pandas database to operate the data, add GDP information, and add seasonal indicators every week. Unique coding for commodities, countries and stores. At the same time, Fourier feature is added.

	gdp	wd2	wd3	wd4	wd5	wd6	wd7	Finland	Norway	KaggleRama	...	hat_sin1	hat_cos1	sin2	cos2	mug_sin2	mug_cos2
0	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.000000e+00	0.000000	3.442161e-02	0.999407	3.442161e-02	0.999407
1	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	1.721336e-02	0.999852	3.442161e-02	0.999407	0.000000e+00	0.000000
2	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0	...	0.000000e+00	0.000000	3.442161e-02	0.999407	0.000000e+00	0.000000
3	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	...	0.000000e+00	0.000000	3.442161e-02	0.999407	3.442161e-02	0.999407
4	5.457200	0.0	0.0	1.0	0.0	0.0	0.0	1.0	0.0	1.0	...	1.721336e-02	0.999852	3.442161e-02	0.999407	0.000000e+00	0.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26293	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	-2.449294e-16	1.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000
26294	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	-0.000000e+00	0.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000
26295	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	-0.000000e+00	0.000000	-4.898587e-16	1.000000	-4.898587e-16	1.000000
26296	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	-2.449294e-16	1.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000
26297	6.319788	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	...	-0.000000e+00	0.000000	-4.898587e-16	1.000000	-0.000000e+00	0.000000

26298 rows × 26 columns

Modeling and Result

Models: linear regression
Public score: 6.48662
Rank: 498/1543

Conclusion

The data structure is relatively simple, so the linear regression model is used for prediction. It is found that the loss value has relatively large deviation at some points, which may be related to the Spring Festival.