MAFS 6010Z Project 3

M5 Forecasting – The prediction for Walmart's Accuracy

Group members:

LIAO Xinzhen (20813829), MA Rui (20736954),

XU Yuan (20801498), SHI Yiyuan (20745230)

Youtube link :

https://youtu.be/46LtAz_H1pU

Contents

01

Project Introduction

Walmart

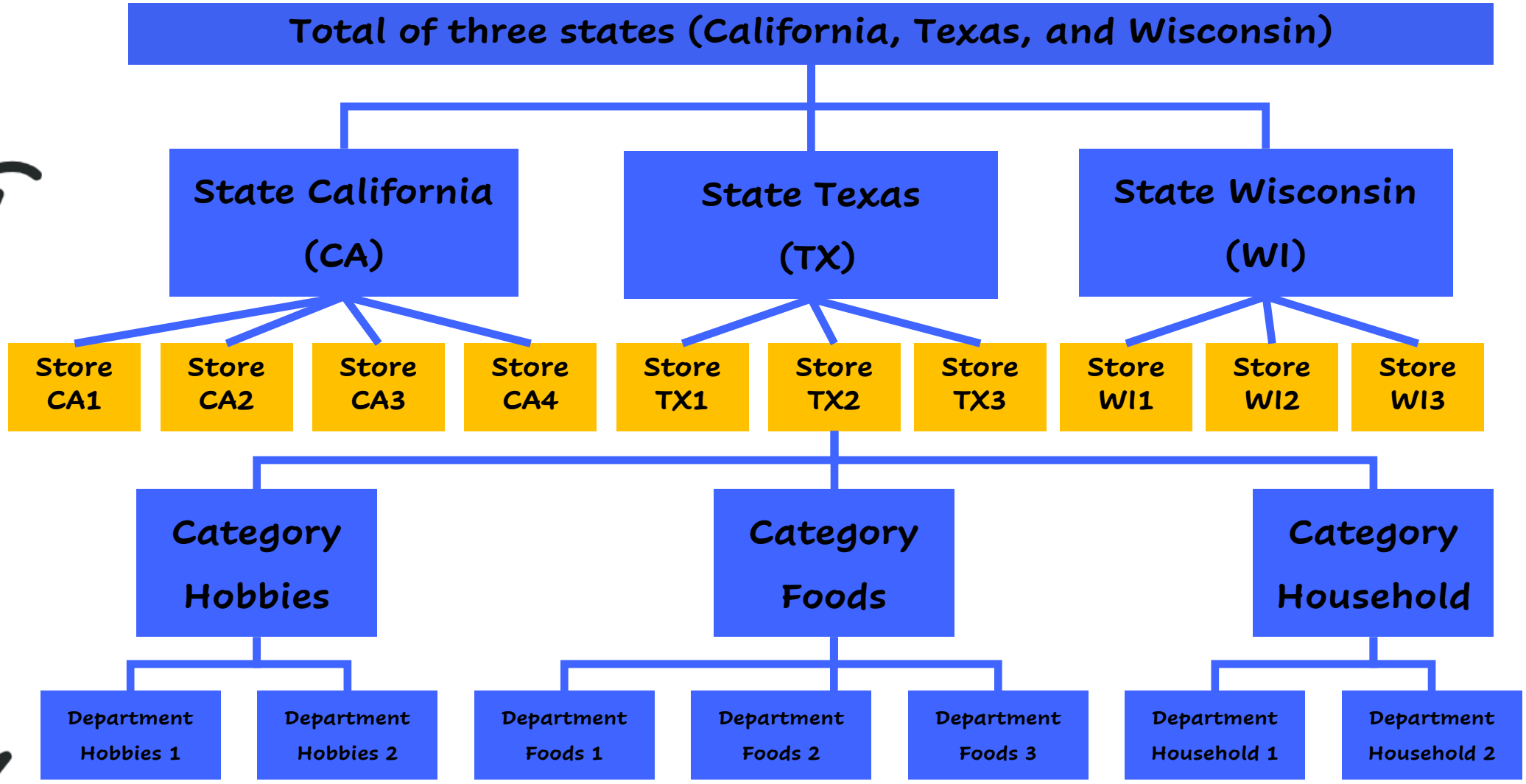
Have Fun

Even though Walmart recently divested from several markets, like Brazil, United Kingdom and Japan, it remains the No.1 in the top 50 global retailers.

Historically, Walmart successfully went to public in 1972 and further entered China in 1996.

In 2018, Walmart topped in the Fortune 500 list with a revenue of US$485.8 billion, which is the fifth consecutive year that Walmart has been ranked as the top 500 company in the world.

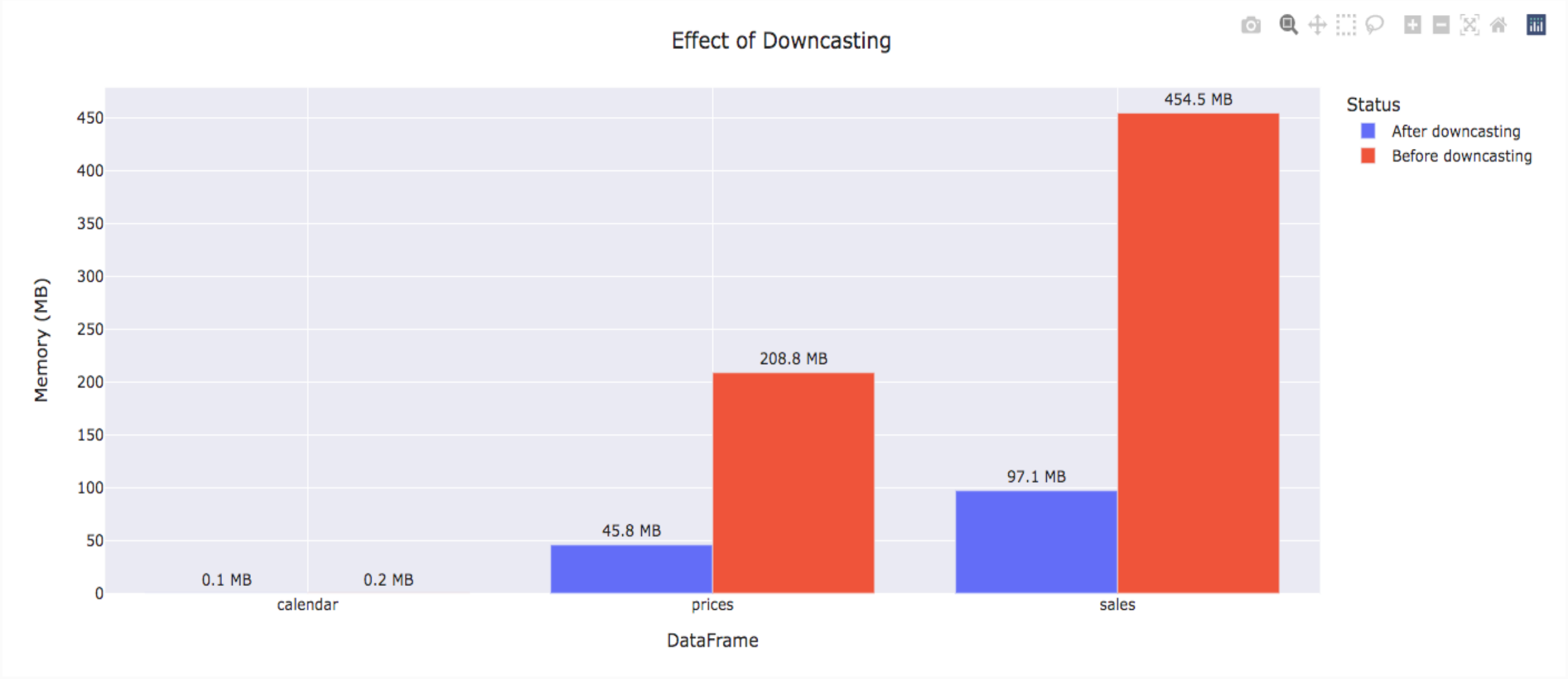
Data Structure



02

Data

Data Downcast

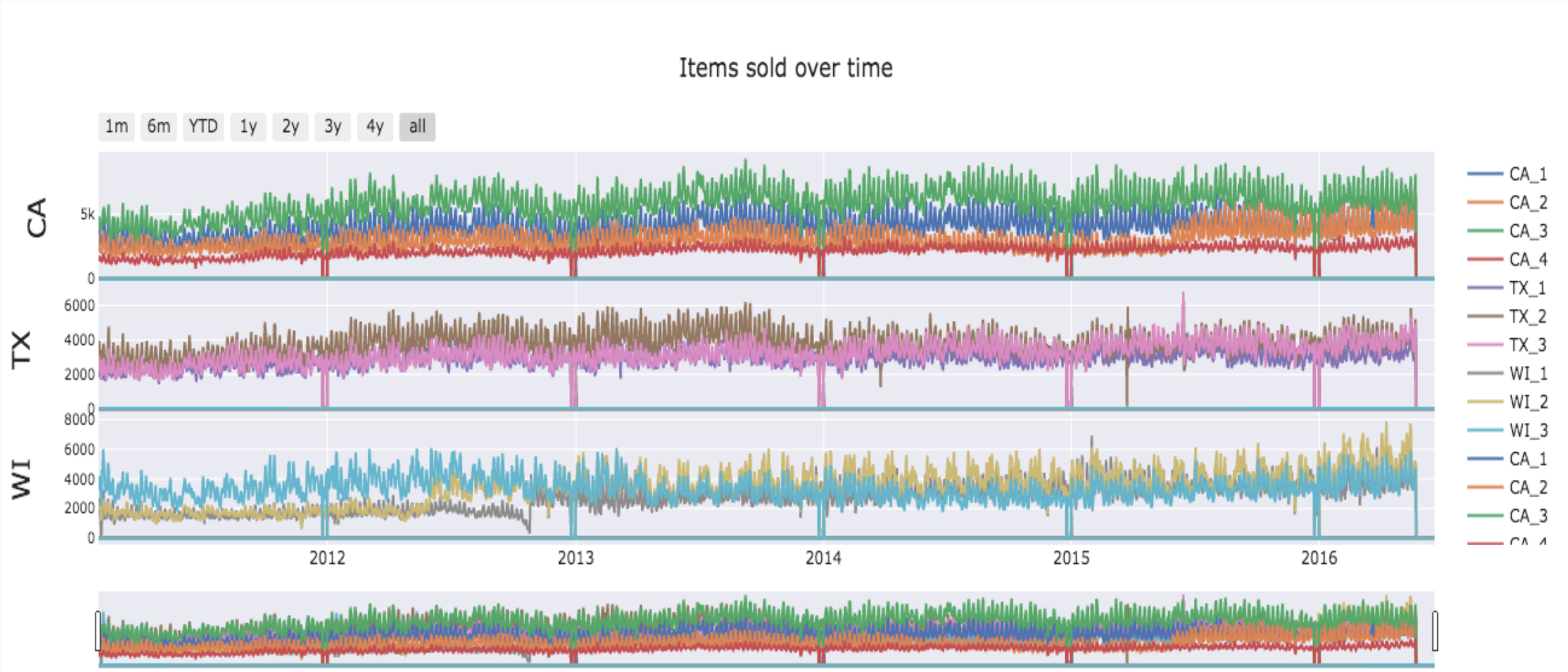


Data View –Distribution of items prices among stores

Data View –Item Prices Between Categories

Data View –Number of Items Sold Among Stores

Data View –Items Sold Over Time



03

LightGBM Model

A

- Training
- Less than 1914

B

- Validating
- From 1914 to 1942

C

- Testing
- Greater than 1942

Data Split

Our group divides the samples into training, validation and testing samples. The samples of days before 1914 are used to train the model, and the samples of days from 1914 to 1942 are used for validating. The remaining samples of days greater than 1942 includes out-of-sample test data sets.

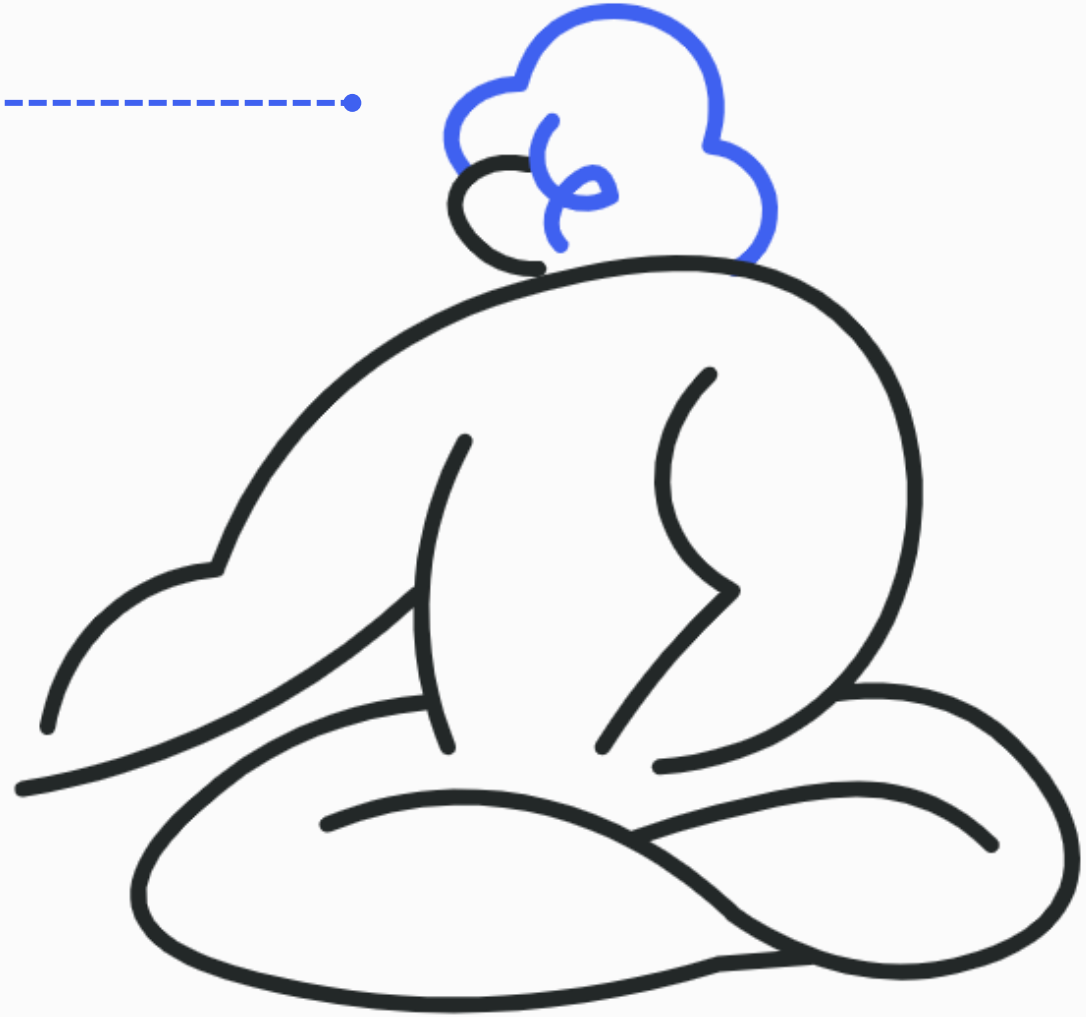
LightGBM Model

framework that implements the GBDT algorithm

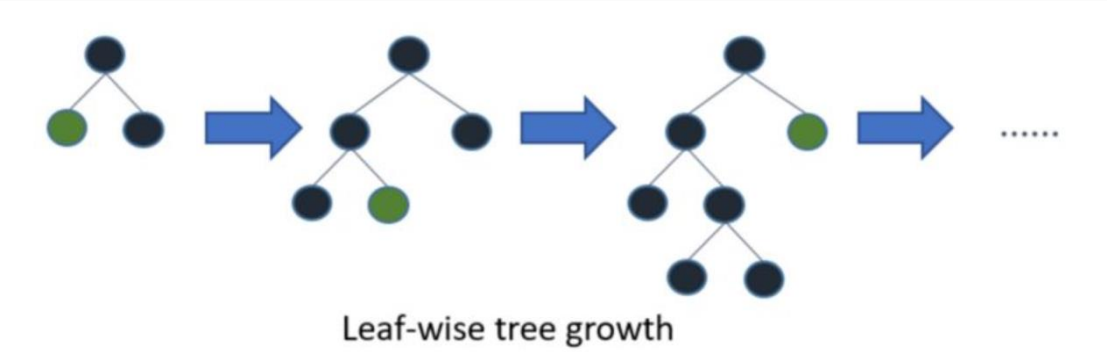
LGBM

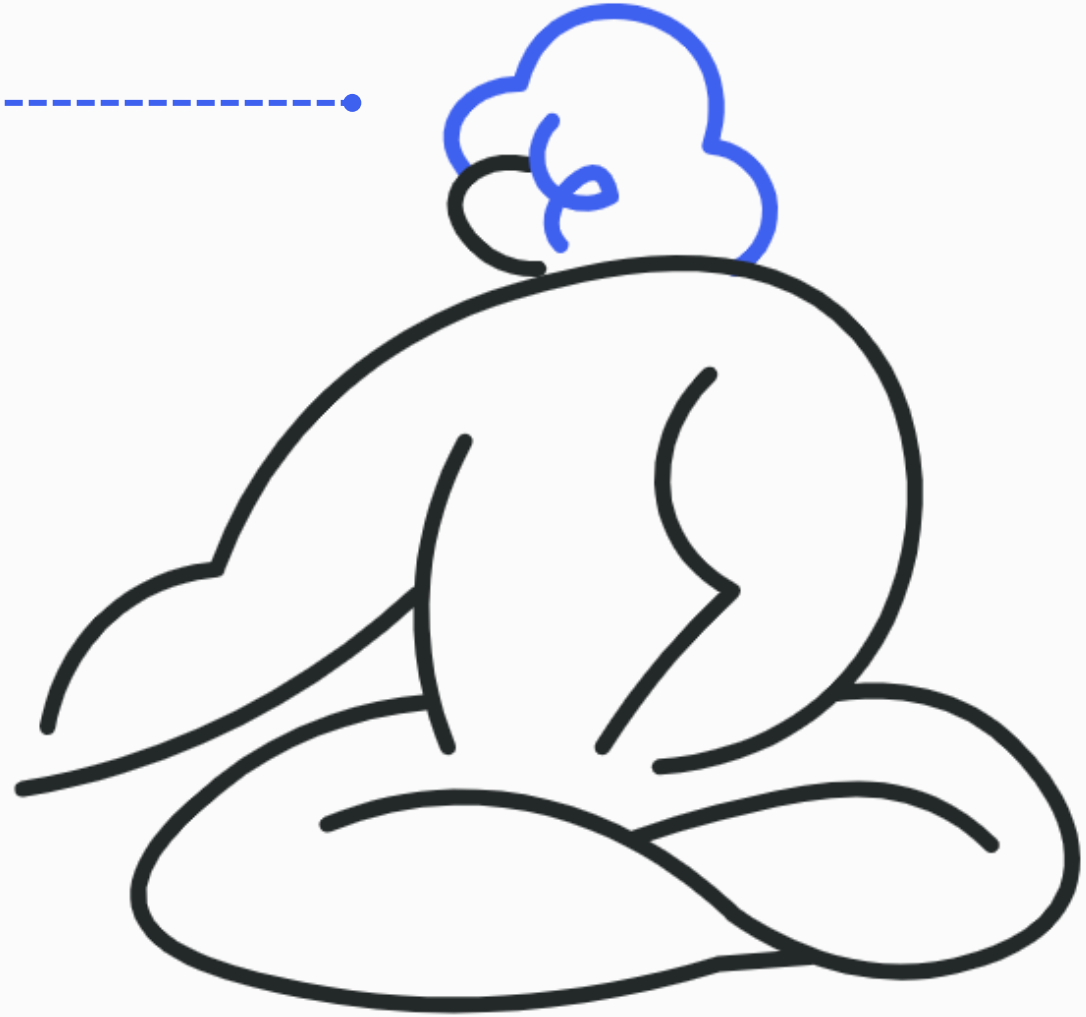
Different from Other Tree-based Algorithms

The growth method of LightGBM tree is vertical, and other algorithms are horizontal, which means that Light GBM grows the leaves of the tree, and other algorithms grow the level of the tree. LightGBM selects the leaves with the largest error for growth. When the same leaves are grown, the leaf-growing algorithm can reduce more loss than the layer-based algorithm.



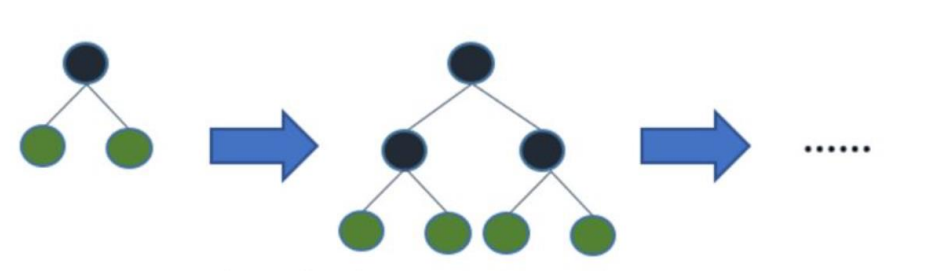
Different from Other Tree-based Algorithms

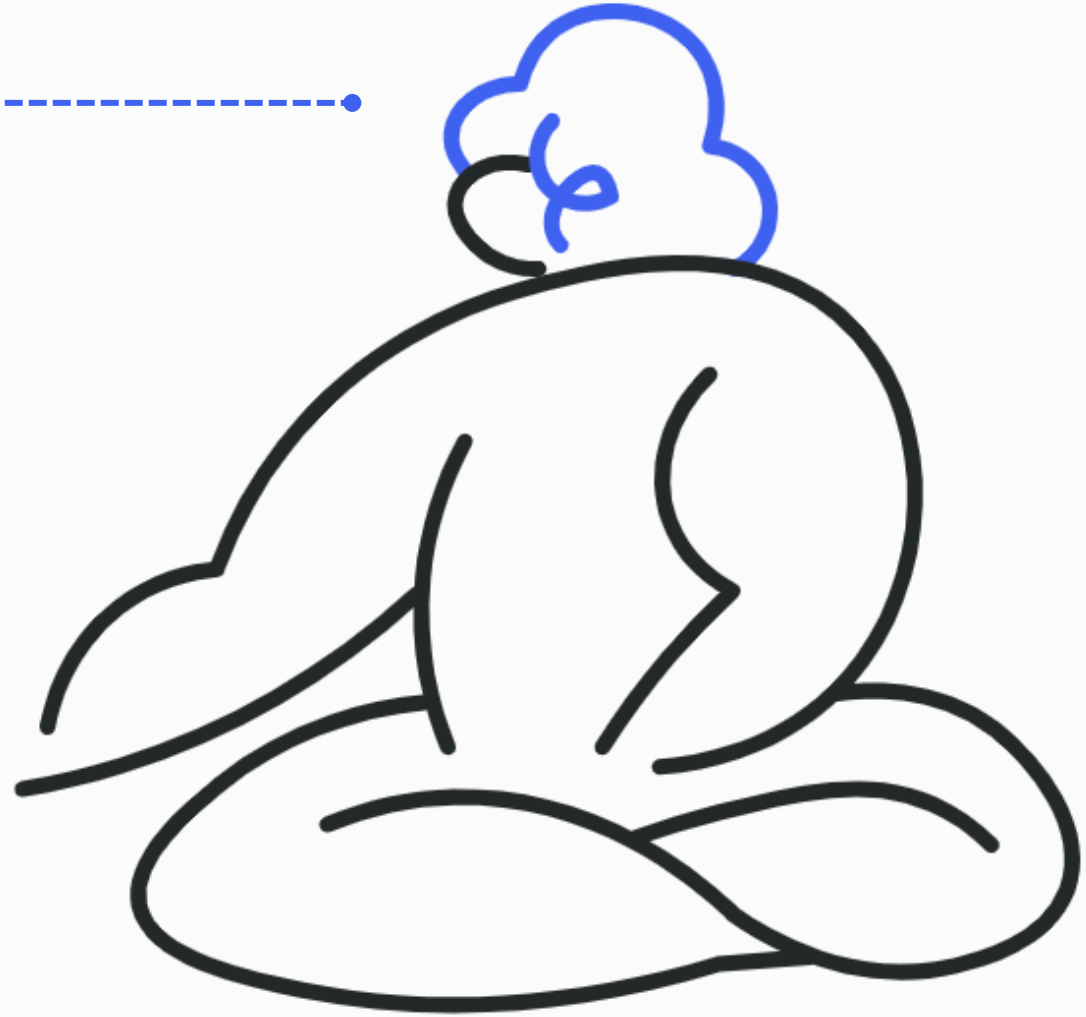




- How LightGBM works:

Different from Other Tree-based Algorithms





- How other boosting algorithm works:

Suitable Data Set and Advantages

Have Fun

LightGBM can handle a large amount of data and occupies very little memory when running. LightGBM focuses on the accuracy of the results and also supports GPU learning.

Core Parameter

n_estimators

learning_rate

subsample

colsample_bytree

max_depth

num_leaves

min_child_weight

04

Model Peformance

Parameter Adjustment

01

learning_rate

When we keep the other parameters the same, the forecast score of 0.03 learning_rate is less than that of 0.3 learning_rate.

02

n_estimators

When we compare and adjust the parameters between 500 and 1000, we find that the fit score levels of 500 and 1000 are not much different, so we choose 1000 as our n_estimators.

03

max_depth

Select in 4, 8 and 12 to prevent setting a too large value from causing serious over-fitting. max_depth is set to 4, which has relatively good performance.

04

num_leaves

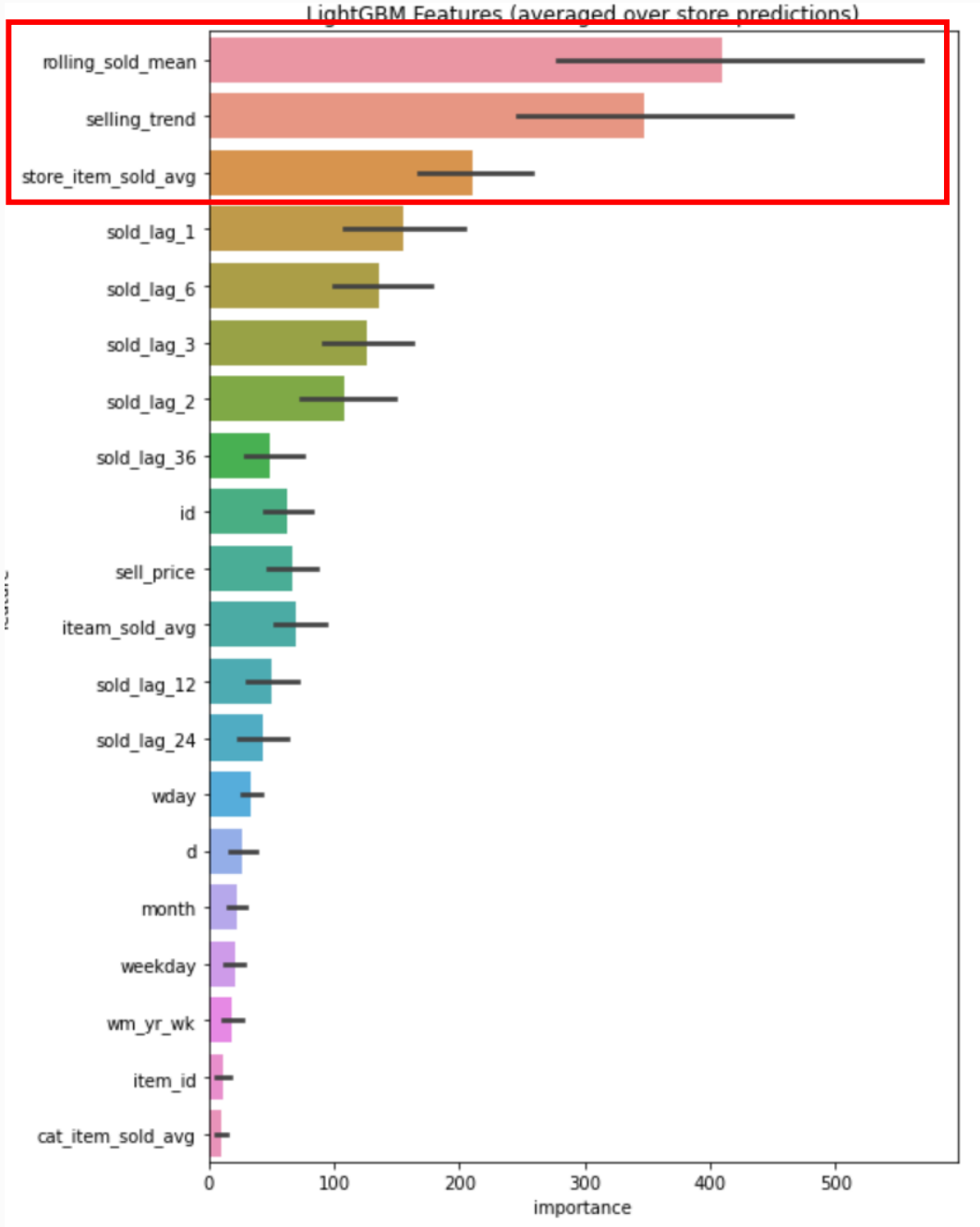
While keeping other parameters as ideal, setting num_leaves to 50 and 100 has little effect on the result, so choose a smaller num_leaves to match the smaller max_depth.

04 Parameter Adjustment

n_estimators	learning_rate	subsample	colsample	max_depth	num_leaves	min_child_weight	score
1000	0.03	0.8	0.8	4	100	300	5.35417
1000	0.3	0.8	0.8	4	50	300	5.39065
500	0.3	0.8	0.8	8	50	300	5.35651
1000	0.3	0.8	0.8	12	50	300	5.35847
1000	0.3	0.8	0.8	12	100	300	5.36140
1000	0.3	0.8	0.8	4	50	300	5.35912
1000	0.3	0.8	0.8	4	100	300	5.35912

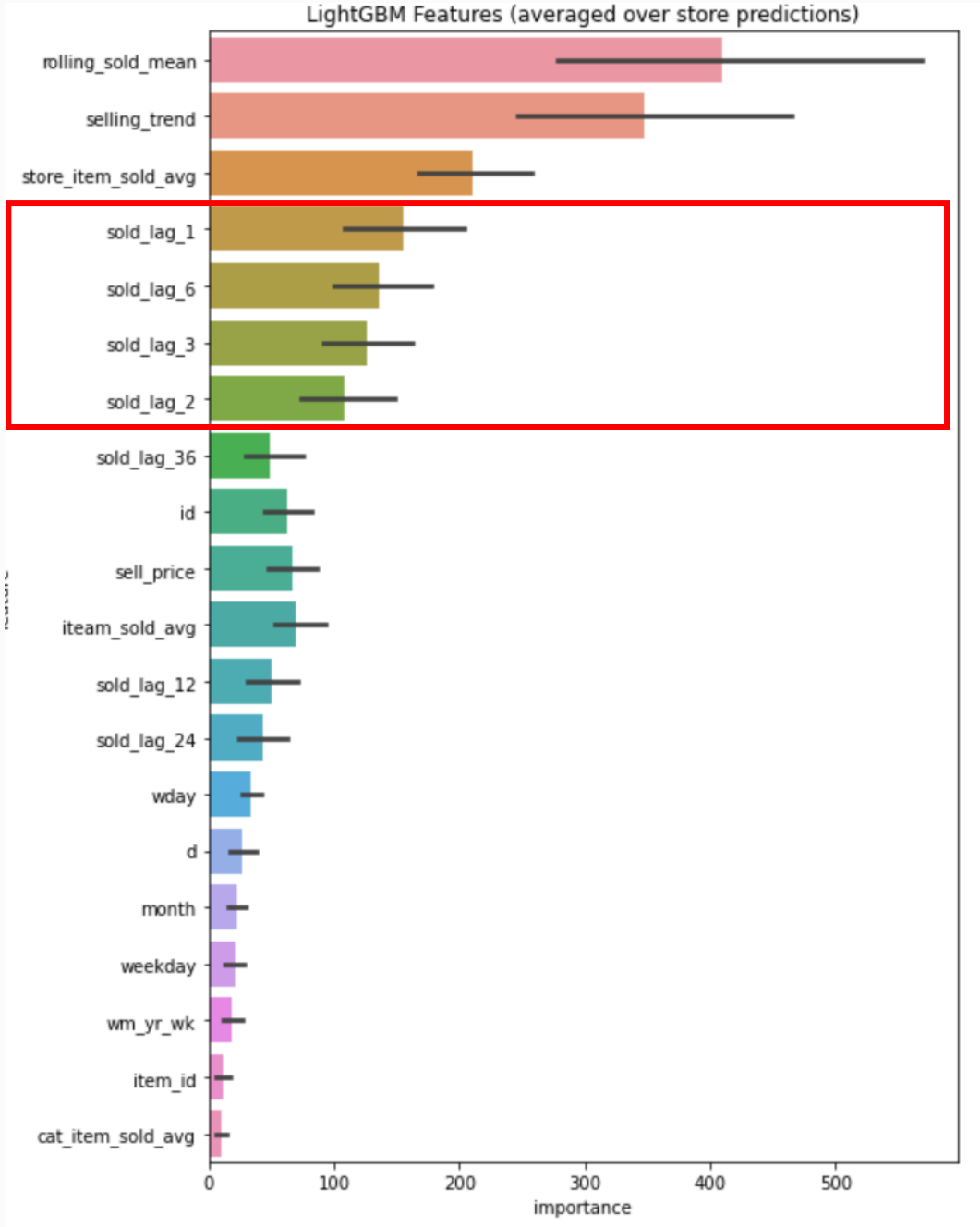
Feature Importance

From the evaluated feature importance results, rolling_sold_mean is the weekly average, selling_trend is the difference between the daily average and the corresponding average of the total number of days of sales, and store_item_sold_avg is the average sales of different products in different stores. These three featuers are very important for the prediction effect of the LGBM model.



Feature Importance

Besides, sold_lag_1, sold_lag_2, sold_lag_3, and sold_lag_6 have significant importance for model prediction, which is also in line with our common sense that the closer to the sales on the forecast day, the better the forecast performance of the sales on the forecast day.



Model Performance

After the above tuning process, the parameters we selected for the LGBM model are:

n_estimators=1000,

learning_rate=0.3,

subsample=0.8,

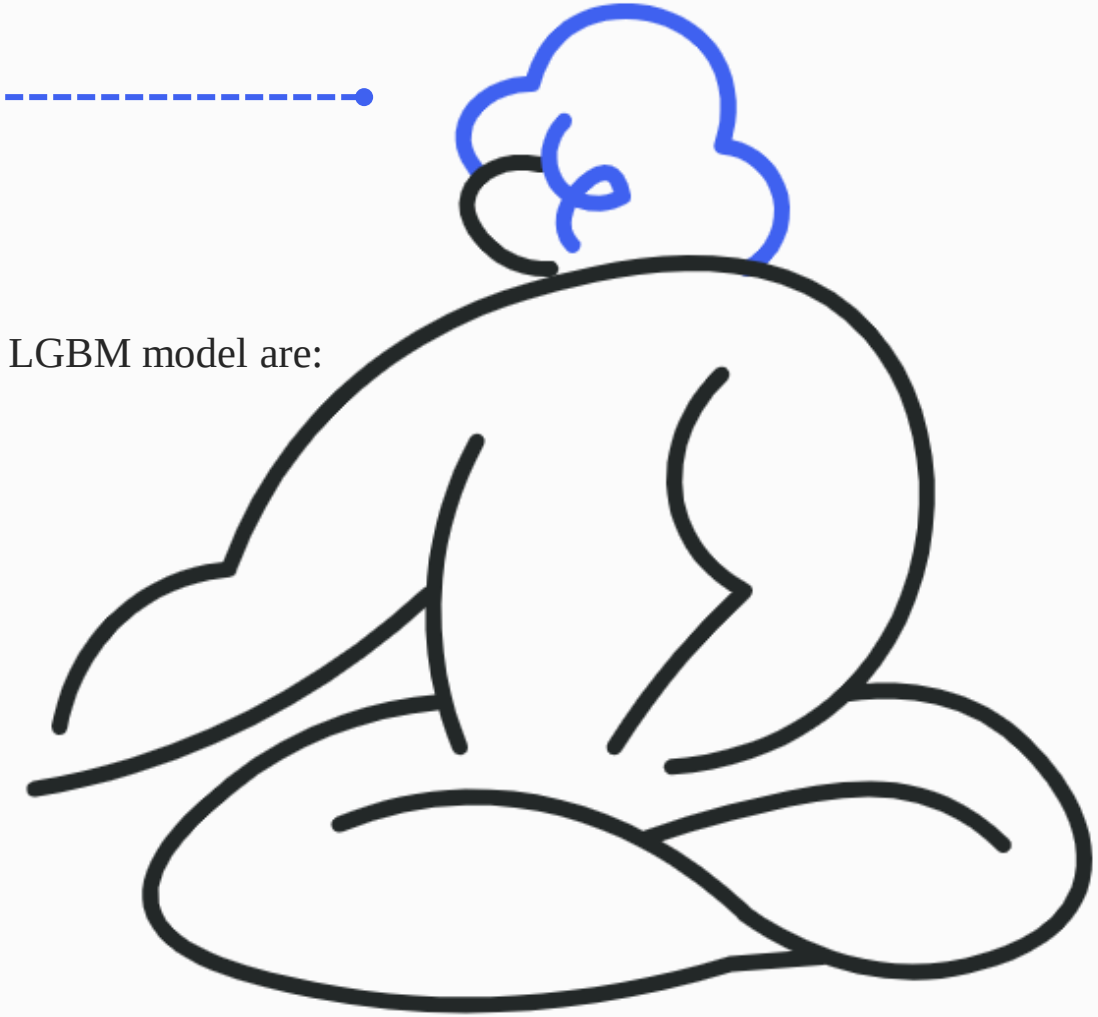
colsample_bytree=0.8,

max_depth=4,

num_leaves=50,

min_child_weight=300,

the final prediction score for the next 28 days is 5.39065.



04

Model Performance

	id	F1	F2	F3	F4	F5	F6	F7	F8	F9	...	F19	F20	F21	F22	F23	F24	F25	F26	F27	F28
0	HOBBIES_1_001_CA_1_validation	0	0	0	2	0	3	5	0	0	...	2	4	0	0	0	0	3	3	0	1
1	HOBBIES_1_002_CA_1_validation	0	1	0	0	0	0	0	0	0	...	0	1	2	1	1	0	0	0	0	0
2	HOBBIES_1_003_CA_1_validation	0	0	1	1	0	2	1	0	0	...	1	0	2	0	0	0	2	3	0	1
3	HOBBIES_1_004_CA_1_validation	0	0	1	2	4	1	6	4	0	...	1	1	0	4	0	1	3	0	2	6
4	HOBBIES_1_005_CA_1_validation	1	0	2	3	1	0	3	2	3	...	0	0	0	2	1	0	0	2	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
30485	FOODS_3_823_WI_3_validation	0	0	0	2	2	0	0	0	2	...	1	0	3	0	1	1	0	0	1	1
30486	FOODS_3_824_WI_3_validation	0	1	1	1	0	0	0	0	1	...	0	0	0	0	0	0	1	0	1	0
30487	FOODS_3_825_WI_3_validation	0	0	1	1	0	2	1	1	0	...	0	0	1	2	0	1	0	1	0	2
30488	FOODS_3_826_WI_3_validation	1	3	0	1	2	1	0	2	1	...	1	1	1	4	6	0	1	1	1	0
30489	FOODS_3_827_WI_3_validation	0	0	0	0	0	1	1	1	2	...	1	2	0	5	4	0	2	2	5	1

Thanks For Your Reading~