Forecasting the Sales of Walmart

ZHOU Xiaomin 20749212

TIAN Xinyu 20750015

SUN Ke 20747903

HUANG Yuning 20738524

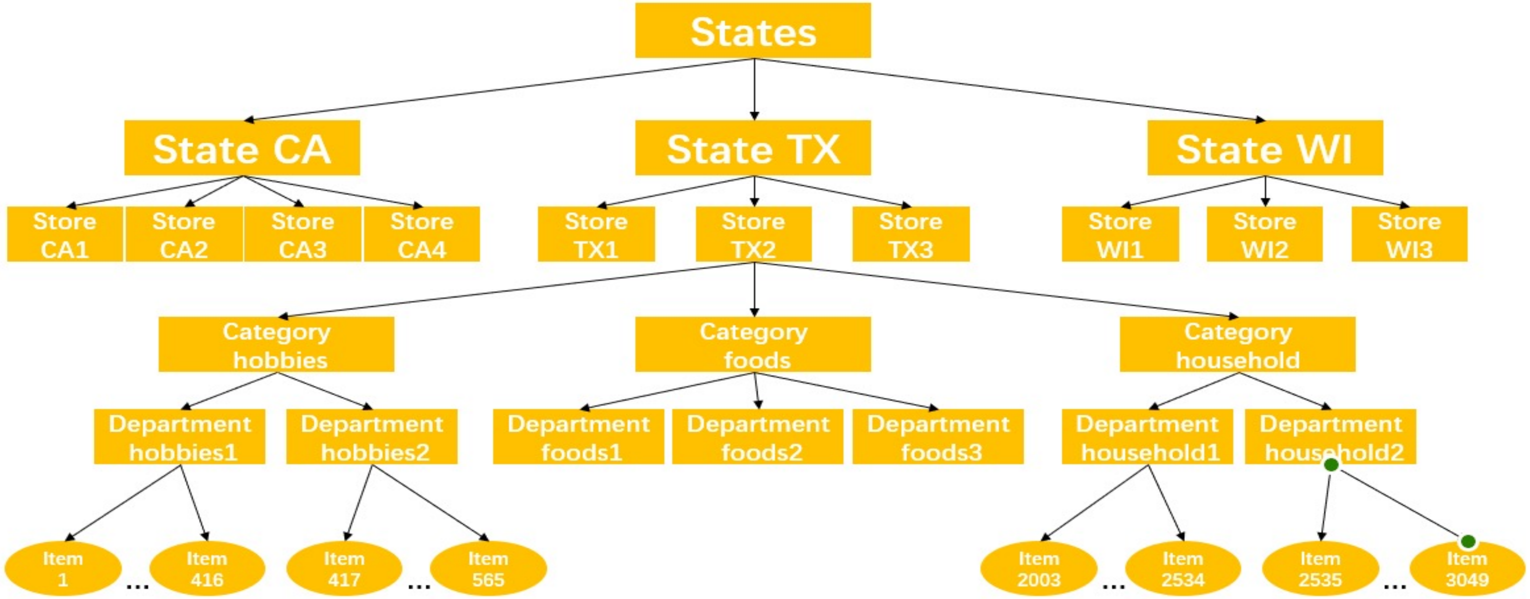
1. Introduction

Introduction – Data Description

CSV *"sales_train.csv"*

- **3 states:** California(CA), Texas(TX), Wisconsin(WI)
- **3 product categories:** Hobbies, Foods, Household
- **3049 products for each store**

Aggregation Level	Number of series
state	3
store	10
category	3
department	7
state & category	9
state & department	21
store & category	30
store & department	70



Introduction – Problem Description

CSV *“calendar.csv” & “sell_price.csv”*

Explanatory variables:

- **date:** The date in a “y-m-d” format.
- **wm_yr_wk:** The id of the week the date belongs to.
- **weekday:** The type of the day (Saturday, Sunday, …, Friday).
- **wday:** The id of the weekday, starting from Saturday.
- **month:** The month of the date.
- **year:** The year of the date.
- **event_name_1:** If the date includes an event, the name of this event.
- **event_type_1:** If the date includes an event, the type of this event.
- **event_name_2:** If the date includes a second event, the name of this event.
- **event_type_2:** If the date includes a second event, the type of this event.
- **sell_price:** The price of the product for the given week/store. The price is provided per week (average across seven days). If not available, this means that the product was not sold during the examined week. Note that although prices are constant at weekly basis, they may change through time (both training and test set).
- **snap_CA, snap_TX, and snap_WI:** A binary variable (0 or 1) indicating whether the stores of CA, TX or WI allow SNAP purchases on the examined date. 1 indicates that SNAP purchases are allowed.

Evaluation:

$$\boldsymbol{RMSSE} = \sqrt{\frac{1}{h}\frac{\sum_{t=n+1}^{n+h}\left(Y_t - \widehat{Y}_t\right)^2}{\frac{1}{n-1}\sum_{t=2}^{n}\left(Y_t - Y_{t-1}\right)^2}},$$

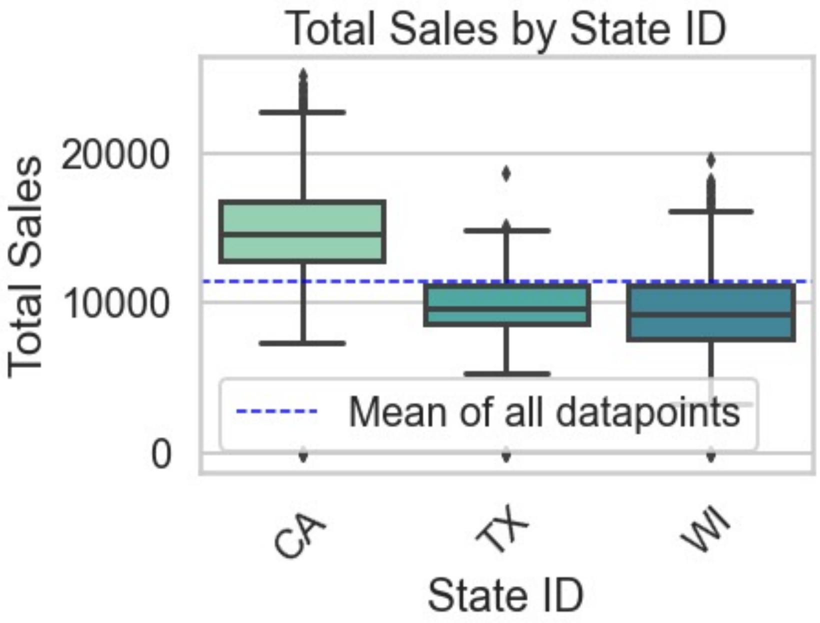
Y_t: actual future value of the examined time series at point t,
$\widehat{Y}_t$: generated forecast
n: length of the training sample
h: forecasting horizon (28)

- **Training set:** d_1 to d_1913
- **Validation /Evaluation set:** d_1914 to d_1941
- **Test set:** d_1942 to d_1969

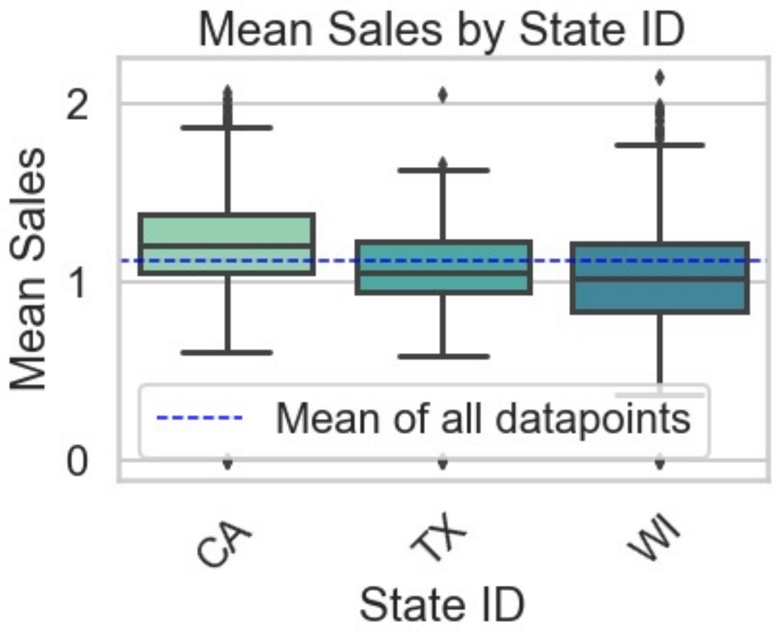
2. EDA

2.1 Sales of Different States

CA: 4 stores TX: 3 stores WI: 3 stores



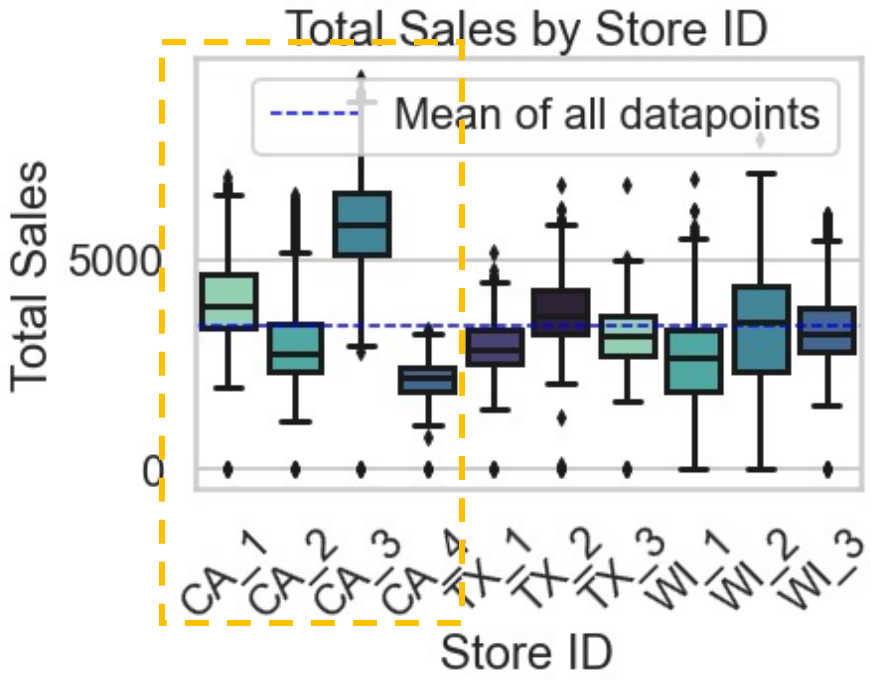
The mean sales have no significant difference.

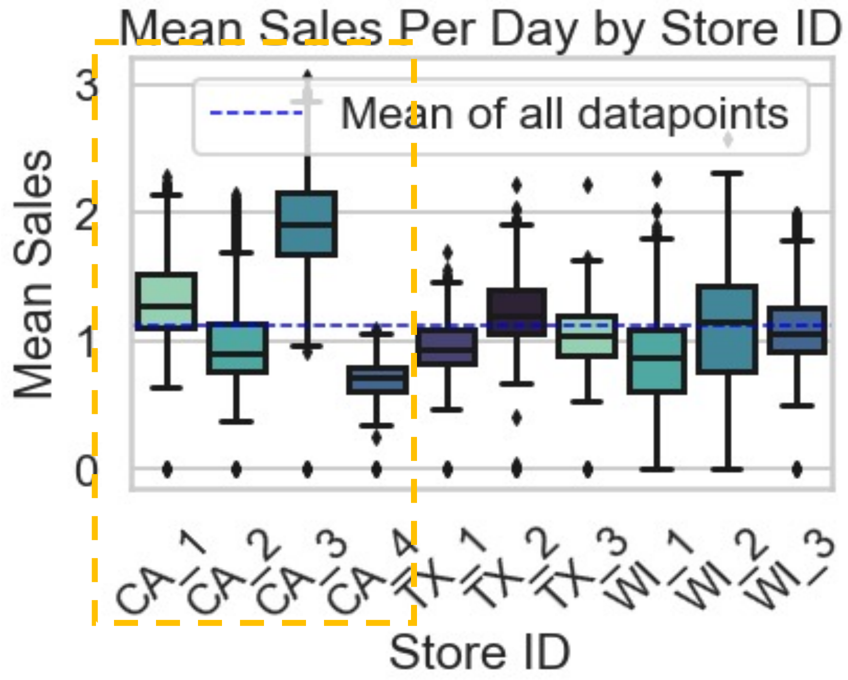


The total sales of CA are significantly higher than that of TX and WI.

2.2 Sales of Different Stores

Q: Why CA has higher mean sales?





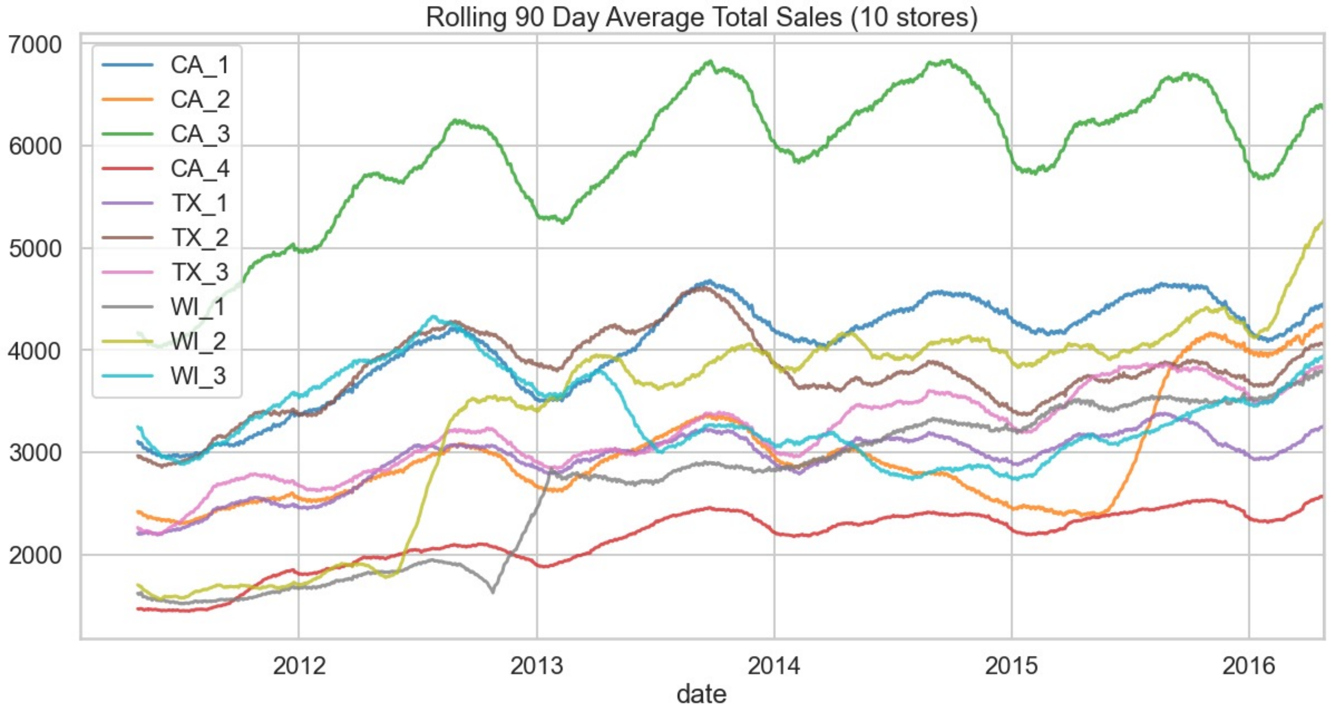
CA_3 : makes the most sales → leads to the high overall mean sales of CA

CA_2 : similar to other stores

CA_4 : relatively low

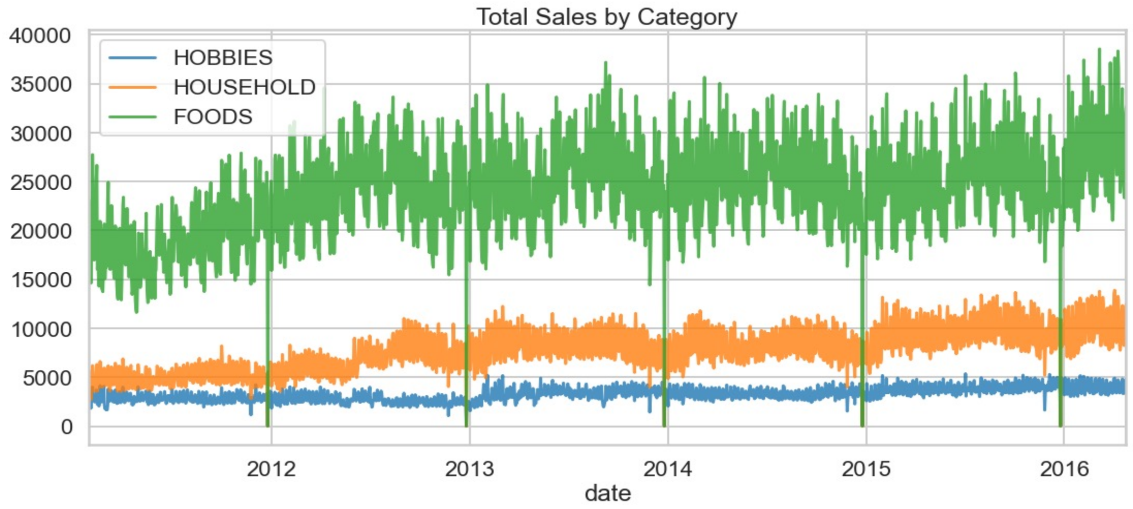
2.3 90-day Average Total Sales of Stores

Rolling 90-day average total sales for each store

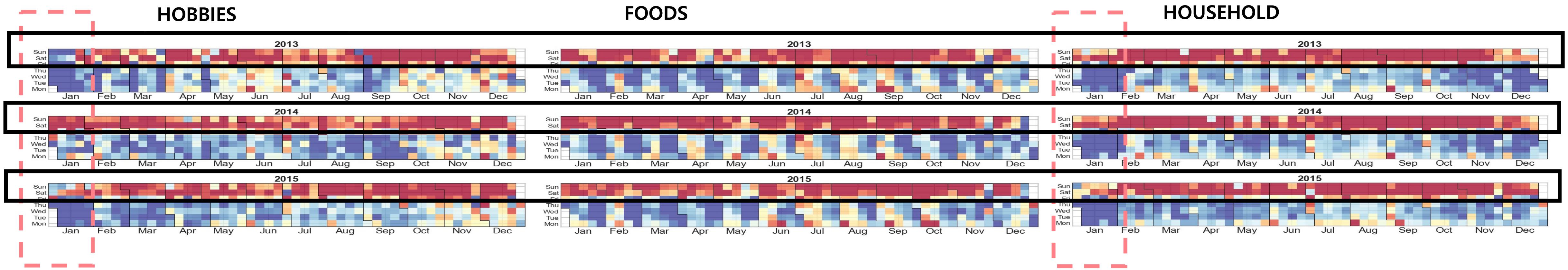


- Some stores have wide fluctuations in average total sales.

2.4 Sales of Different Categories

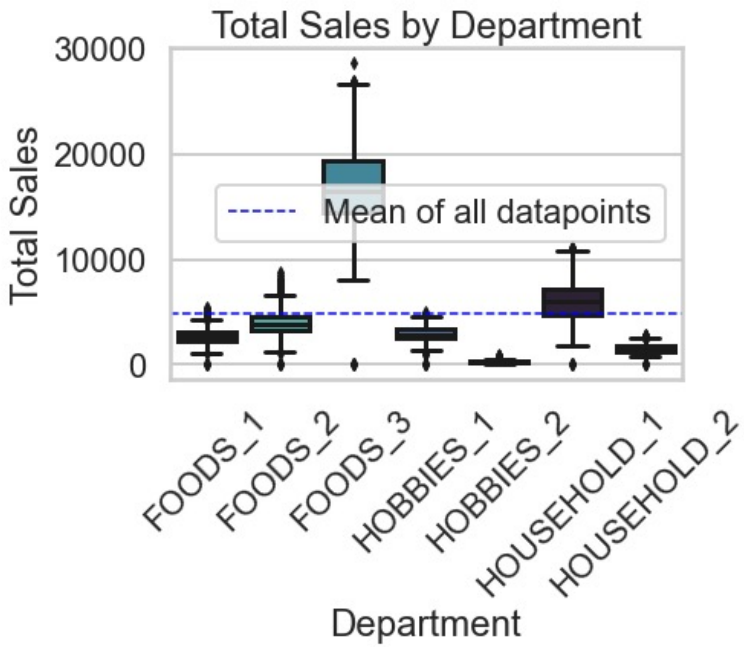


- FOODS has the most sales, followed by HOUSEHOLD and HOBBIES.
- **FOODS** sales are higher in the middle of the year and generally decline in the second half of the year.
- Sales of **HOUSEHOLD** and **HOBBIES** hit a low in January
- For all categories, **weekends** contribute more sales than weekdays.

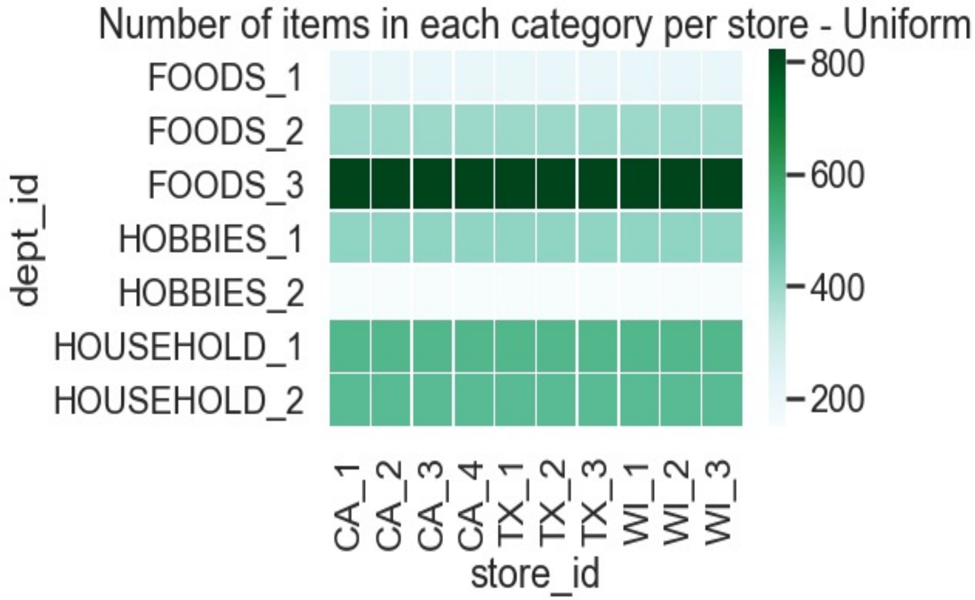


Red color represents more sales, while blue represents less sales

2.5 Sales of Different Departments



Aggregate

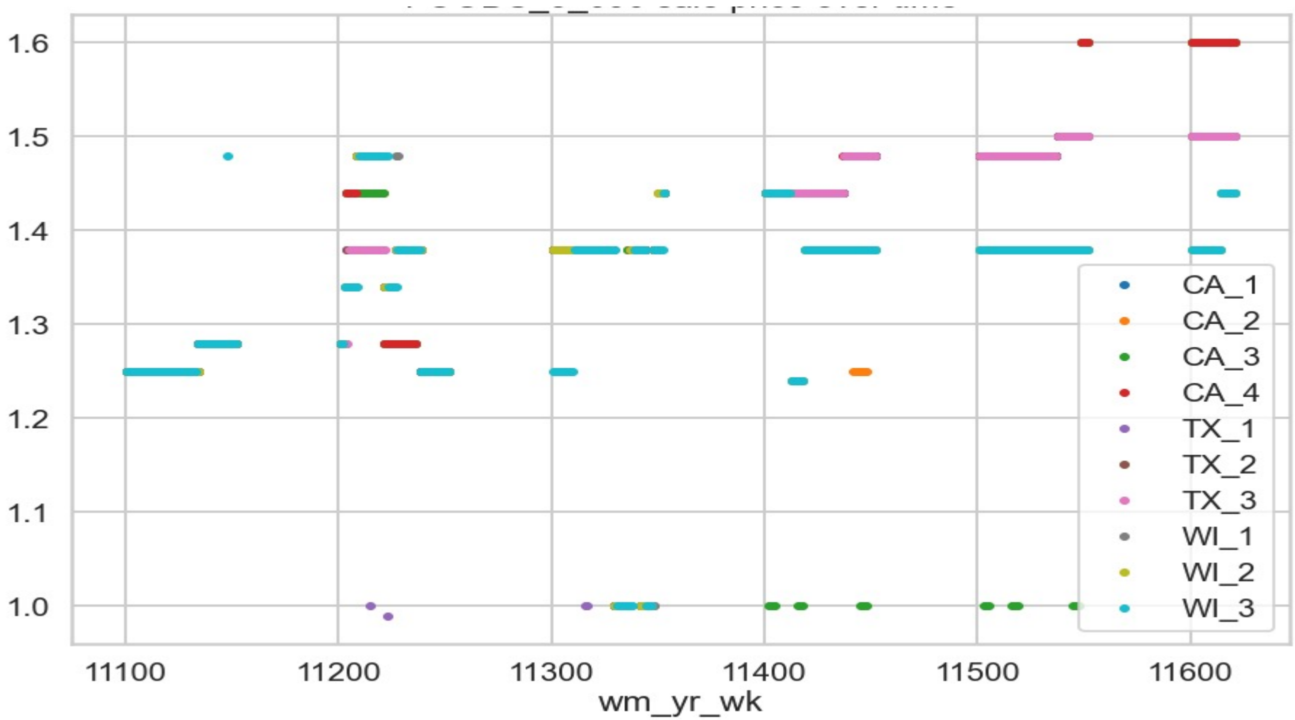


FOODS_3 and HOUSEHOLD_1 make the most total sales

All stores have the same kind of items and are not selling more of one category or another

2.6 Price of Sample Item

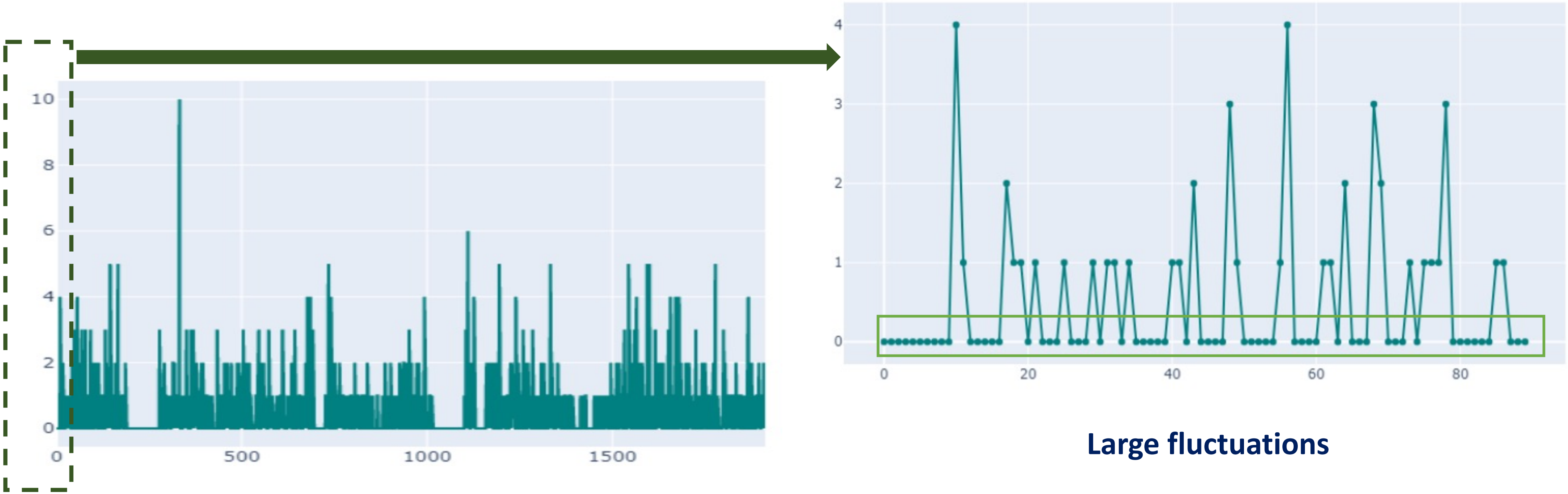
FOODS_3_090



- The price of FOODS_3_090 has increased over time.
- In the same period, different stores have different sell prices.

2.7 Instability

FOODS_1_001_TX_3

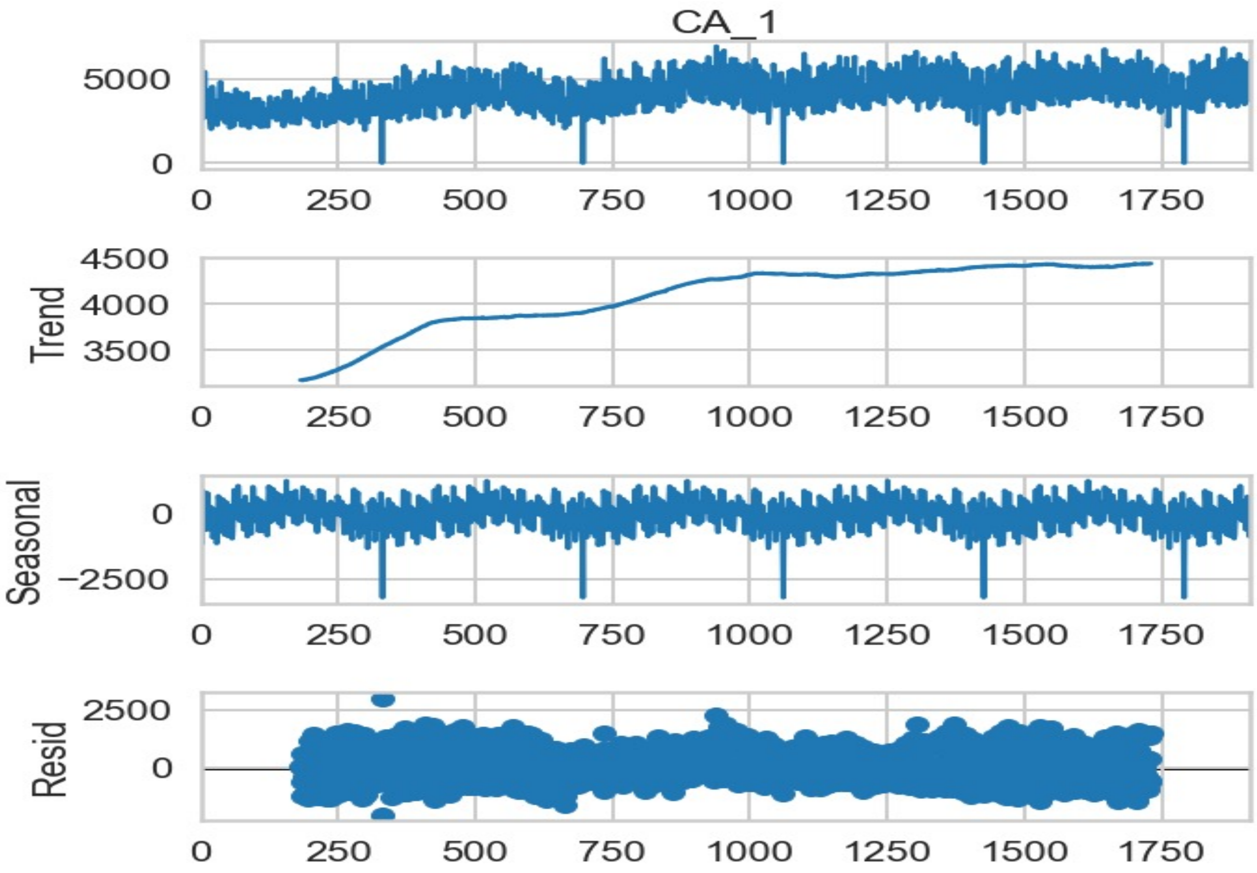


Large fluctuations

- No sales: The product may not be available on that day or the stores are closed.

2.8 Seasonality and Trend

CA_1

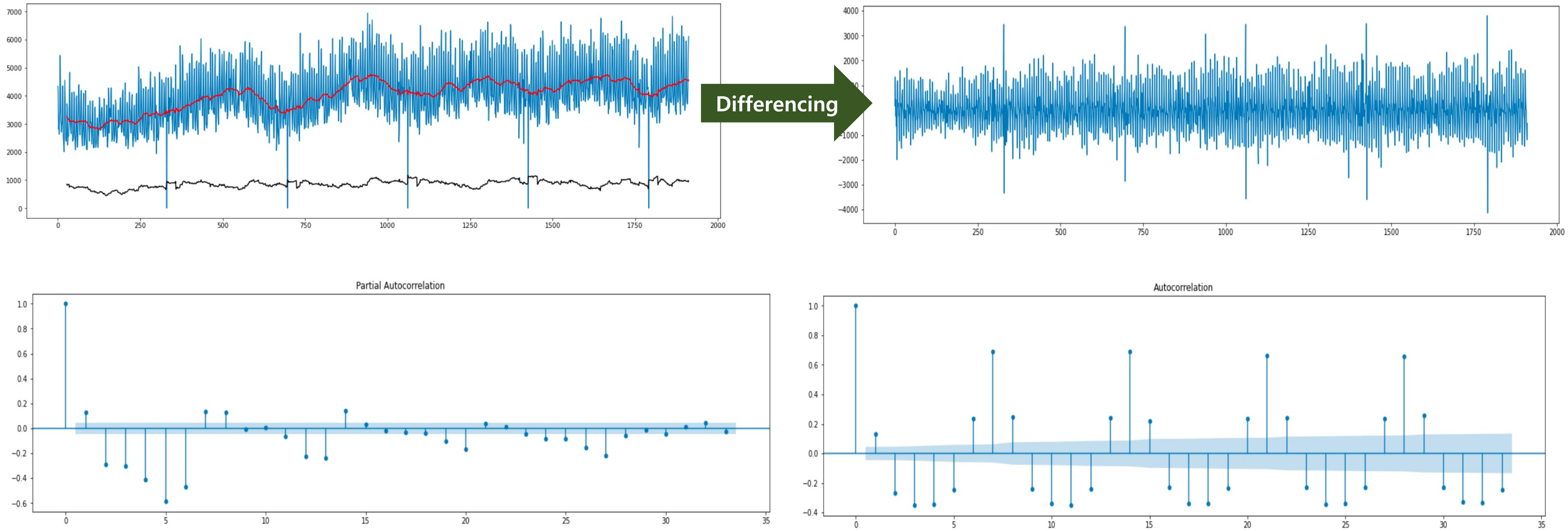


3. Modelling

3.1 SARIMA

Step1: Construct models for total sales of each store

Total sales of CA_1



Model: SARIMA(5, 1, 0)x(1, 0, [1], 7)

3.1 SARIMA

Step2: Apply the fitted the model to all the items in the store

RMSE of validation set at the store level

Store	RMSE
CA_1	5.9030
CA_2	2.8008
CA_3	3.2848
CA_4	1.6776
TX_1	2.3938
TX_2	3.2404
TX_3	3.2241
WI_1	2.1808
WI_2	4.3800
WI_3	3.0816

- Using more explanatory variables?
- Constructing more features capturing time shift effect?

3.2 LightGBM

Step1: Feature construction

Features	Description	Meaning
sold-lag-7	sales shifted 7 steps downwards for each item	Captures the week-on-week similarity.
sold-lag-28	sales shifted 28 steps downwards for each item	Captures the weekly similarity from a month-to-month perspective.
rmean-7-7	rolling mean sales of a window size of 7 over lag-7	Captures the information regarding the sales of the whole previous week ending 7 days ago.
rmean-7-28	rolling mean sales of a window size of 7 over lag-28	Captures the information regarding the sales of the entire previous 4 weeks ending 7 days ago.
rmean-28-7	rolling mean sales of a window size of 28 over lag-7	Captures the information regarding the sales of the whole week ending 4 weeks ago.
rmean-28-28	rolling mean sales of a window size of 28 over lag-28	Captures the information regarding the sales of the entire previous 4 weeks ending 4 weeks ago.
item-sold-avg	mean sales for each item	
store-sold-avg	mean sales for each store	

The other features used to construct models:

- d month year wm_yr_wk
- id item_id dept_id cat_id store_id state_id
- weekday wday
- sell_price
- snap_CA snap_TX snap_WI
- event_name_1 event_type_1 event_name_2 event_type_2

3.2 LightGBM

Step2: Parameters tuning and modelling

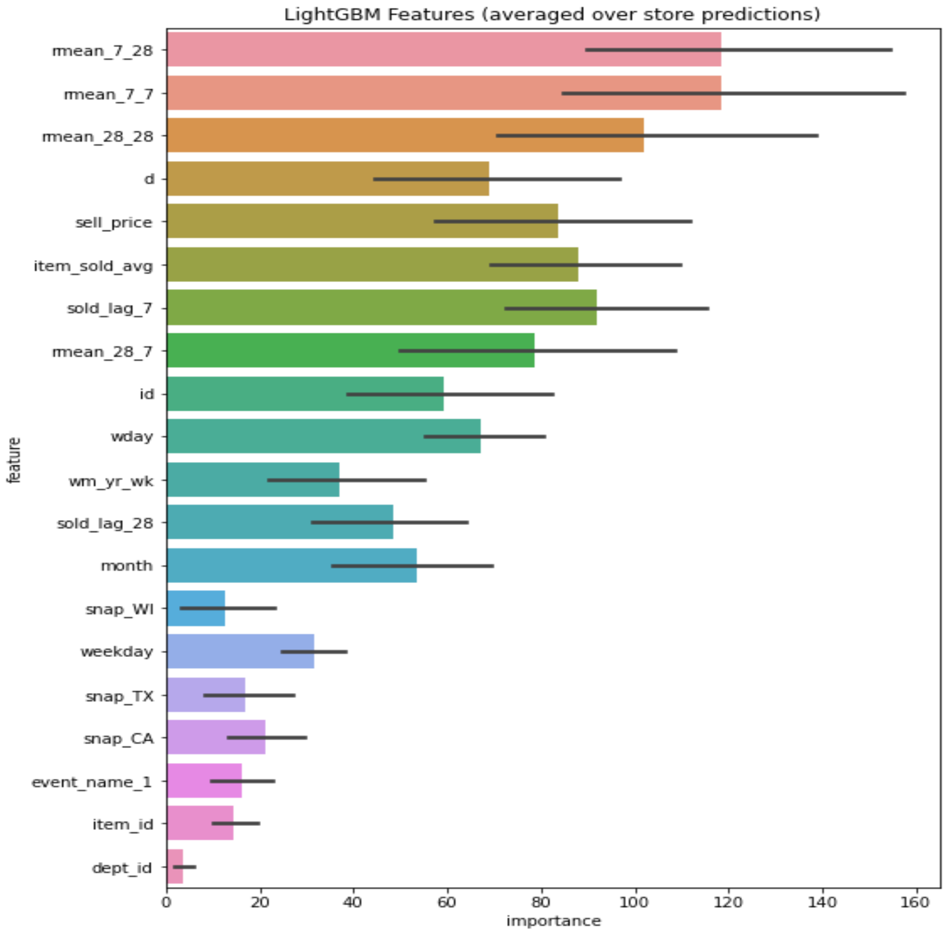
Parameters

- n_estimators:1000
- learning_rate: 0.3
- subsample:0.8
- colsample_bytree=0.8,
- max_depth=8,
- num_leaves=50,
- min_child_weight=300

RMSE of validation set at the store level

Store	RMSE
CA_1	2.10675
CA_2	1.95357
CA_3	2.518
CA_4	1.41278
TX_1	1.70231
TX_2	1.84975
TX_3	1.95191
WI_1	1.66553
WI_2	2.88884
WI_3	1.98551

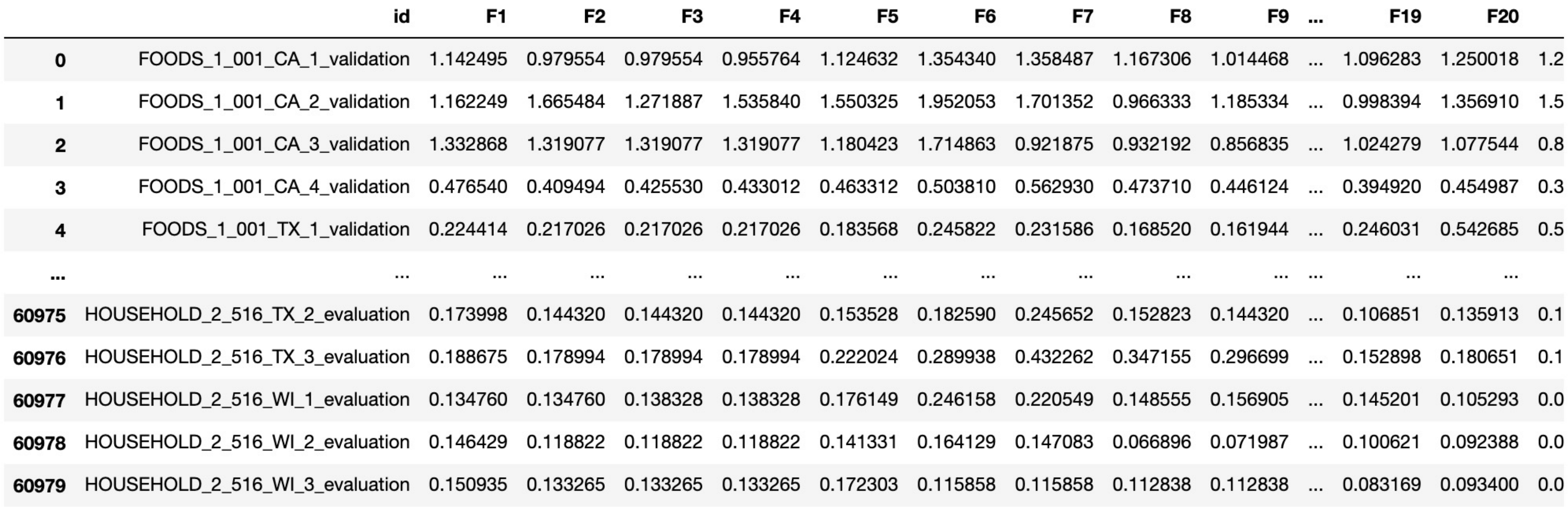
Feature importance



4. Prediction

4.Prediction

1. Predict sales of each item on d_1914 to d_1941
2. Predict sales of each item on d_1942 to d_1969



	id	F1	F2	F3	F4	F5	F6	F7	F8	F9	...	F19	F20	
0	FOODS_1_001_CA_1_validation	1.142495	0.979554	0.979554	0.955764	1.124632	1.354340	1.358487	1.167306	1.014468	...	1.096283	1.250018	1.2
1	FOODS_1_001_CA_2_validation	1.162249	1.665484	1.271887	1.535840	1.550325	1.952053	1.701352	0.966333	1.185334	...	0.998394	1.356910	1.5
2	FOODS_1_001_CA_3_validation	1.332868	1.319077	1.319077	1.319077	1.180423	1.714863	0.921875	0.932192	0.856835	...	1.024279	1.077544	0.8
3	FOODS_1_001_CA_4_validation	0.476540	0.409494	0.425530	0.433012	0.463312	0.503810	0.562930	0.473710	0.446124	...	0.394920	0.454987	0.3
4	FOODS_1_001_TX_1_validation	0.224414	0.217026	0.217026	0.217026	0.183568	0.245822	0.231586	0.168520	0.161944	...	0.246031	0.542685	0.5
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
60975	HOUSEHOLD_2_516_TX_2_evaluation	0.173998	0.144320	0.144320	0.144320	0.153528	0.182590	0.245652	0.152823	0.144320	...	0.106851	0.135913	0.1
60976	HOUSEHOLD_2_516_TX_3_evaluation	0.188675	0.178994	0.178994	0.178994	0.222024	0.289938	0.432262	0.347155	0.296699	...	0.152898	0.180651	0.1
60977	HOUSEHOLD_2_516_WI_1_evaluation	0.134760	0.134760	0.138328	0.138328	0.176149	0.246158	0.220549	0.148555	0.156905	...	0.145201	0.105293	0.0
60978	HOUSEHOLD_2_516_WI_2_evaluation	0.146429	0.118822	0.118822	0.118822	0.141331	0.164129	0.147083	0.066896	0.071987	...	0.100621	0.092388	0.0
60979	HOUSEHOLD_2_516_WI_3_evaluation	0.150935	0.133265	0.133265	0.133265	0.172303	0.115858	0.115858	0.112838	0.112838	...	0.083169	0.093400	0.0

Thanks for watching