

# *Predict Future Sales*

*~A Competition on Kaggle*

資科四 106703055 黃浩瑋

資科四 106703043 林琰歲

資科四 106703018 戴冕

# 資料介紹

## **items.csv**

*item\_name item\_id item\_category\_id*

## **item\_categories.csv**

*item\_category\_name item\_category\_id*

## **shops.csv**

*shop\_name shop\_id*

## **sales\_train.csv**

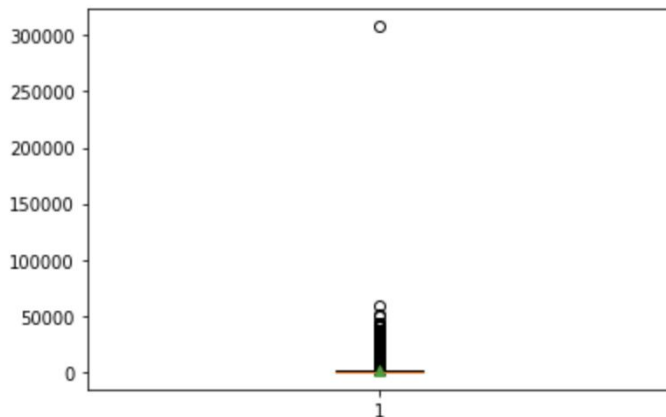
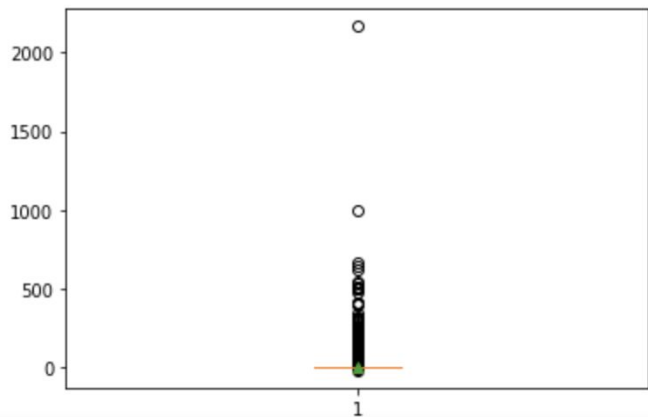
*date date\_block\_num shop\_id item\_id item\_price item\_cnt\_day*  
2,935,849筆

## **test.csv**

*ID shop\_id item\_id*  
214,200筆

# 資料前處理-異常值處理

item\_cnt\_day&item\_price



# 資料前處理-shop值處理

## 相同shop處理

	shop_name	shop_id
39	РостовНаДону ТРК "Мегацентр Горизонт"	39
40	РостовНаДону ТРК "Мегацентр Горизонт" Островной	40

## shop劃分為city和category

	shop_name	shop_id	city	category
0	!Якутск Орджоникидзе, 56 фран	0	Якутск	Орджоникидзе,
1	!Якутск ТЦ "Центральный" фран	1	Якутск	ТЦ
2	Адыгея ТЦ "Мега"	2	Адыгея	ТЦ
3	Балашиха ТРК "Октябрь-Киномир"	3	Балашиха	ТРК
4	Волжский ТЦ "Волга Молл"	4	Волжский	ТЦ

出現頻率低的category用etc表示

## 資料前處理-item\_categories進行分類/篩選

	item_category_name	item_category_id	item_type	split	subtype
0	PC - Гарнитуры/Наушники	0	etc	[PC , Гарнитуры/Наушники]	Гарнитуры/Наушники
1	Аксессуары - PS2	1	Аксессуары	[Аксессуары , PS2]	PS2
2	Аксессуары - PS3	2	Аксессуары	[Аксессуары , PS3]	PS3
3	Аксессуары - PS4	3	Аксессуары	[Аксессуары , PS4]	PS4
4	Аксессуары - PSP	4	Аксессуары	[Аксессуары , PSP]	PSP

## 資料前處理-item\_name進行篩選/細分

	item_name	item_id	item_category_id	name	type
164	1с аудиокниги аркадий аверченко классика русск...	164	44	цифровая версия	цифровая
165	1с аудиокниги артур конан дойл долина страха	165	45	0	0
166	1с аудиокниги артур конан дойл истории о шерло...	166	45	0	0
167	1с аудиокниги аткинсон у сила мысли или магнет...	167	44	pc цифровая версия	pc

	item_id	item_category_id	name
0	0	40	4
1	1	76	48
2	2	40	4
3	3	40	4
4	4	40	4

## 進階資料處理-從「train」資料集製作 Dataframe「matrix」

	date_block_num	shop_id	item_id	item_cnt_month	shop_category	shop_city	item_category_id	name	subtype_code	item_type
0	0	2	19	0.0	4	0	40	4	4	5
1	0	2	27	1.0	4	0	19	60	10	3
2	0	2	28	0.0	4	0	30	77	55	3
3	0	2	29	0.0	4	0	23	90	16	3
4	0	2	32	0.0	4	0	40	4	4	5
...	...	...	...	...	...	...	...	...	...	...
11056272	34	45	18454	0.0	4	20	55	4	2	7
11056273	34	45	16188	0.0	4	20	64	4	42	8
11056274	34	45	15757	0.0	4	20	55	4	2	7
11056275	34	45	19648	0.0	4	20	40	4	4	5
11056276	34	45	969	0.0	4	20	37	4	1	5

11056277 rows × 10 columns

# 進階資料處理-Matrix

- . 每種資料的排列組合

- . 新增變數—Item\_cnt\_month, Revenue

Item\_cnt\_month:每個月不同店家, 所賣出的item數量-種類。

Revenue:每個店家在該item類別的月收入

- . 新增Lag Feature



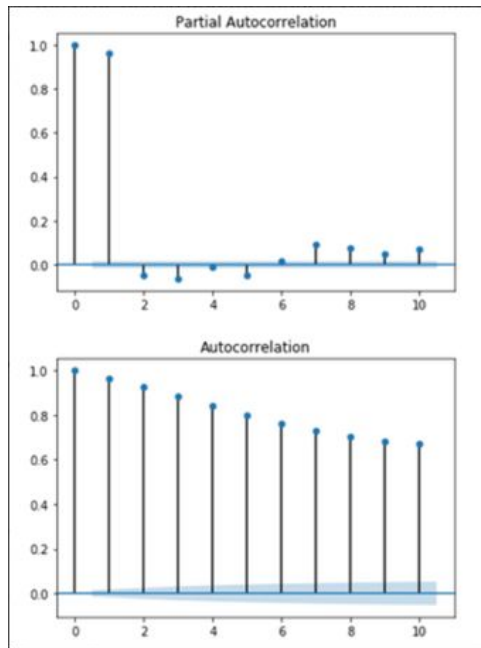
# Lag Feature

- . 時間性的問題處理方法, 可以用在股價預測、銷售量等
- . 前面的月份會影響現在的銷售量

# Lag Feature Estimate- ACF vs PACF

ACF(Autocorrelation Function): ACF會計算時間跟自己本身變因的 correlation 關係。

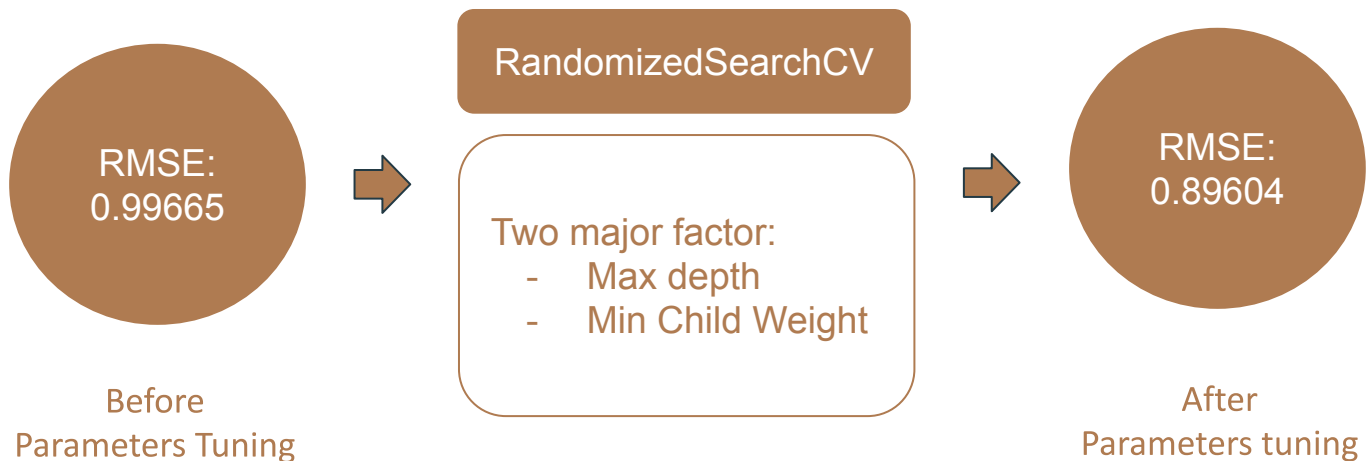
PACF(Partial Autocorrelation Function): PACF也會計算時間與自己本身變因的關係，不過會把已經解釋的變因移除。



# 新增完Lag Feature後的matrix

item_cnt_month_lag_1	0.000000	0.000000	NaN	0.000000	1.000000
item_cnt_month_lag_2	0.000000	0.000000	NaN	1.000000	1.000000
item_cnt_month_lag_3	0.000000	0.000000	NaN	0.000000	4.000000
date_avg_item_cnt_lag_1	0.286865	0.286865	NaN	0.286865	0.286865
date_item_avg_item_cnt_lag_1	0.021744	0.130493	NaN	2.826172	1.260742
date_item_avg_item_cnt_lag_2	0.086975	0.152222	NaN	11.046875	4.781250
date_item_avg_item_cnt_lag_3	0.065247	0.173950	NaN	18.734375	13.648438
date_shop_avg_item_cnt_lag_1	0.071838	0.071838	NaN	0.071838	0.071838
date_shop_avg_item_cnt_lag_2	0.091064	0.091064	NaN	0.091064	0.091064
date_shop_avg_item_cnt_lag_3	0.059875	0.059875	NaN	0.059875	0.059875
date_shop_item_avg_item_cnt_lag_1	0.000000	0.000000	NaN	0.000000	1.000000
date_shop_item_avg_item_cnt_lag_2	0.000000	0.000000	NaN	1.000000	1.000000
date_shop_item_avg_item_cnt_lag_3	0.000000	0.000000	NaN	0.000000	4.000000
date_shop_subtype_avg_item_cnt_lag_1	0.449463	0.387207	NaN	0.018585	0.035919
date_city_avg_item_cnt_lag_1	0.071838	0.071838	NaN	0.071838	0.071838
date_item_city_avg_item_cnt_lag_1	0.000000	0.000000	NaN	0.000000	1.000000
delta_price_lag	0.367676	0.256348	0.0	0.212402	0.191040
delta_revenue_lag_1	37326.816406	37326.816406	NaN	37326.816406	37326.816406

# XGBoost





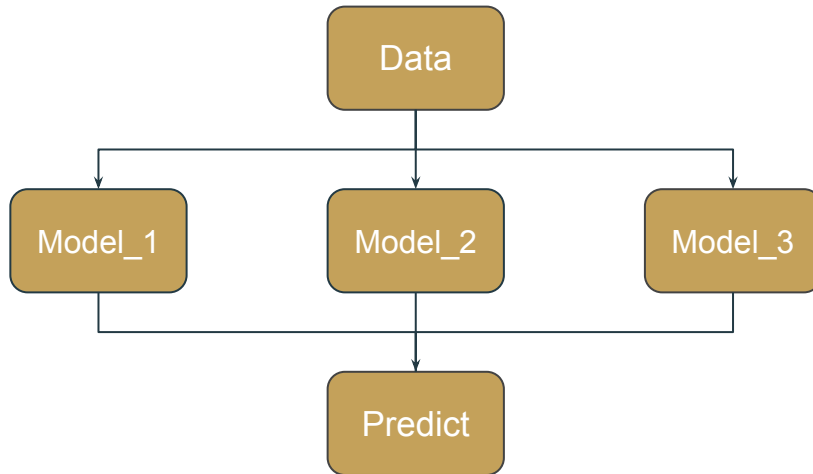
*What about other models?*

## Other Models

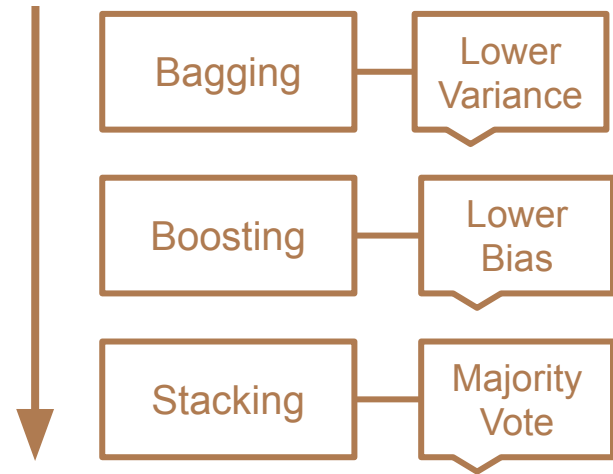
<i>Random Forest</i>	➔	RMSE: 1.01704
<i>SVD Regressor</i>	➔	RMSE: 1.06846
<i>Ridge Regressor</i>	➔	RMSE: 1.07151

NOT better than XGBoost!!

# Essemble Model



## 3 Methods




# Performance

- Q: How to merge the prediction of every model?  
A: How about use the Mean of all predictions?



RMSE:  
0.98507

- 
- Q: Is it good enough?  
A: We can use the predictions of the models be the input of an additional model. (e.g. Linear Regression)



RMSE:  
0.90700



# Output details & Null Model

	ID	item_cnt_month
0	0	0.509988
1	1	0.381447
2	2	0.862875
3	3	0.473632
4	4	4.759768
...	...	...
214195	214195	0.058575
214196	214196	0.003565
214197	214197	0.056603
214198	214198	0.006932
214199	214199	0.036925

214200 rows × 2 columns



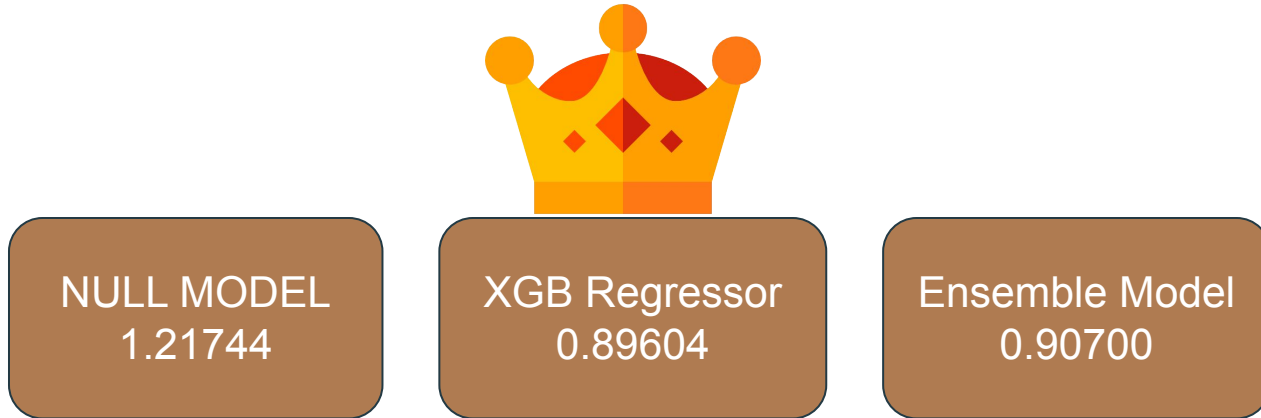
Mean:  
0.286274



	ID	item_cnt_month
0	0	0.286275
1	1	0.286275
2	2	0.286275
3	3	0.286275
4	4	0.286275
...	...	...
214195	214195	0.286275
214196	214196	0.286275
214197	214197	0.286275
214198	214198	0.286275
214199	214199	0.286275

214200 rows × 2 columns






# Model Comparison



# Ranking



The First Place with 0.75980

1014	Jiachen Shi		0.89601	14	9mo
1015	omarB202791		0.89601	2	14d
1016	tourist		0.89601	8	2y
1017	houheixue		0.89603	6	2y
1018	1091DS_106703055		0.89604	14	2h



Q&A