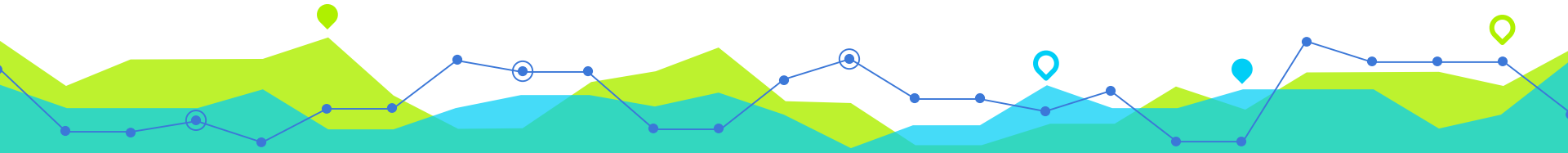


M5 forecasting

陳志瑄

2021/1/28

目錄

- 資料理解與目的
- 探索性分析的過程
- 建立模型與驗證
- 結論
- 附錄

資料理解與猜測

- Calendar：為日期資料(2011/1-2016/6)包含節假日、政府補助日SNAP等等資料 (或許節日會影響買氣、購買衝動)
- Sell_price：為各產品每日的銷售價錢(或許價錢會影響購買意願)
- sales_train_validation：前 1913 天的銷售量
- sales_train_evaluation：前 1941 天的銷售量(比 validation 多28天)

目的：預測3049種商品*(10家店)的28天銷售量(1942~1969天)

資料範例參照附錄-資料合併

資料分析流程

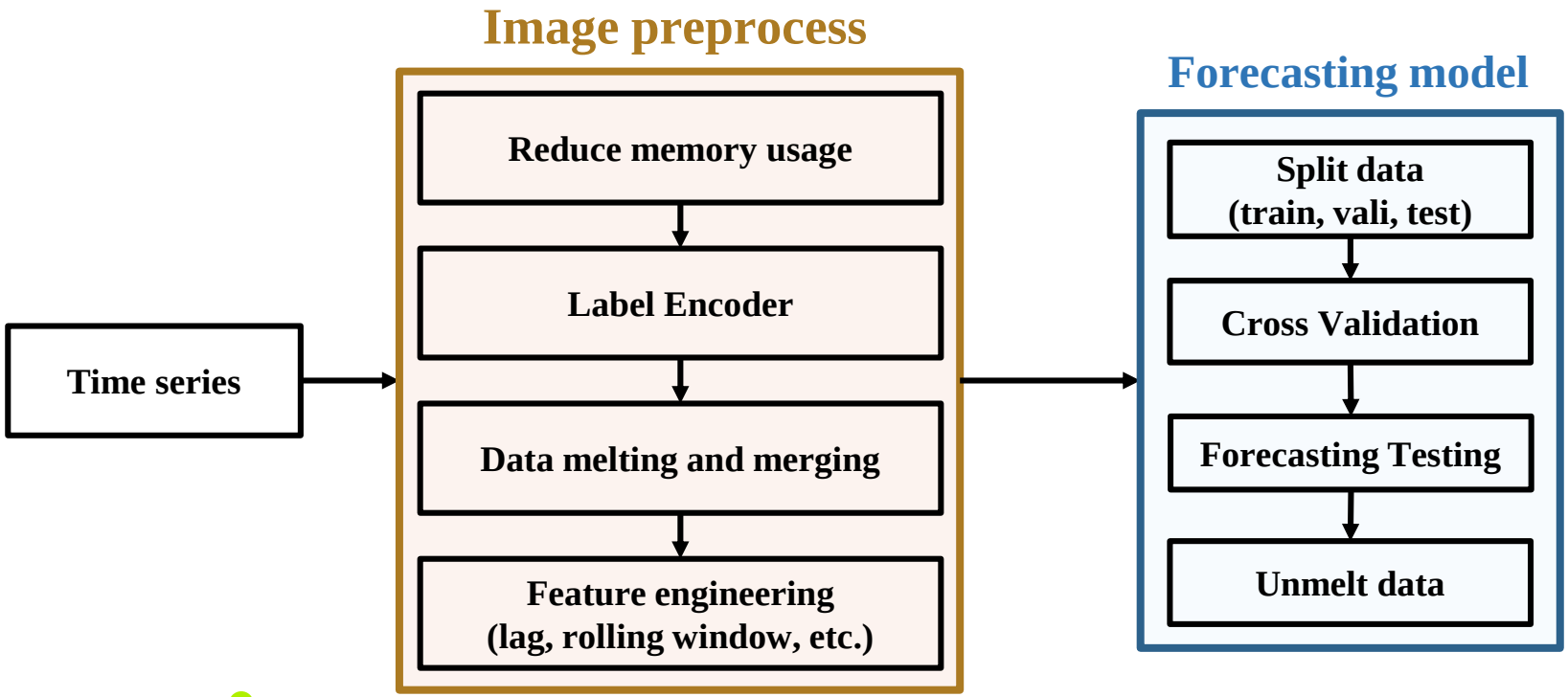
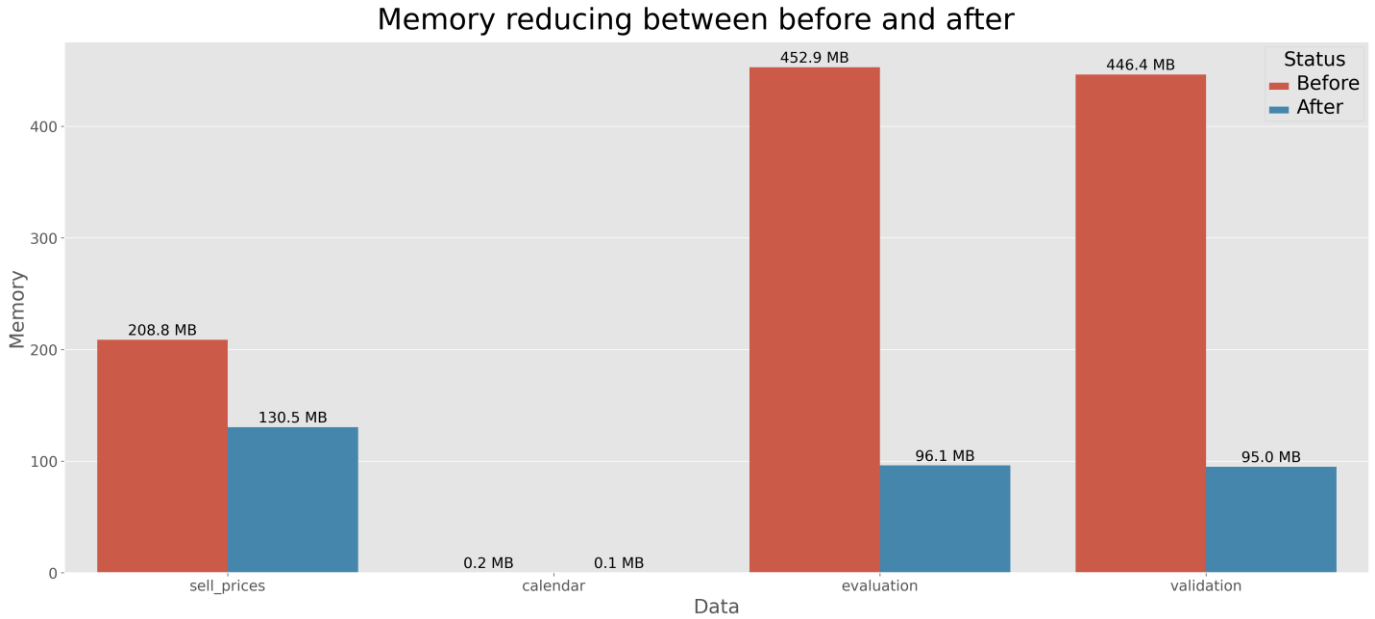


Image preprocess
Forecasting model
Time series
Reduce memory usage
Label Encoder
Data melting and merging
Feature engineering
(lag, rolling window, etc.)
Split data
(train, vali, test)
Cross Validation
Forecasting Testing
Unmelt data

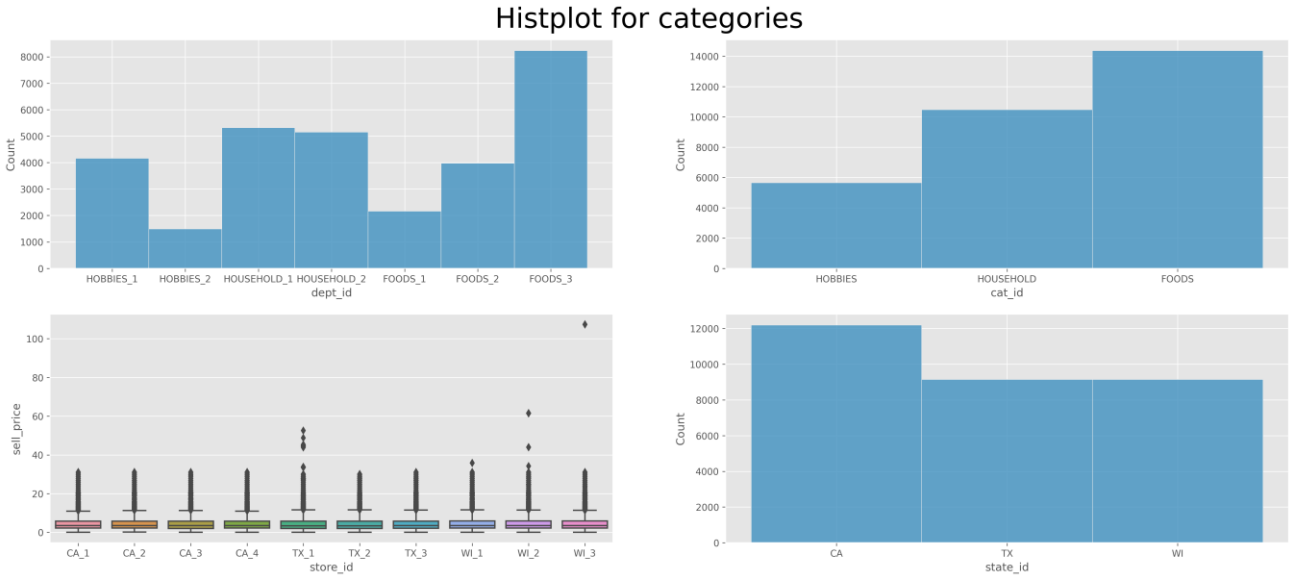
Memory reduction



鑒於本次資料集稍微龐大，因此須採用將資料型態改至最小範圍(int64 → int16)等等，來減少記憶體使用。最大可減少70%的記憶體

Ref： https://www.kaggle.com/fabiendaniel/elo-world

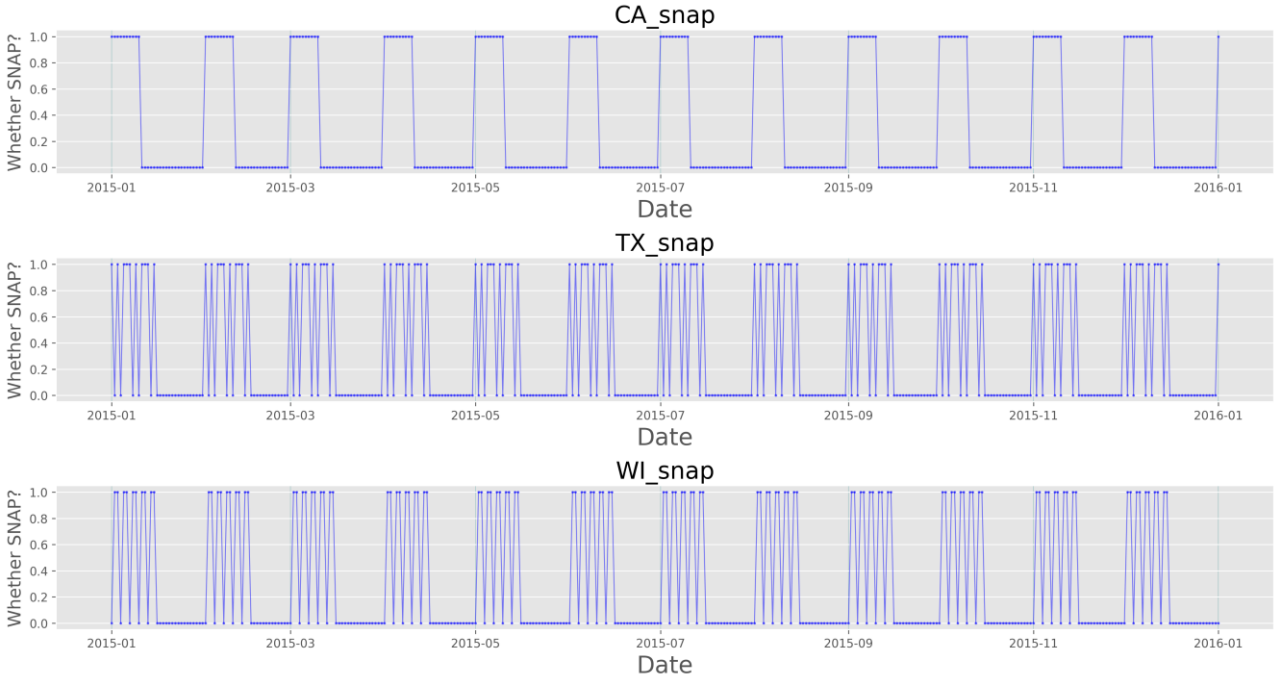
各資料直方圖



- 3 大產品分類
  (Hobbies, Foods, 和Household)
  且 FOODS的占比最多
- 7 小產品分類
- 3 大洲分布 (CA比TX跟WI多)
- 10 分店
- 3,049 種商品
- 共有 30,490 個商品

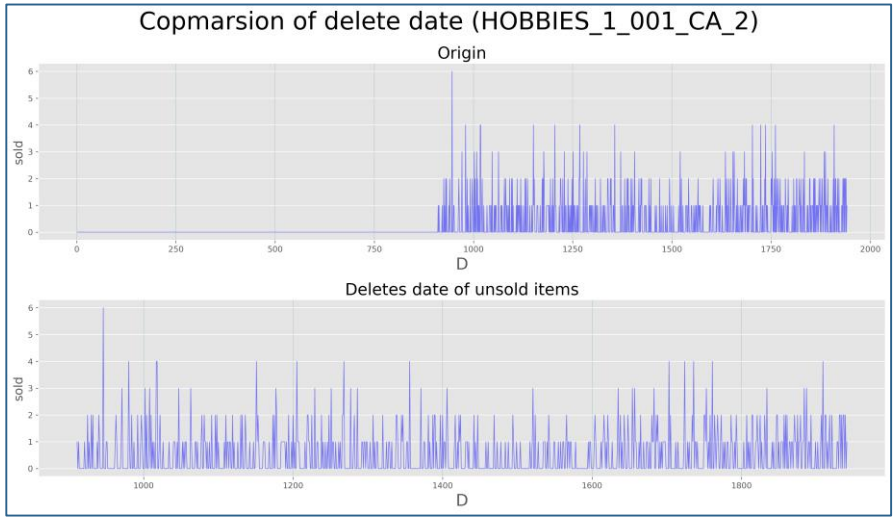
Supplemental Nutrition Assistance Program (SNAP)

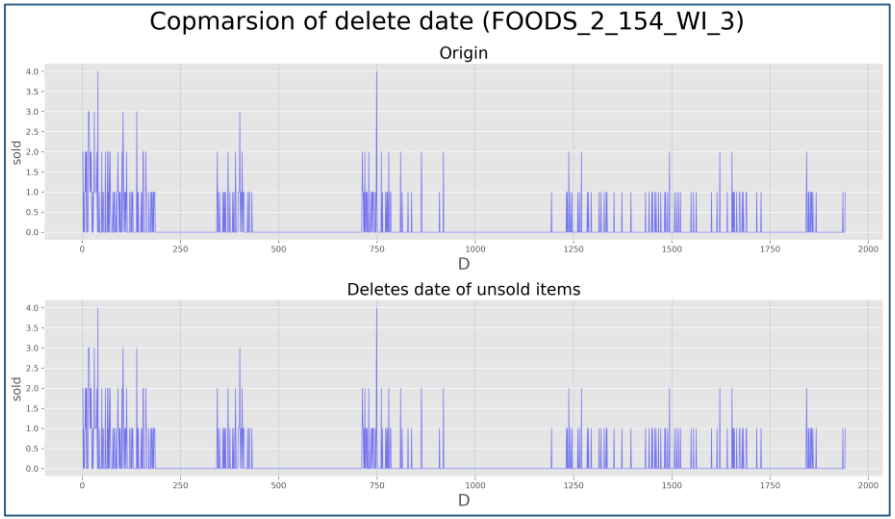
美國補充營養援助計畫(Supplemental Nutrition Assistance Program；SNAP)亦被稱為食物券計畫(Food Stamp Program)，主要為無收入及低收入美國居民提供購買食品的補助。



將calendar資料的snap經由表格觀察以及畫圖發現，snap在各州的每個月是有週期性的，而且每年是固定的，因此需考慮此因素當作特徵進行預測。

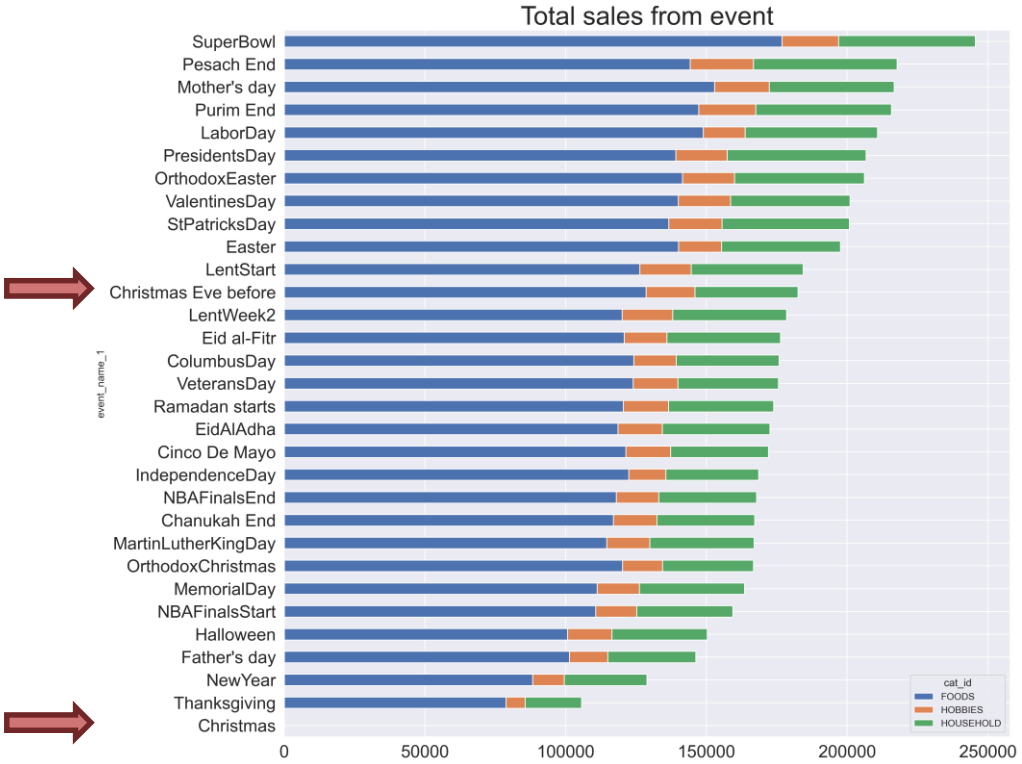
EDA -資料截取





資料對每個商品統一給了1941天的銷售量，因此會出現許多商品會有一段時間的銷售量皆為0，其原因是為尚未開賣。我採取了對每個商品進行抓取開賣時間，並將未開賣的資料捨去。舉例 HOBBIES_1_001_CA_2 (左)前900天捨去，但像 FOODS_2_154_WI_3 (右) 就第一天就開賣，就不需要進行截取。

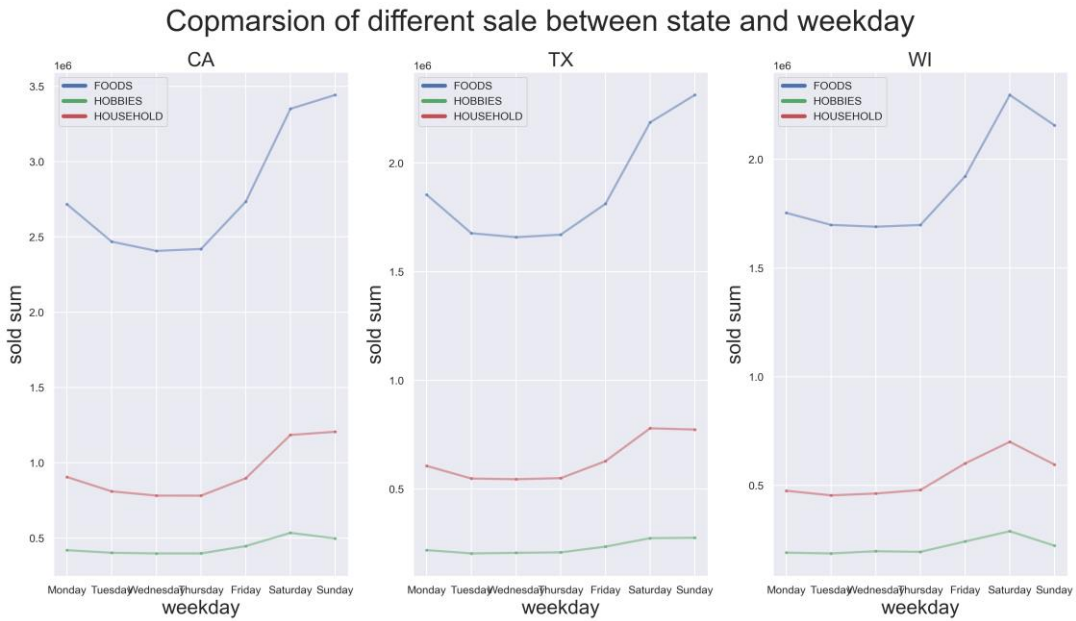
EDA-Event sales



左圖為根據不同的節日以不同種的商品來計算銷售量。很明顯得知在超級盃(Super Bowl)的食物銷售量為最多，可能原因為大家在看賽事的時候喜歡配著東西吃。

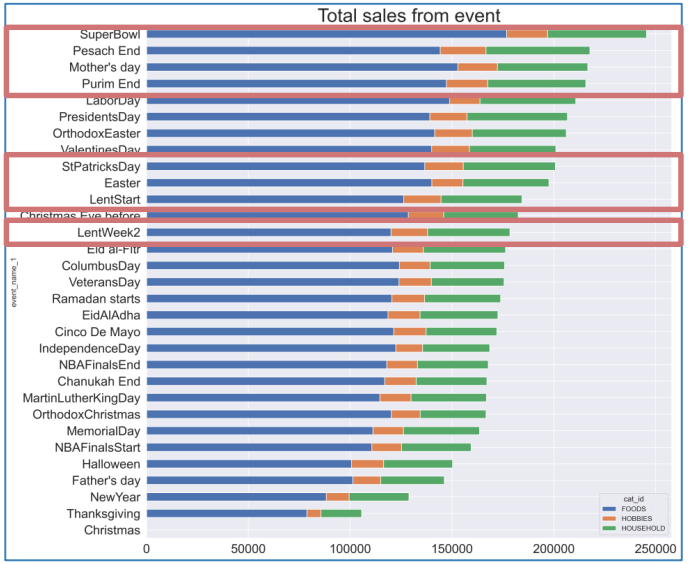
然而推斷 Christmas、Thanksgiving、New Year 可能因為歐美最重大的特殊節日會有許多家庭大量購買食物來慶祝，實際狀況是為倒數三名。但 Christmas 當天 Walmart 是停業的，因此我加入了聖誕夜前一天(Christmas Eve before) 的銷售狀況，也並沒有預期的高。推測原因為消費習慣與想像中不同，多數人不在家裡烹調而是選擇在外用餐來慶祝。

EDA-Weekly sales



左圖為根據不同的天以不同種的商品來計算銷售量。並不意外的，銷售量在五、六、日的時候會大幅提升，其餘維持平常的銷售量。

EDA-Monthly sales



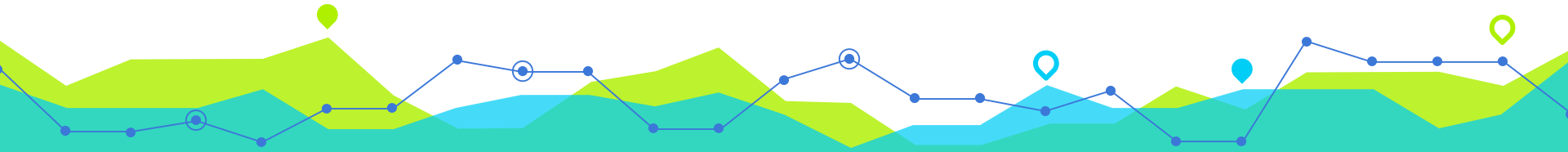
左圖為根據不同的月以不同種的商品來計算銷售量。三、四、五月為推測為這三個月的特殊節日相對後半年較為多，且其總銷售的皆為前幾名，因此才會造成銷售旺季。

EDA-Yearly sales

左圖為根據不同的年以不同種的商品來計算銷售量。可以觀察到每種商品在每州的總銷售量皆是逐年上升，為每年的總體趨勢。

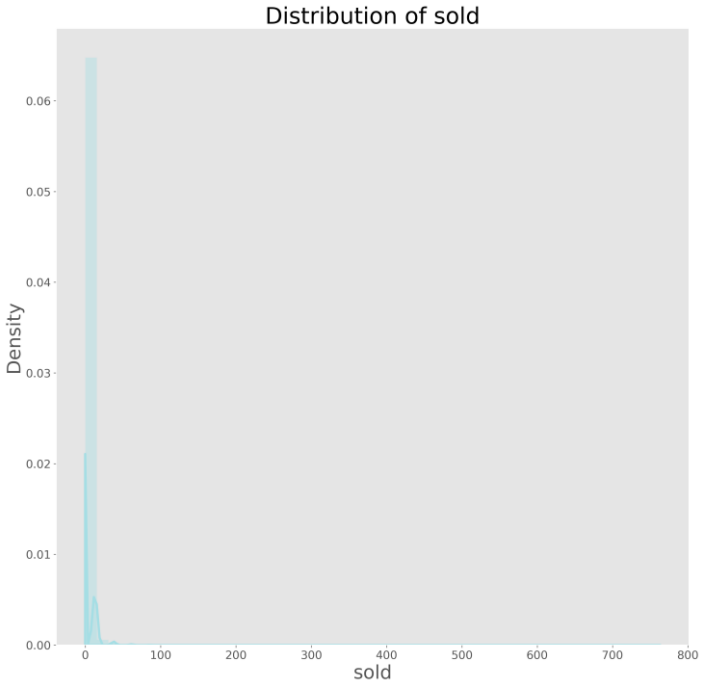
2011至2012的快速上升原因詳見EDA -資料截取。

(2016年資料未滿半年，故不考慮)



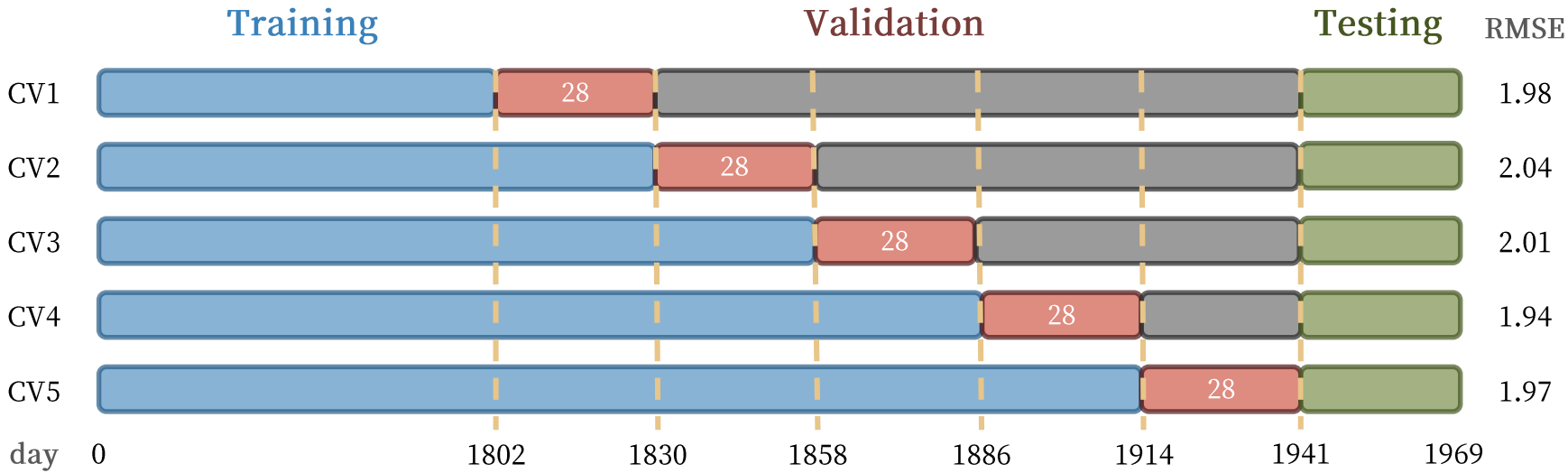
模型與驗證

Loss function 選擇



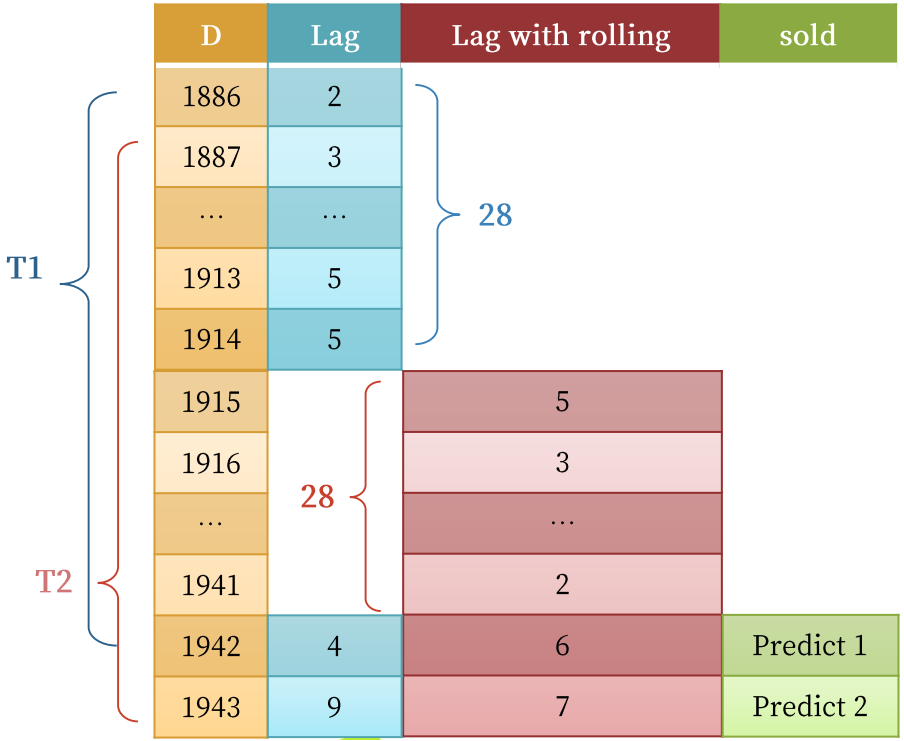
- 本次比賽的評估方法為 RMSE 的變形-Weighted Root Mean Squared Scaled Error (RMSSE) 的，但在 Discussion 有許多人討論到採用此方法當做 loss function 的結果並不如預期。我的推論原因是銷售量(預測值)是並不屬於常態分配，所以不該使用 RMSE, MAE, etc.應看預測值的分布來採取不同的 loss function。此資料接近 Poisson ，但 Tweedie 更為合適。
- Tweedie 分布為 Poisson 分布及 Gamma 分布的混合型。其概念是屬於 0 的用Poisson，非 0 的用Gamma。
- Tweedie 分布最明顯的一個特點是以一定的概率生成數值為 0 的樣本。左圖為本次資料(截取開始商品販售之後)所畫出的直方、機率密度圖，由圖得知銷售量為 0 的機率佔了大多數，這是 Tweedie 分布的特性，因此採用 Tweedie loss 作為 loss function。

5-Folds Cross Validation



與一般 Classification 問題的 Cross Validation 不同的地方在於，Time series 的 Forecasting 預測會有時序性問題，因此採用的資料分割也會有所不同 。此次 5-Folds Cross Validation 的平均RMSE為1.988

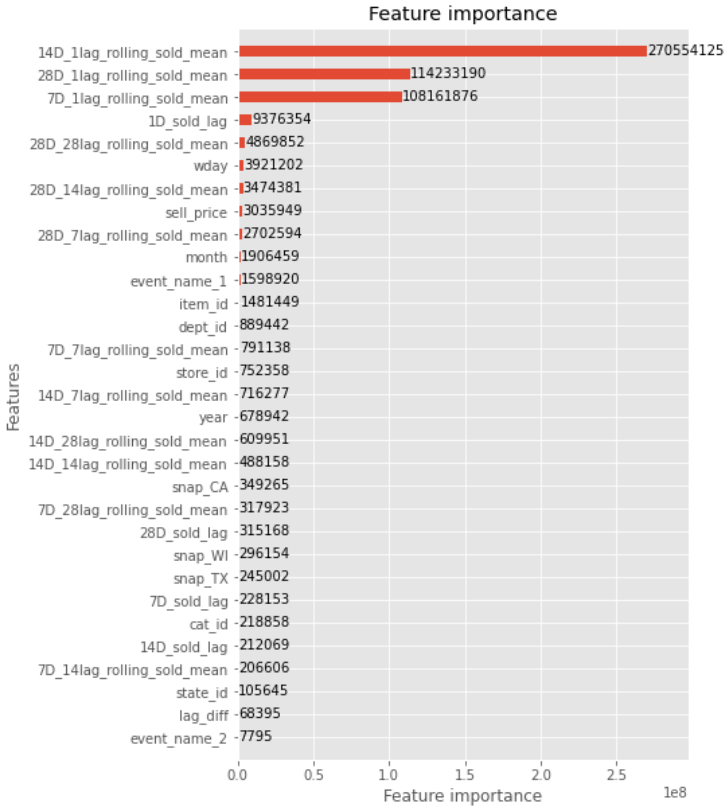
Testing 資料說明



Testing 資料示意圖

- **Testing 資料準備**：由於採用了 lag 及 lag with rolling mean 的特徵，在預測新一筆資料(Predict 1)的時候必須要往前提取 28 + 28 天數的資料(lag 28 天 + 對 lag 28 天使用 rolling mean 28 天)由此資料才會完整，與 training 資料一樣
- **遞迴預測**：在每次預測新一筆的時候，都會重新對其特徵做 lag 及 lag with rolling mean 且這些特徵範圍選取到上一個的 testing 資料，e.g.,預測 T1 往前選取 56 天資料；預測 T2 也往前選取 56 天資料，但會包含 T1 的資料重新做特徵工程。

Feature importance



左圖為進行 lightGBM 的 Feature importance 可以看出在 14D_11ag_rolling_sold_mean, 28D_11ag_rolling_sold_mean, 7D_11ag_rolling_sold_mean 中的表現佔了極大部分，此三個 feature 皆為做 1 lag 再將其往前7、14、28天做平均，前半個月、一個月或者一個星期的銷售量趨勢會影響到當天所要預測的銷售量。

結論與問題討論

1. 每個州每種商品在前半年是銷售旺季，與節日有關，尤其 FOODS 最為明顯。
2. 每個州每種商品在每年的總銷售量都有逐漸上升，為長時間的趨勢，並無大斷層或者急速上升的狀況。
3. 每一種損失函數背後都有一種假設，滿足假設的前提下，利用 loss function 訓練出來的模型才有比較好的效果。
4. 在特徵工程中，rolling mean的表現特別重要，可見在前數天銷售平均會影響預測當天的狀況
5. 資料應加入是否缺貨狀況，否則可能會誤判實際的銷售結果。
6. 未來可選用多模型進行 stacking 以達到更好的結果。

附錄

附錄-資料重整 (melt)

Item_id	d_1	d_2	d_3
S_1_001	1	0	2
S_1_002	0	3	0
S_1_003	0	0	0

30490 rows X 1947 columns

Item_id	d	sold
S_1_001	1	1
S_1_001	2	0
S_1_001	3	2
S_1_002	1	0
S_1_002	2	3
S_1_002	3	0
S_1_003	1	0
S_1_003	2	0
S_1_003	3	0

59181090 rows X 8 columns

為了使 sales 與 Calendar 及 Sell_price 在合併資料的時候比較容易，先採用 melt 方法重整資料

附錄-資料合併

2

1

Sales (melt)

id	item_id	dept_id	cat_id	store_id	state_id	d	sold
001_CA_1_evaluation	HOBBIES_1_001	HOBBIES_1	HOBBIES	CA_1	CA	d_1	0
002_CA_1_evaluation	HOBBIES_1_002	HOBBIES_1	HOBBIES	CA_1	CA	d_1	0
003_CA_1_evaluation	HOBBIES_1_003	HOBBIES_1	HOBBIES	CA_1	CA	d_1	0
004_CA_1_evaluation	HOBBIES_1_004	HOBBIES_1	HOBBIES	CA_1	CA	d_1	0
005_CA_1_evaluation	HOBBIES_1_005	HOBBIES_1	HOBBIES	CA_1	CA	d_1	0

Salendar

date	wm_yr_wk	weekday	wday	month	year	d	event_name_1	event_type_1	event_name_2	event_type_2	snap_CA	snap_TX	snap_WI
2011/1/29	11101	Saturday	1	1	2011	d_1					0	0	0
2011/1/30	11101	Sunday	2	1	2011	d_2					0	0	0
2011/1/31	11101	Monday	3	1	2011	d_3					0	0	0
2011/2/1	11101	Tuesday	4	2	2011	d_4					1	1	0
2011/2/2	11101	Wednesday	5	2	2011	d_5					1	0	1

Sell_price

store_id	item_id	wm_yr_wk	sell_price
CA_1	HOBBIES_1_001	11325	9.58
CA_1	HOBBIES_1_001	11326	9.58
CA_1	HOBBIES_1_001	11327	8.26
CA_1	HOBBIES_1_001	11328	8.26
CA_1	HOBBIES_1_001	11329	8.26

首先透過 1 將sales、calendar串接

再將其結果透過 2 與 Sell_price 串接

附錄-時間序列的特徵工程

- Component encoding：利用資料的時間相關性，給予對應時間相關的特徵，如年、月、周、星期，或是時、分、秒。
- Characteristics in time series ：
  - Feature Lag N period ：取先前時間點(Lag)的單一數值當作特徵。Ex：以月為週期的資料中，前30天的資料點（Lag30）
  - Feature Lag N periods Aggregate ：取先前一段時間點(Window)的數值，經過統計方式的計算（平均值、最大值、標準差）後當作特徵。Ex：過往七天銷售量的平均，通常代表了銷售量的趨勢。
  - Feature Lag N periods Interaction ：不同時間點資料彼此的變化。Ex：前兩天銷售額的變化（Lag2-Lag1）
- Dummy variables ：在零售行業遇到特殊時節時，營業的整體供給需求會有所變化。如聖誕節、感恩節、SNAP

附錄-Feature - Lag and rolling mean

Date	sold	Lag(3D)
2011/2/1	10	
2011/2/2	3	
2011/2/3	5	
2011/2/4	7	10
2011/2/5	2	3
2011/2/6	6	5
2011/2/7	0	7

Lag by 3 days

Date	Lag X	Rolling_mean(3D)
2011/2/1	10	
2011/2/2	3	
2011/2/3	5	6
2011/2/4	7	5
2011/2/5	2	4.666667
2011/2/6	6	5
2011/2/7	0	2.666667

Rolling mean by 3 days

透過 Lag 計算過往1、7、28天的銷售量，time series 的資料特性，有可能當下數據會因為前一天至前幾天的數據受影響。以左表來說，2011/2/4 銷售量 7 可能會受到2011/2/1 銷售量 10 的影響。

透過 rolling windows 計算 lag 1、7、14、28的往前7、14、28天的平均值。