M5 Forecasting - Accuracy

宋振維

About Me

- **資深數據分析師 @ UDN Group**
- **工作經歷**
  - 集團**個人化推薦系統**專案負責人，並獲頒傑出表現獎
  - 利用**深度學習**方法，建立個人化新聞推薦模型，提升成效 1.3
  - 建構並優化精準廣告模型，提升 EDM 開信率達 80%、提升數位廣告點擊率 50%
  - 規劃與開發**推薦系統雲端架構**（數據處理與儲存、模型訓練與部署、API 服務）
  - 開發集團**會員儀表板**，輔助集團各單位制定會員的發展策略

M5

CONTENTS

目錄

01

競賽內容介紹

競賽說明．數據介紹

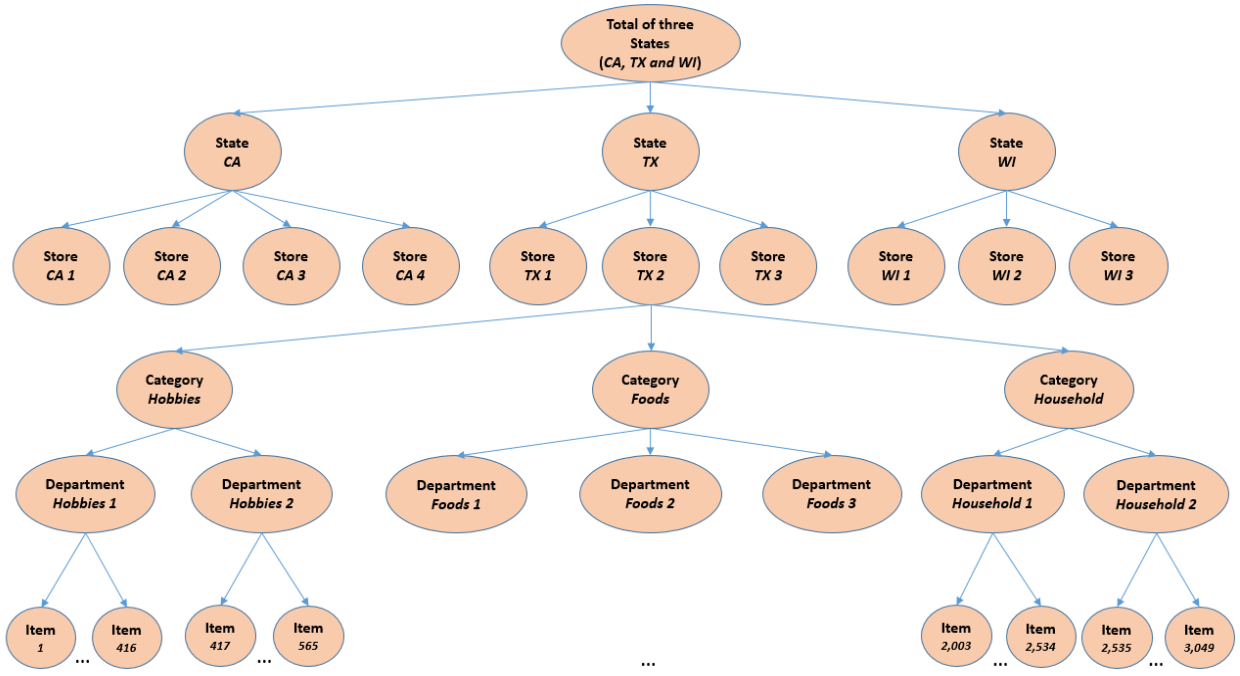
競賽說明

目的

- 替 42840 組商品銷售量的時間序列提供最準確地預測方法

內容

- 層級的單位銷售數據
  從產品商店級別，可再依據產品部門，產品類別，商店和州等類別匯聚成 12 個層級
- 除銷售數據外，另提供解釋性變量（售價、特殊事件、星期等）



數據介紹

sales_train.csv

- *item_id*：產品的 ID (3049 個)
- *dept_id* ： 產品所屬的部門 ID (7 個)
- *cat_id* ：產品所屬的類別 ID (7 個)
- *store_id* ： 銷售的門市 ID (10 間)
- *state_id* ： 門市所在的州 ID (3 個)

calendrer.csv

- *event_name_1/2*：事件名稱
- *event_type_1/2*：事件的類型
- *snap_CA/TX/WI*：各州是否允許 SNAP 購買

sales_price.csv

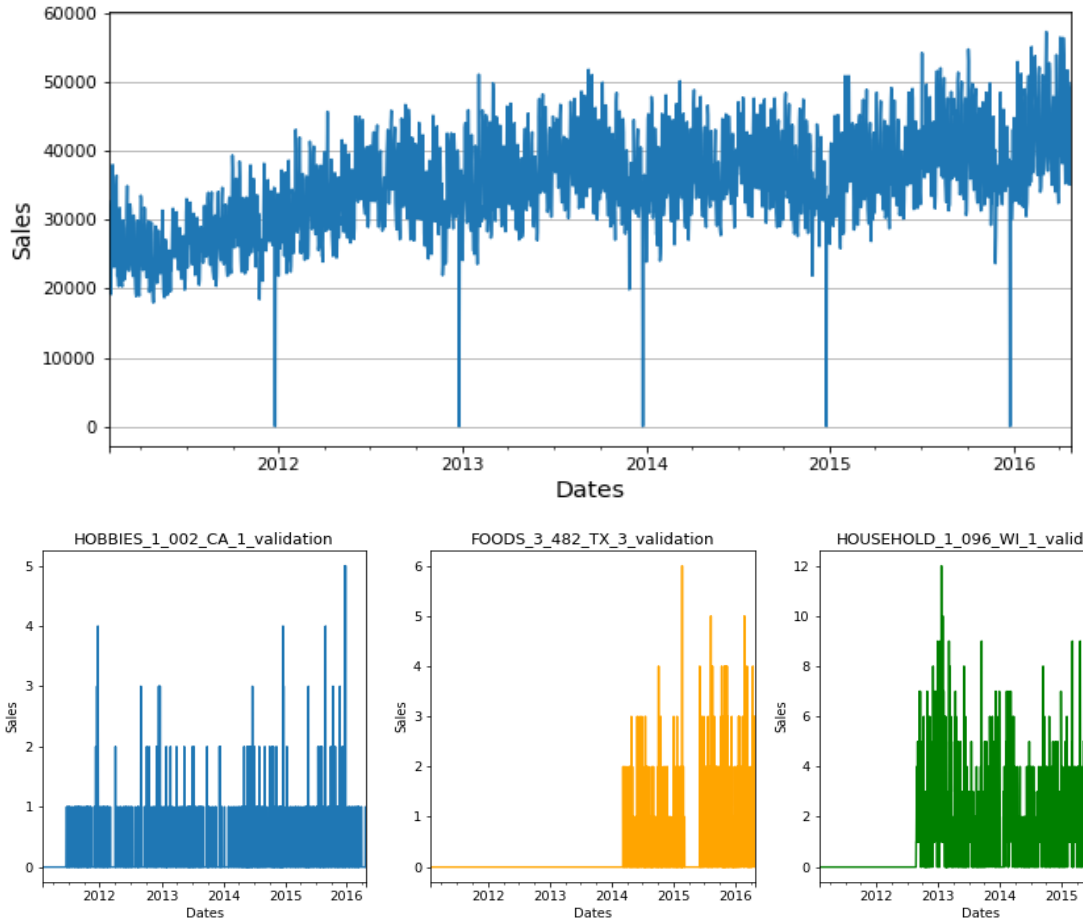
- *store_id, item_id*：每間店各產品的價格不同
- *wm_yr_wk*：價格以週為單位(7天的平均)
- *sell_price*：若無銷售則不會有價格

02

探索性數據分析

分層分析 · 交叉分析 · 時間週期 · 事件影響

分層分析



商品的各店銷售

- 大多數商品每天的銷售量很低
- 商品並非每段時間都有銷售量，時間序列有間斷的狀況

整體匯總

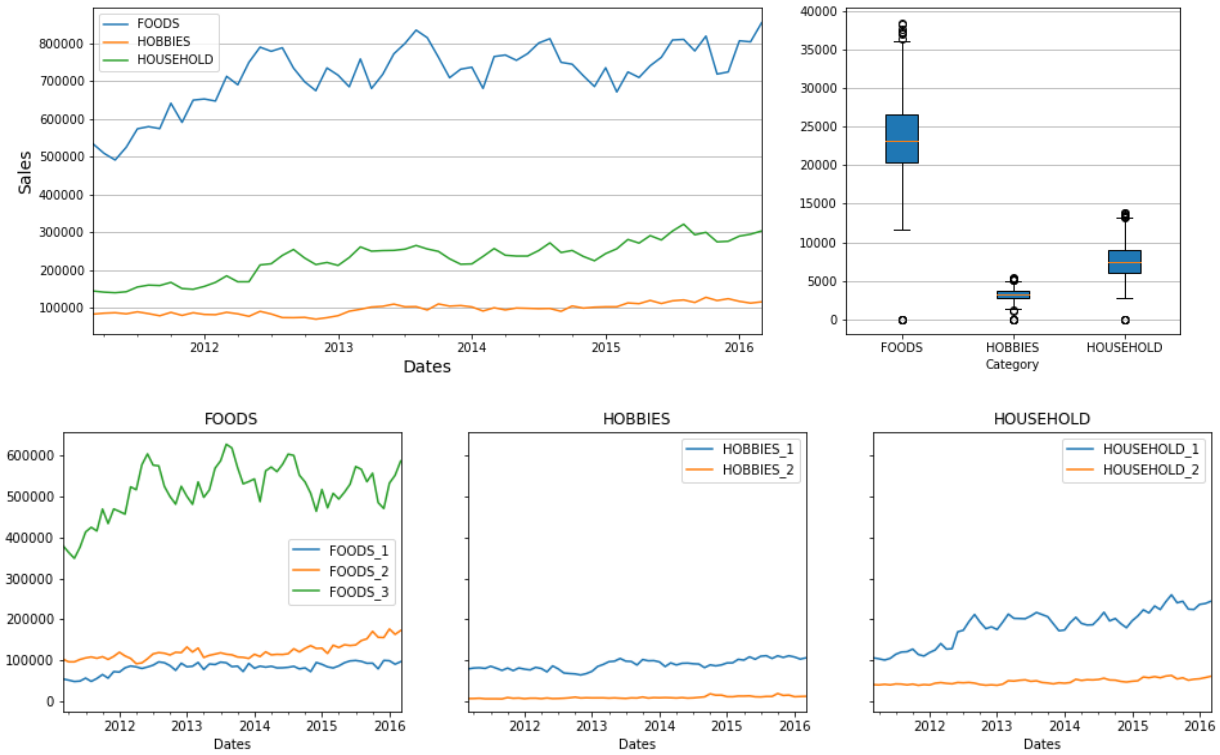
- 產品銷售具在每年都有季節性變化
- 每年聖誕節為異常值銷售量低

分層分析

各州/各店銷售

- 加州具有較高的銷售量，尤其以 CA_4 最為明顯
- 威州與德州銷量曾相當接近，但於威州於 2015 下半年反超德州
- 加州 4 店的銷量差距明顯
- 德州 3 店銷量走勢相當接近
- WI_2 銷量在 2012 年中跳躍性成長

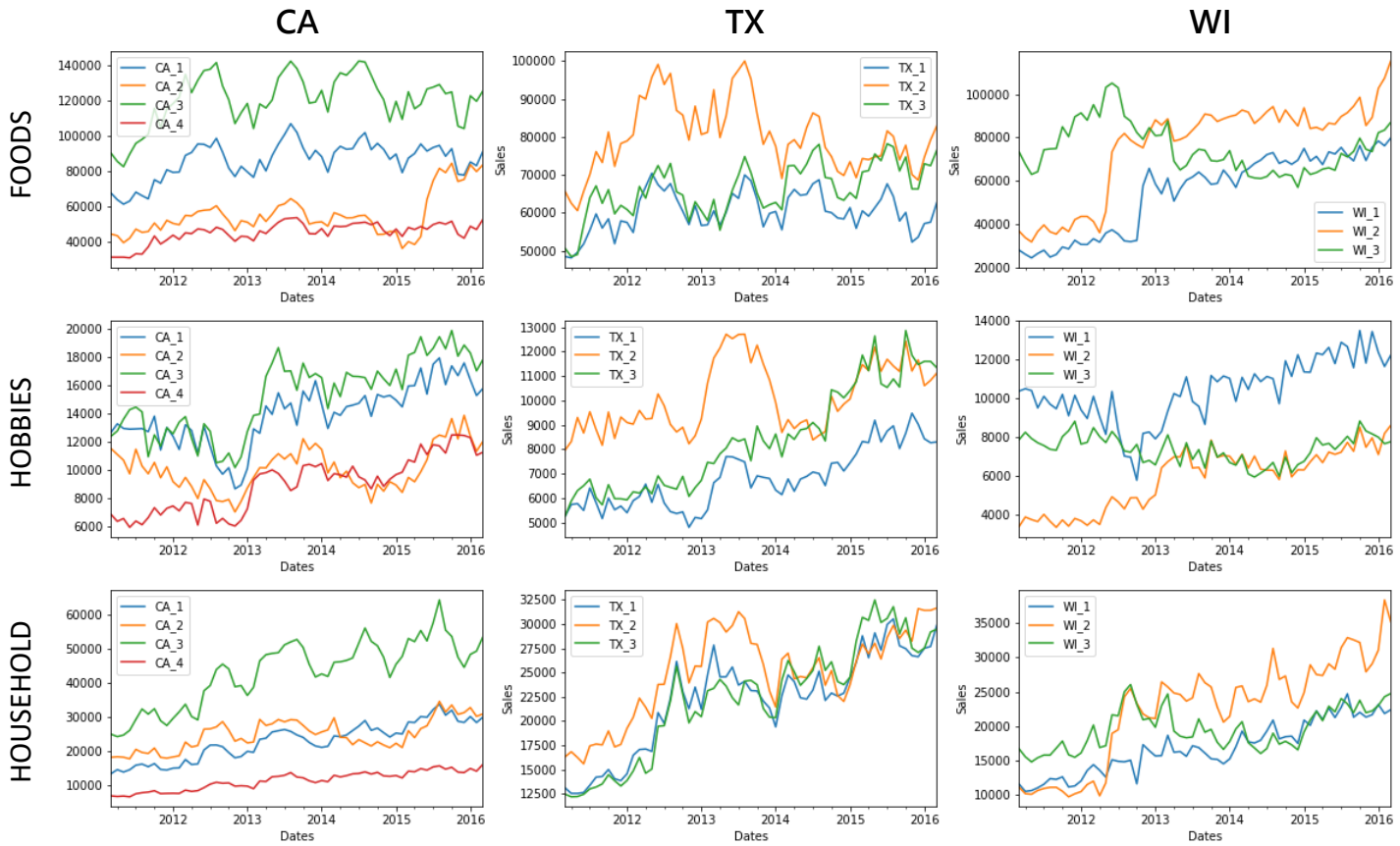
分層分析



各商品類別/部門銷售

- FOODS 銷量相對高很多，其次是 HOUSEHOLD 類，HOBBIES 則明顯低很多
- 各商品部們中，皆以類別 1 具有較高的銷量
  - FOODS_1 銷量差距最大，且具有明顯週期性
  - HOUSEHOLD_1 成長趨勢明顯

交叉分析

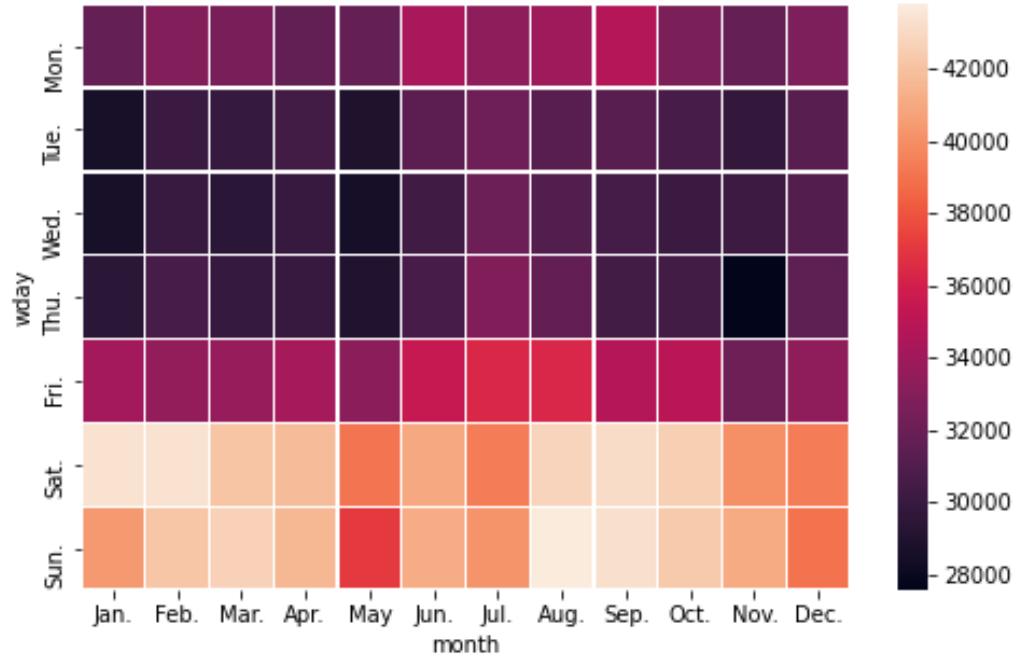


商品類別 X 各店銷售

- TX_2 店的 FOODS 與 HOBBIES 商品在 2012~2014 年銷量明顯較高
- 各州在 HOUSEHOLD 商品上，每間店的銷售走勢接近
- WI_2 店在 2012 年 FOODS 與 HOUSEHOLD 銷售皆跳躍性成長
- WI_1 店在 HOBBIES 類銷量優於威州其他店面，但在其他商品類別卻相反

時間週期

銷售熱點圖

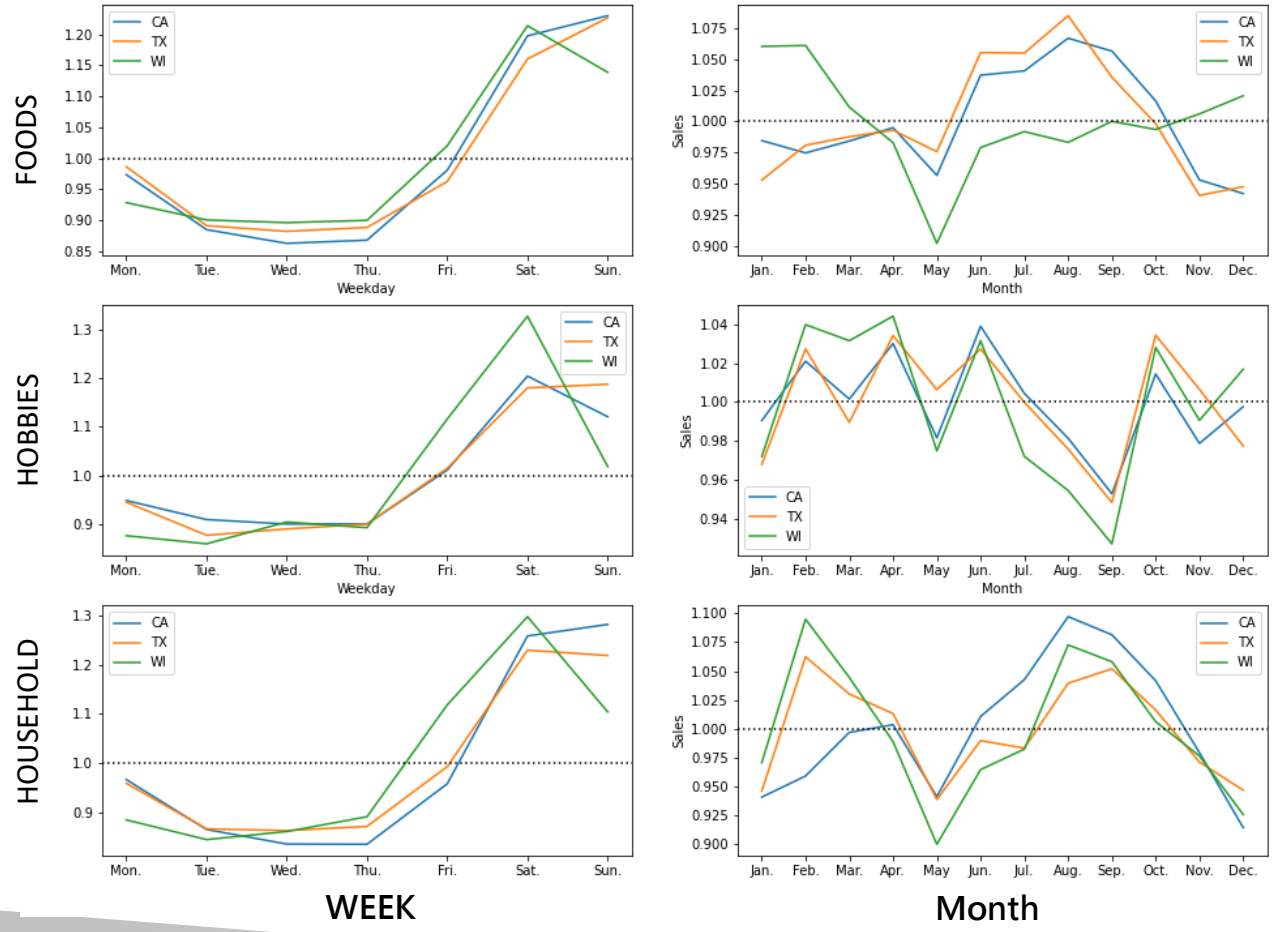


Month X Weekday

- 人們傾向**周末假日**進行消費
- 冬天 **11, 12, 1 月**與夏季的 **5, 6 月**銷量較低，尤其在週間最為明顯

時間週期

相對購買力比較

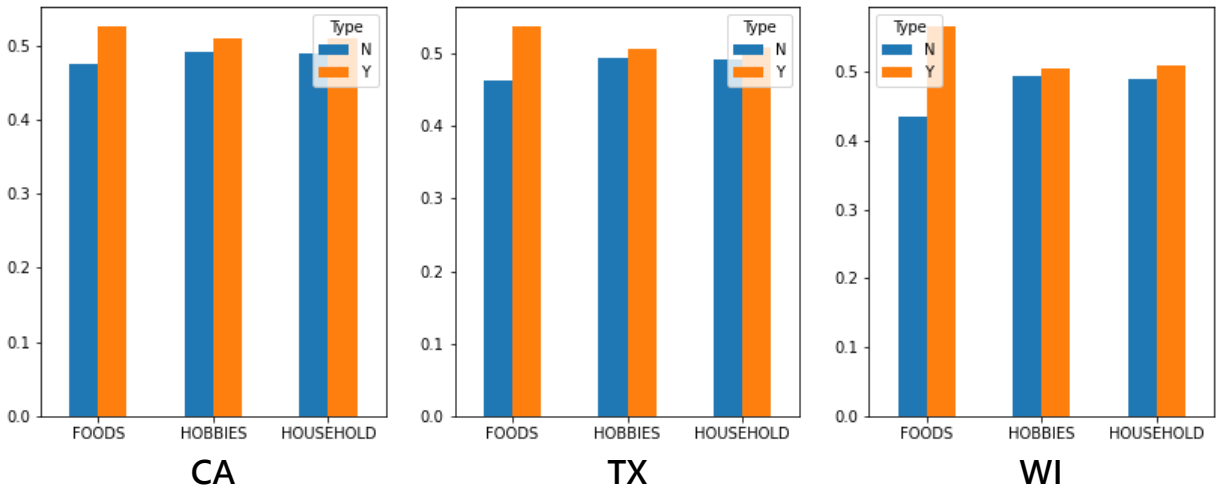


商品類別 X 州

- 相對購買力 = 當日銷量 / 平均銷量
- 不分類別人們**傾向周末**進行消費，但威州的週日相對購買力較低
- 人們在夏季 (6~9 月) 對於 FOOD 與 HOUSEHOLD 的需求較強，但對 HOBBIES 相反

事件影響

有無 SNAP 銷售量占比

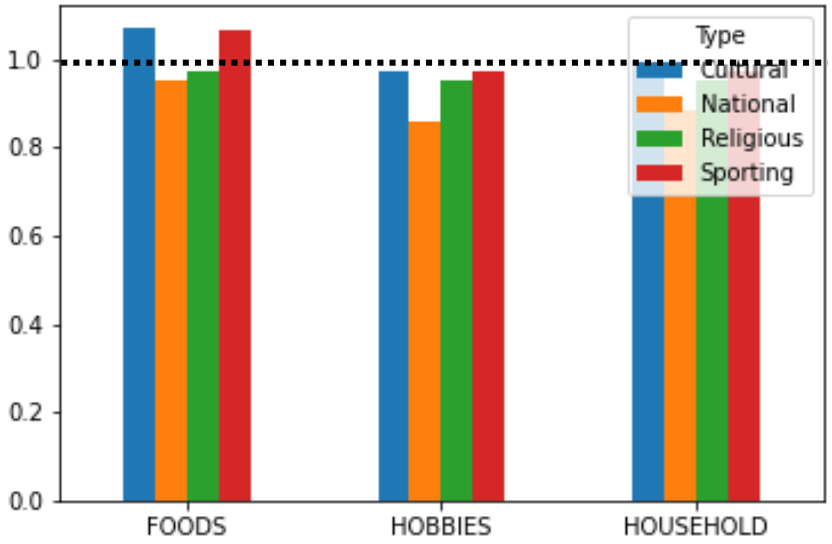


SNAP

- 補充營養援助項目，替窮苦家庭提供食品經濟補貼
- SNAP 顯然對**德州與威州在 FOODS 類別**的銷售有影響，加州的落差則不明顯

事件影響

相對購買力比較



活動與節慶

- 相對購買力 = 銷量 / 各產品類別平均銷量
- 當碰到**文化節慶**(情人節、萬聖節、母親節)或**運動賽事**(季後賽、超級盃)時，食物的需求會相對提升

03

模型建構與預測

模型方法．特徵工程．預測評估

模型方法

選用策略

- 選項：傳統時間序列 / 機器學習 / 深度學習
- 納入銷量時序以外的**多元特徵**
- 高準確率前提下，保有**訓練速度快**且**記憶體使用低**

} *LightGBM Regression*

模型架構

- 各間店或每個類別的商品銷量走勢不盡相同
- 一次預測 28 天內的銷量，而非每天獨立預測
- 在每個 iteration 建立**樣本抽樣**與**特徵抽取**，避免 overfitting

特徵工程

類別層級

- 產品資訊：產品 ID、產品部門 ID、產品類別 ID
- 門市資訊：門市 ID、州別 ID

時間與事件

- 事件：事件名稱、事件類型、是否有 SNAP
- 時間：星期、相對年份、月、週、日、週末
- 其他：產品銷售時長

特徵工程

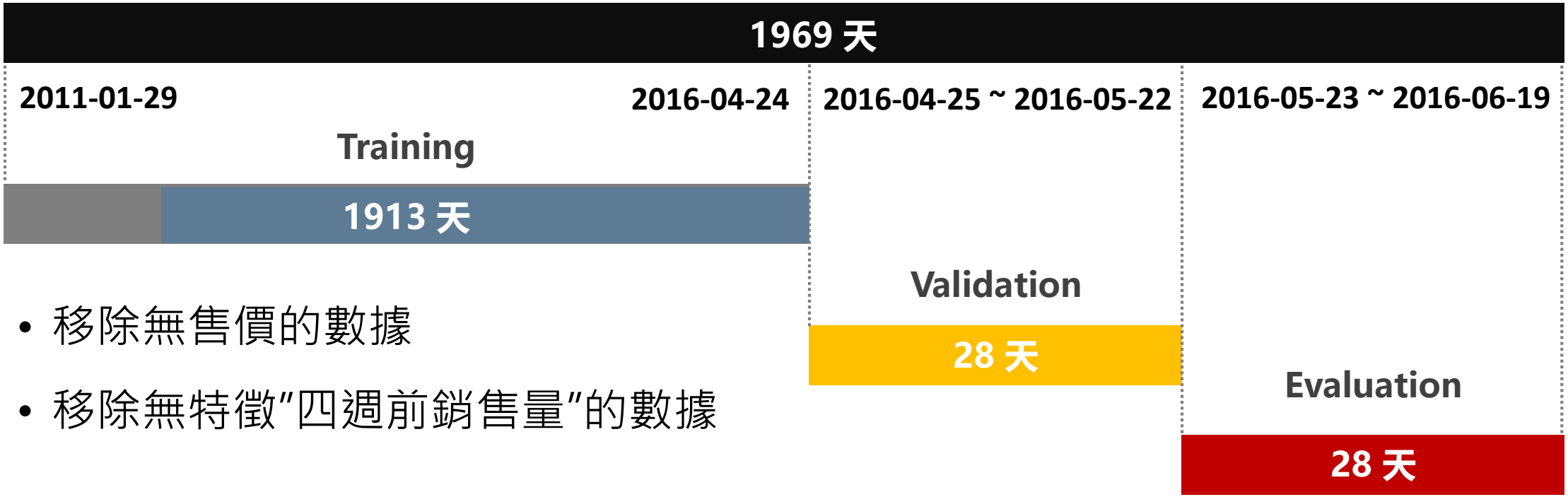
產品售價

- 產品售價
- 產品最高/低售價
- 產品平均售價
- 產品售價標準差
- 同時銷售的同類產品數
- 同時銷售且價格一致的同類產品數
- 產品價格成長率 (日、週、月)
- 相對同部門產品的售價

產品銷售量

- 四週前銷售量
- 各 州/店/產品類別/產品部門/產品
  → 四週前平均銷量
- 各 州/店/產品類別/產品部門/產品
  → 四週前銷量標準差
- 每個產品類別在各 州/門市
  → 四週前平均銷量
- 每個產品類別在各 州/門市
  → 四週前銷量標準差

資料期間



- 移除無售價的數據
- 移除無特徵"四週前銷售量"的數據

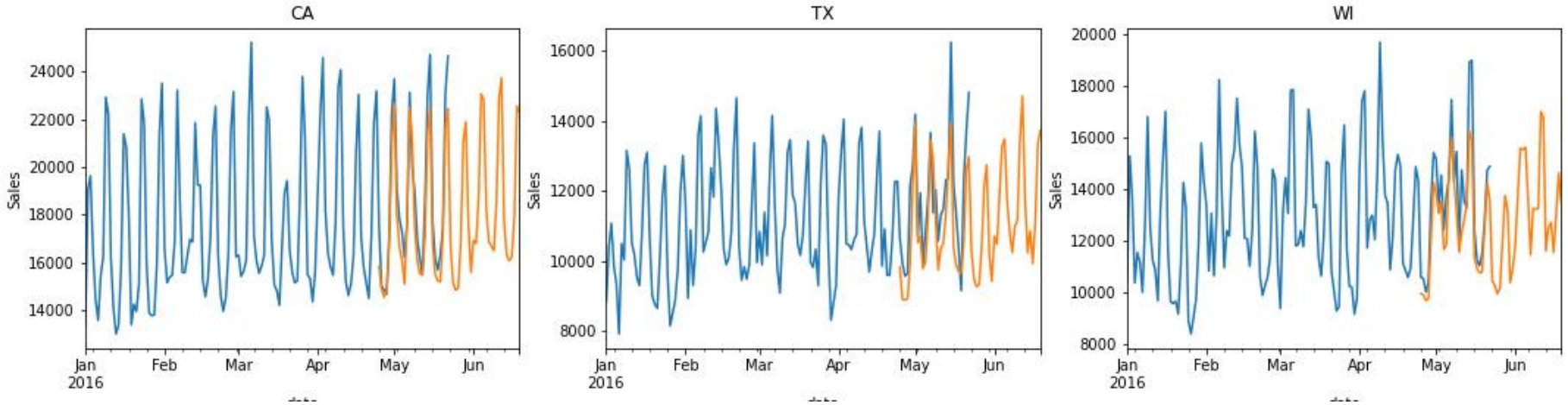
預測結果

		100 Trees	300 Trees	500 Trees	1000 Trees
門市 + 產品類別	**有銷量特徵**	**0.74772**	**0.66673**	**0.64552**	**0.62364**
門市 + 產品類別	無銷量特徵	0.77352	0.76431	0.71063	0.65676
門市	有銷量特徵	0.86045	0.71135	0.68681	0.65429

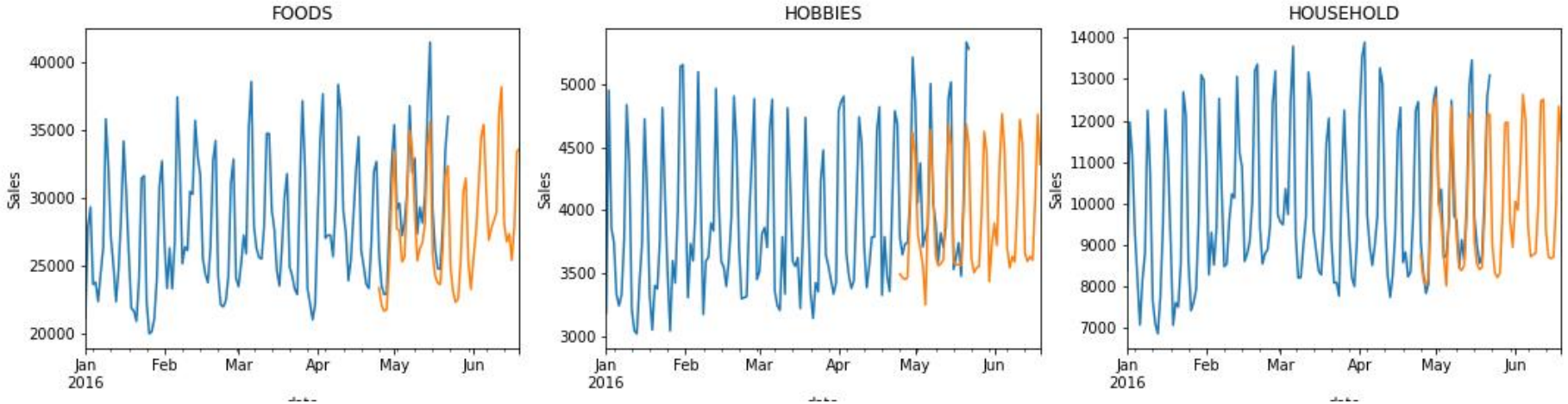
- 3 種策略皆可利用增加樹的數量來提升預測效果也
- **歷史銷量**提供更多資訊給模型，但也必須利用更多的樹來提升預測效果
- 更**細粒度**的模型建立，對於銷售的預測預準確

預測結果

州別


CA
TX
WI
Sales
Jan
2016
Feb
Mar
Apr
May
Jun

產品類別


FOODS
HOBBIES
HOUSEHOLD
Sales
Jan
2016
Feb
Mar
Apr
May
Jun

04

結論與展望

可優化項目．延伸應用

結論與展望

Next Step

- 對銷量時序進行拆解

  → 加入特徵 or 預測趨勢

- 加入外部數據（如天氣）
- 各模型進行各自的特徵篩選
- 嘗試以 DNN的方式（e.g. Seq2Seq）

延伸應用

- 預測 ATM 的**提款**或**整體**使用需求

  → 補鈔的時點與量

  → 決定最佳機器維護時點

- 自然人/法人還款金額預測
- 網站流量預測 → 推薦清單更新時機

附件

程式碼

- https://github.com/royalucifer/m5_accurancy_competitions