

線上消費性產業 資訊分析系統

—以電商平台 Olist 為例

指導老師：黃登揚、蔡智勇老師

成員：鄭偉鏢、梁雪樺、陳偉祥、李儒育

黃怡家、簡智弘、何舜華、王怡文



2021/12/29 財團法人自強工業科學基金會

TEAM MEMBER



何舜華
物流小組



梁雪樺
銷售小組



黃怡家
評價小組



王怡文
物流小組



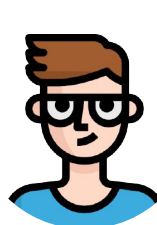
陳偉祥
銷售小組



簡智弘
評價小組



李儒育
銷售小組



鄭偉鏢
組長



專題成果網站

<https://ec-study.allen-cheng.com/>

議程

- ▶ 專題：目標、挑戰、實作
- ▶ 成果：
 - ▷ 銷售組
 - ▷ 物流組
 - ▷ 評價組
- ▶ 總結：將成果設計成完整的營運系統

專題的目標

用這門課教授的技術
完成數據分析與機器學習的研究案例

最大限度地讓每個組員都能參與到所有環節

不只單純的研究成果
更要真實的呈現歷程

專題的挑戰

等同三組專題的工作量

除了原始資料相同以外，各小組的
觀點、想法、研究方式、模型全部不同

同時管理三個專題的資訊、議題、任務與目標

如期完成外，還要將成果整合



Object



How we made it?



Deadline

專題的實作

展現「勝任工作的能力」

把專題當業界工作裡的「專案」看待

用專案管理思維，模擬真實工作環境

「邊移動邊開火」逐步推進完成度

參考過去從事軟體工作使用過的「敏捷開發」

減少障礙、提升效率、緊盯進度如期完成

Olist電商概況

Olist成立於2015年，致力於將小型零售商與各大電商網站連接

超過10,000商家加入，平台以收取月費、產品佣金、運費為商業模式

銷售於180個國家、10幾個不同的購物網，包括巴西前三大及Amazon

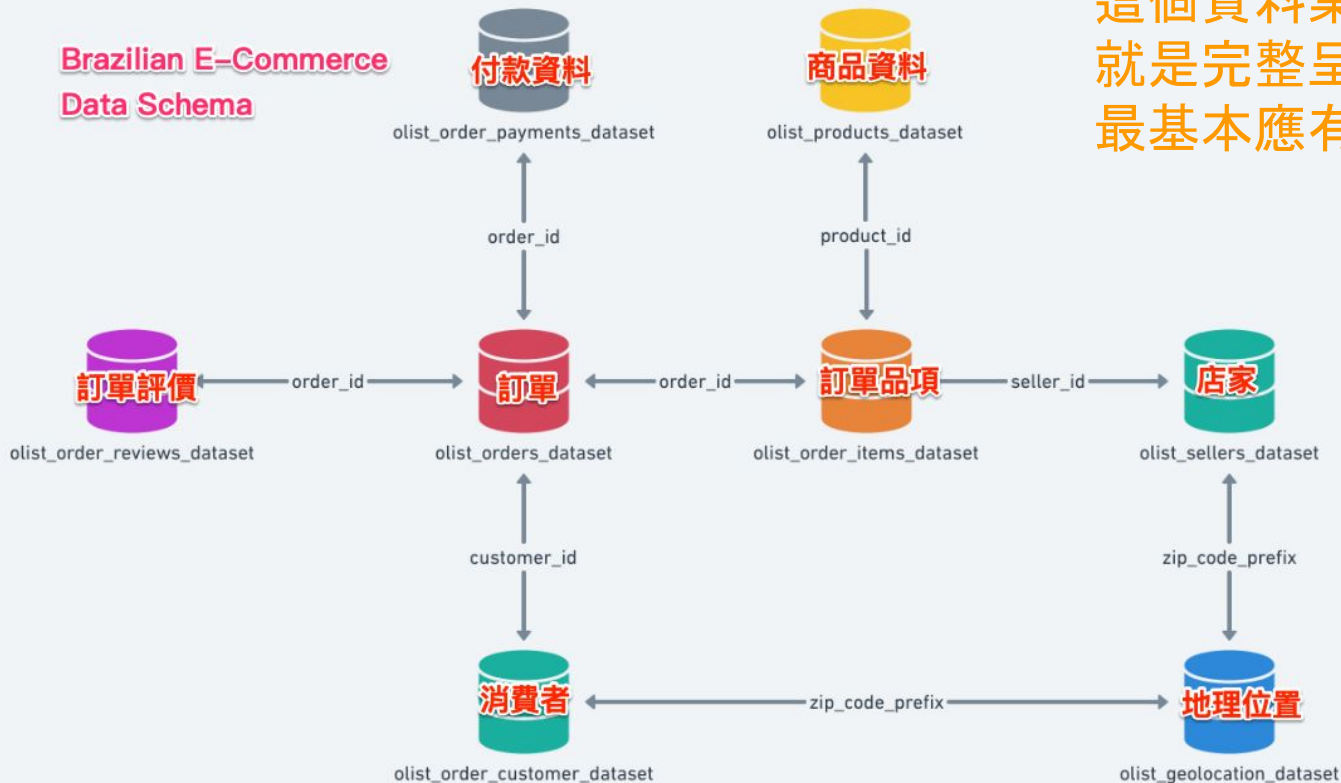
物流、銷售管理、廣告投放、客戶服務等一站整合式服務



Olist dataset (連結)

8 個表格、52 個欄位、約9 萬多筆資料

這個資料集最大特色
就是完整呈現了一個電商平台
最基本應有的資料結構



資料分析架構



kaggle

MySQL™

python™

pandas

scikit
learn
Machine Learning with Scikit-Learn

colab

jupyter

AMCHARTS

HTML

研究議題





各組的成果

銷售組

影響銷售量的關鍵因子

陳偉祥、李儒育、梁雪樺



研究動機

分析出**關鍵銷售量因素**，將會在實務上的效果帶來更大利益。

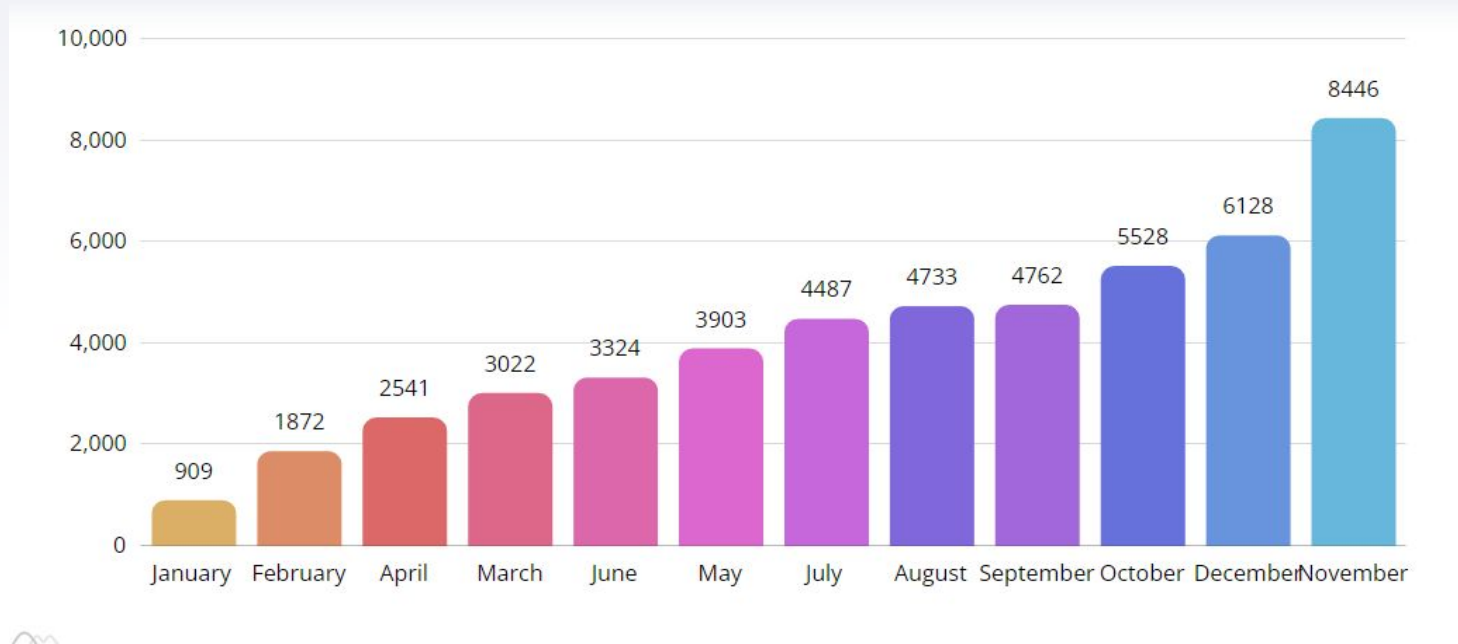
- ▶ 能夠更快速應對消費者的需求，**針對痛點規劃營運及精準行銷**
- ▶ 有助與商戶**聯合推出行銷方案**
- ▶ 了解有實力商戶類型，**有助開發商戶上架**



詳細說明:

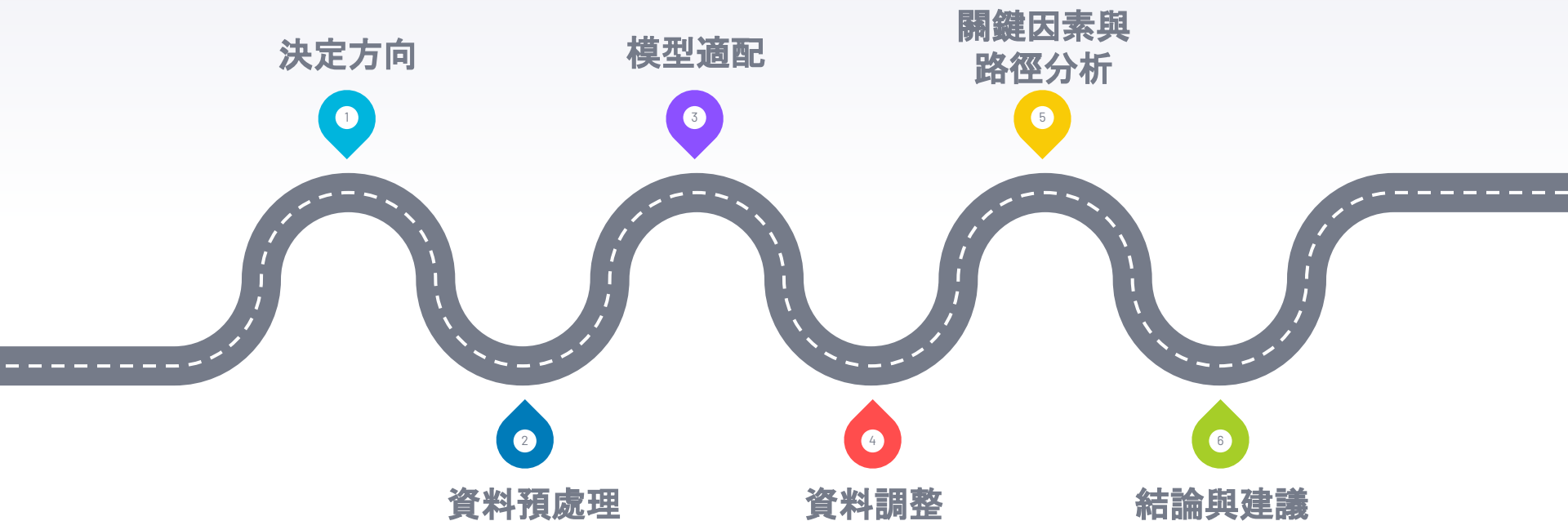


銷售量趨勢圖



1. 銷售量**集中在Q4**, 佔全年40%
2. 找出銷售量成因可助 **提高其他時段銷量**

資料處理及分析流程

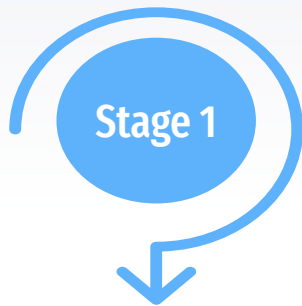


研究過程





資料預處理流程



連結資料表

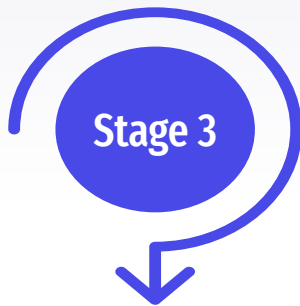
將kaggle Olist資料集有關
運營資料整理成一份綜合
資料

-資料欄位共65



空值資料處理

確認資料是否有 NA 值及
刪除



離群值處理

1. 對欄位做敘述統計分析
2. 窺探資料分布情形並將極端值刪除
3. 刪除與分析較無關的欄位

-剩餘欄位共26欄



特徵轉換

1. 為賣家銷售量重新編碼。Q3以上=高 Q1以下=低,其餘刪除
2. 利用單價、交易量進行潛在類別分析(LCA)將相近進行歸類,71類歸類為6類產品
3. RFM分析, R(最近的一次消費), F(消費頻率), M(消費金額)來評量客戶潛在價值
4. 利用交易量高低、產品單價高低生成出4種商品型態

-轉換後欄位共30欄

資料預處理-特徵轉換

RFM

R 最後消費日 	F 消費頻率 	M 總消費金額 
較近	高	較高
較遠	低	高
較近	高	低
較遠	低	低



4級客戶

3級客戶

2級客戶

1級客戶

商品型態

 交易量	 單價	 形態
大	高	獲利商品
大	低	薄利多銷商品
小	高	奢侈商品
小	低	淘汰商品

4級商品

3級商品

2級商品

1級商品



資料處理後的變數

訂單相關：

item_countorder_purchase_time_day
ororder_purchase_timeday
ororder_purchase_day
order_purchase_dayofweek
order_purchase_hour
order_purchase_month
order_purchase_year
total_freight_value
total_item_price
total_payment_value

買賣家相關：

customer_state_region_type
RFM_type
seller_level
seller_state_region_type

產品內容相關：

product_category6
product_type4

物流相關：

delivery_efficiency
estimated_logistics_using_hours
geo_distance
is_delivered_delayed
is_shipping_delayed
logistics_delay_hours
total_delivered_waiting_day
total_logistics_using_hours
total_package_volume
total_package_weight_g

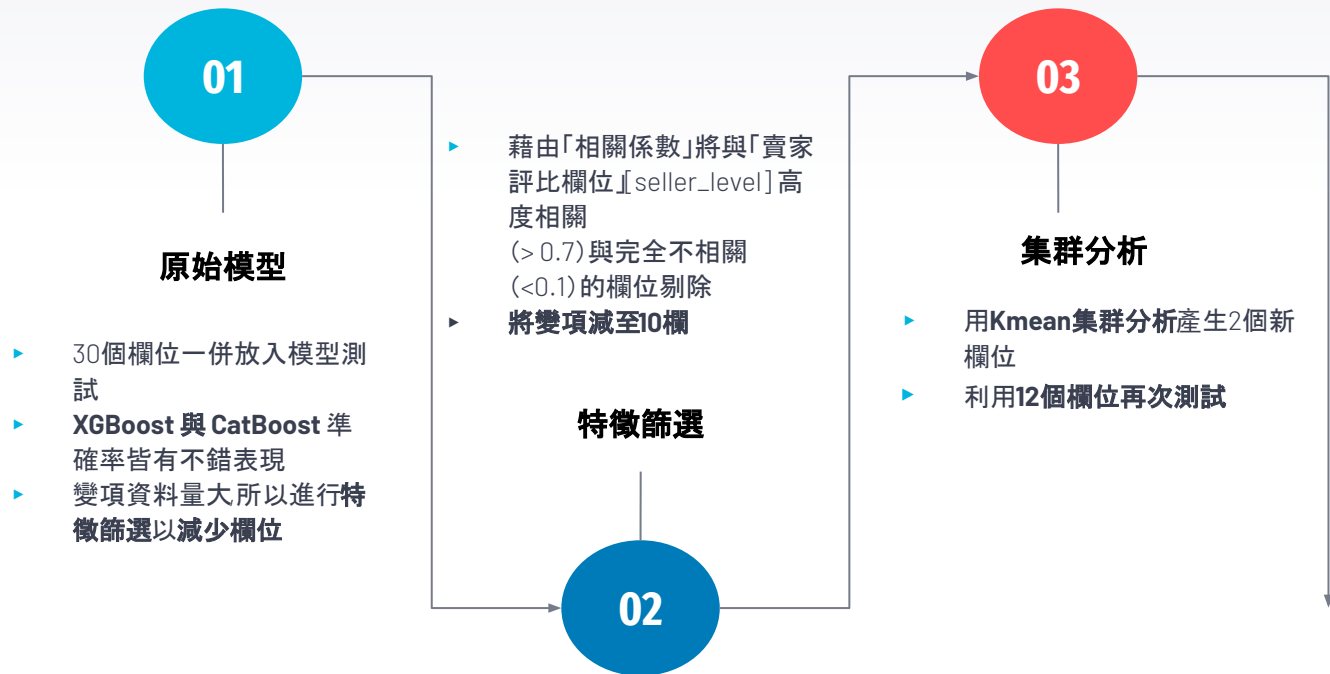
買家評分相關：

review_score
review_type



依變數: seller_level

機器學習



資料調整-集群分析

KMean Value 集群分析應用:

欄位[total_item_price]、[total_payment_value]、[total_freight_value]
進行集群處理成新變項:

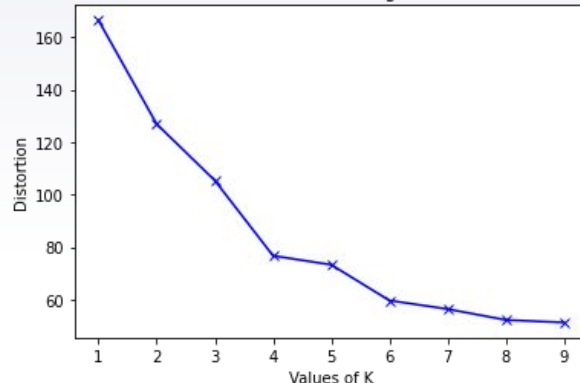
- ▶ 透過Elbow Method 匯出下圖,發現其值於第4類後的差異趨緩,故選擇分4群
- ▶ 透過Kmean 集群分析進行非監督式學習獲得新變項[Kmean_value]

欄位[total_freight_value]、[total_package_volume]、[geo_distance]
[total_package_weight_g]、進行集群處理成新變項:

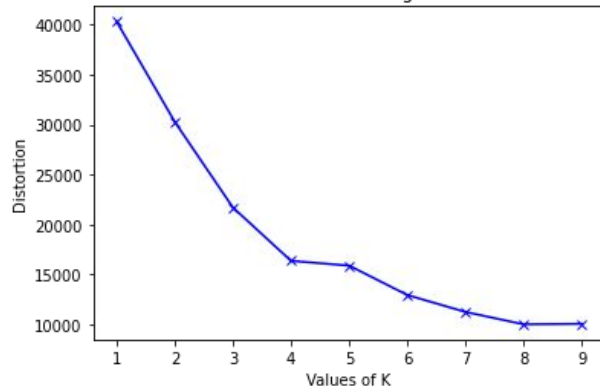
- ▶ 透過Elbow Method 匯出下圖,發現其值於第4類後的差異趨緩,故選擇分4群
- ▶ 透過Kmean 集群分析進行非監督式學習獲得新變項[Kmean_package]



The Elbow Method using Distortion



The Elbow Method using Distortion



研究結果



模型適配

挑選模型及測試結果:

- ▶ 在**CatBoost**測試前後變化約**+0.27%**,變項減少**18欄(-40%)**。欄位縮減後提升機器學習的分析效率,且準確率也上升。
- ▶ **CatBoost**在效能上比 **XGBoost** 和 **LightGBM**等Boosting方法 更優,他同時支援 CPU 和 GPU 運算,提高分析效率。
- ▶ **CatBoost**解決了梯度偏差 (Gradient Bias) 以及預測偏移 (Prediction shift) 的問題,減少Overfitting發生,提高算法的準確性和泛化能力。

程式碼:

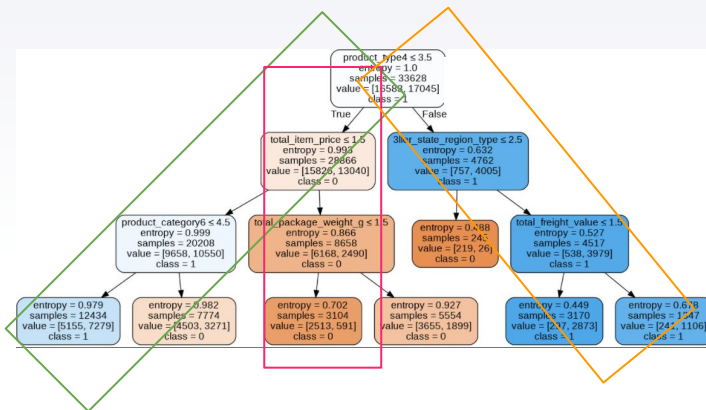


	X:30 原始model	X:10	X:12
模型	測試準確率	測試準確率	測試準確率
CartTree	0.688	0.689	0.682
ExtraTrees	0.702	0.699	0.709
XGBoost	0.799	0.774	0.792
CatBoost	0.726	0.720	0.728

影響銷量關鍵特徵

決策樹的模型顯示, 產品型態為首要關鍵, 其次之為價格、地區、品類及運費

- ▶ 方向1(橙色): 平台賣方所在的地區於巴西屬於人口密集(63%)及買家偏向低運費
- ▶ 方向2(紅色): 買家對價格敏感度較高, 接受度普遍在USD\$92以下
- ▶ 方向3(綠色): 買家偏好於價格在USD\$92以下的商品及某幾項品類



主關鍵
產品型態



次關鍵
運費費用
賣家地點
價格實惠
產品類別



購買綜合因素

詳細說明:



建議



結論與建議

詳細說明:



技術結論：經特徵工程後，模型欄位縮減**-40%(提升機器學習效率)**且準確度提高**+0.27%(提升機器學習準確度)**



倉儲物流: 區域設物流倉庫, 針對某品類增加庫存

開發商戶: 針對不同品類開發更多價格合適商戶進駐

免運費活動: 免運費活動設定免運的門檻可增加買家客單價及下單意慾

主題活動: 非旺季時定期以生活用品周 3C家電周等主題舉辦平台活動

品牌形象: 嬰幼童用品、休閒保健可引入更多品牌及大品牌

專業徵才: 在找業務開發時, 針對熟悉最受歡迎的品類的人員進行招聘。
嬰幼童用品、休閒保健需要對品牌熟悉度高的人員

物流組

預測訂單是否延遲？

何舜華、王怡文、鄭偉鑠



研究動機



雙11貨物爆量「卡關」 電商：延遲3至5天配達

記者 鍾潔科 林政鑫 / 攝影 何佳璵 報導
發佈時間：2021/11/18 18:18
最後更新時間：2021/11/18 19:30



雙11購物節買氣強強滾，不過很多消費者抱怨，到現在都還沒收到包裹！送貨卡關狀況非常嚴重，就連賣家也大喊無奈，明明收到訂單就出貨了，卻因為物流塞車連被買家詬會速度太慢，電商平台也在官網公告，因為貨量實在太多，會延遲3至5天配達，要大家耐心等待。

電商主打「24小時到貨」 屬「交易重要事項」

三級警戒民眾宅在家，電商訂單爆滿，物流公司貨送不完，消費糾紛頻傳！

行政院消保處統計，從5月中旬，已累計86件延遲或沒收到貨投訴，第一名是PChome 24H有58件，佔案件量67%。第二名是蝦皮16件，以及富邦momo有9件。

有網友表示，自5/25在PChome 24H下單，卡貨近一個月，仍未收到貨。也有網友無奈說，「廣告24hr，已經變成240hr了」。在臉書上，也有人成立「[每日關心PChome 24h出貨進度](#)」，以戲謔梗圖揶揄PChome 24H的塞貨狀況。

有71%消費者認為快速到貨能大幅提升下單意願

Olist市場研究表示，購買撤回的原因，有55%延遲交貨是導致

Olist市場研究表示，不到24小時內發貨的產品，轉化率平均提高了48%

雙11檔期，消費者平均多等3~5天，甚至長達一個月才能拿到包裹

研究目的

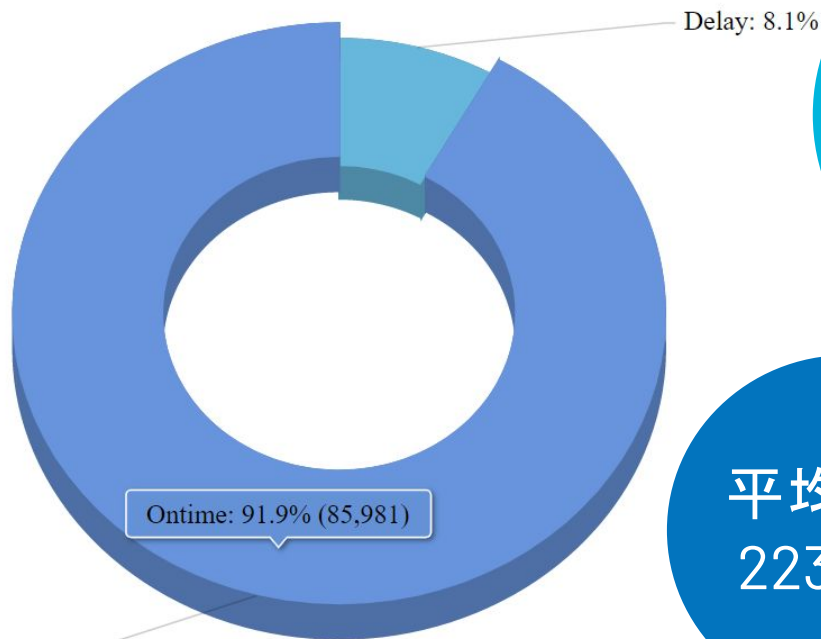
物流3大重點：快！省！準！



電商賣家

機器學習預測「新的訂單出現時，是否會延遲」，以提早做應對

訂單延遲



Delay: 8.1%

2017年
11月24日
訂單延遲
數最多

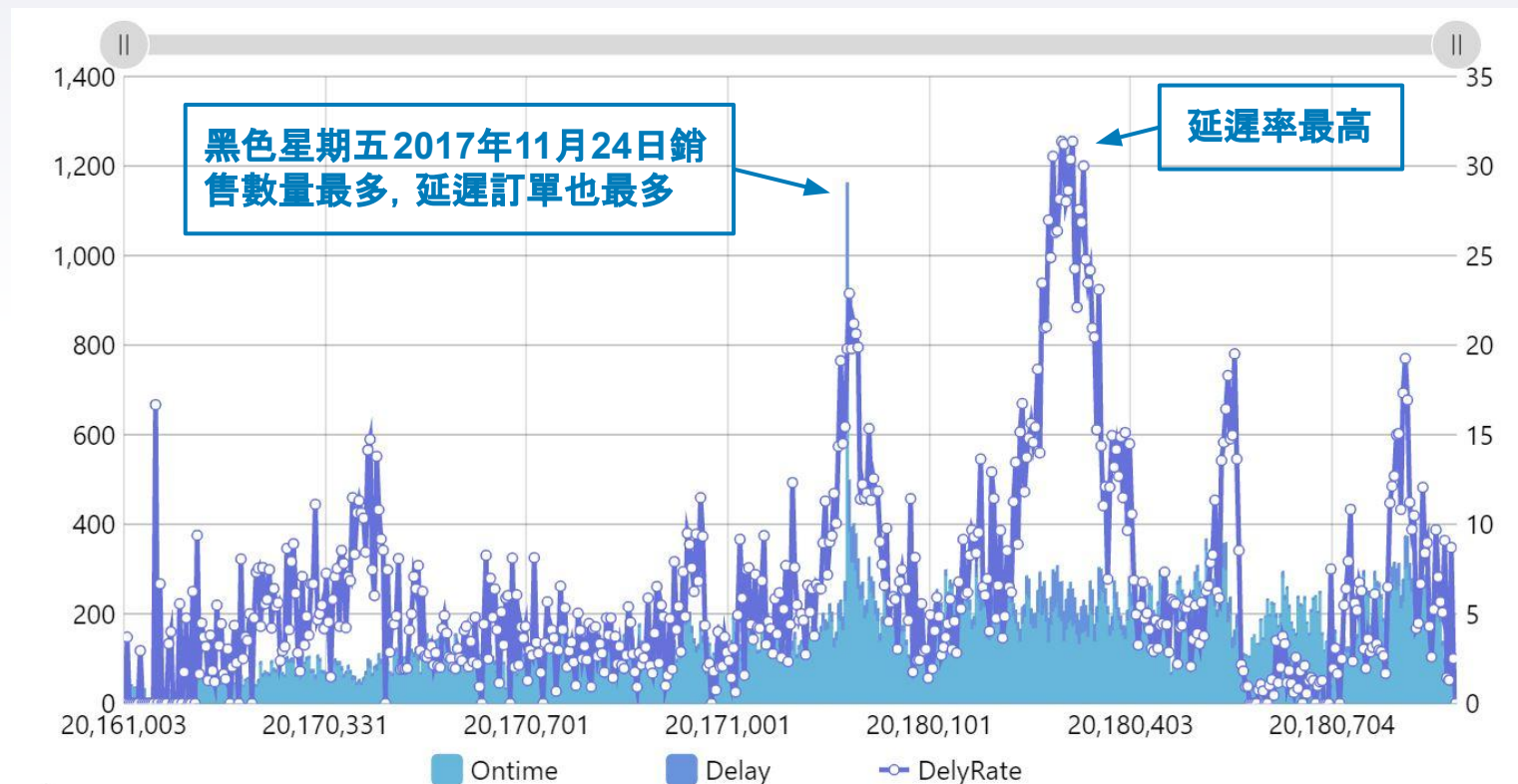
辦公用品
平均等待
時間最久

平均等待
223小時

消費者
從下訂單到
取貨需要等
約9-10天

Delay 8.1% Ontime 91.9%

每日訂單延遲次數及延遲率



▲2月巴西嘉年華、12月聖誕節期間，延遲訂單、延遲率明顯變多、變高

變數選擇

Package_volume
Package_weight

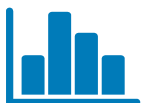


Item_count
Category_type
Category_name

Seller_state(類別)
Customer_state(類別)
Geo_distance

Item_price
Item_freight_value
Total_payment_value

purchase_yearweek
purchase_month
purchase_dayofweek
purchase_Time_day(類別)



原始共15欄, 97,879筆資料

資料預處理

Stage1

空值處理

- ▲ 刪除空白或NULL欄位

Stage 2

資料清洗

- ▲ 觀察資料分布情形
- ▲ 刪除離群值

Stage 3

類別型編碼

- ▲ 將 Seller_state、Customer_state、Time_day轉成Label Encoding或One Hot Encoding

模型調校

最原始dataset

Logistic_analytics_v0
15欄位, 因數據不平衡
(非延遲:延遲=10:1)
模型呈現失衡
延遲訂單準確率: 9%

1

	測試集 預測準確率
非延遲訂單	0.95
延遲訂單	0.19

資料標準化後 dataset

Logistic_analytics_v2
擷取常態分配的 20% ~ 80%
找出主要訓練模型
Random forests or CatBoost
延遲訂單準確率: 20 ~ 29%

3

	測試集 預測準確率
非延遲訂單	0.96
延遲訂單	0.09

2

資料平衡化後 dataset

Logistic_analytics_v1
訓練集資料做平衡採樣
(非延遲:延遲=7:3)
延遲訂單準確率: 15 ~ 20%

	測試集 預測準確率
非延遲訂單	0.94
延遲訂單	0.29

特徵工程



特徵提取

增加新變項: 考量當地季節與文化

▲ 節慶(is_festival)

12月聖誕假期
2月嘉年華月
3-4月聖周慶祝活動
5月母親節
6月同志大遊行與聖約翰節
8月父親節
11月黑色購物節



▲ 雨季(is_rain)

12-4月

	RF 測試集 預測準確率	CAT 測試集 預測準確率
非延遲訂單	0.94	0.93
延遲訂單	0.32	0.31

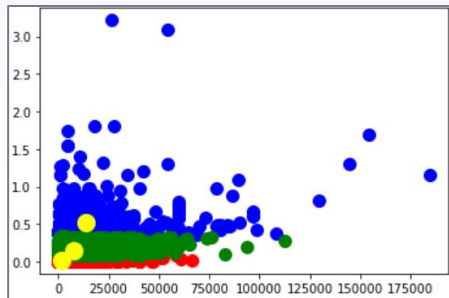
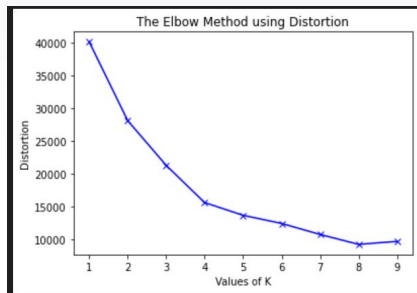
特徵工程

增加新變項: 集群Cluster

▲ 集群1

Package_volume
Package_weight
Item_price

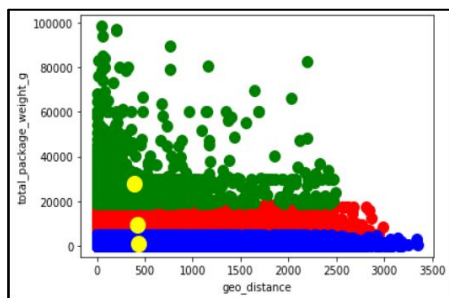
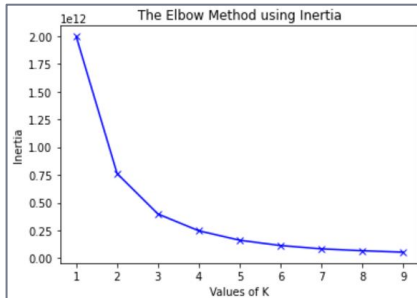
K-means 分3群



▲ 集群2

Geo_distance
Package_weight

K-means 分3群



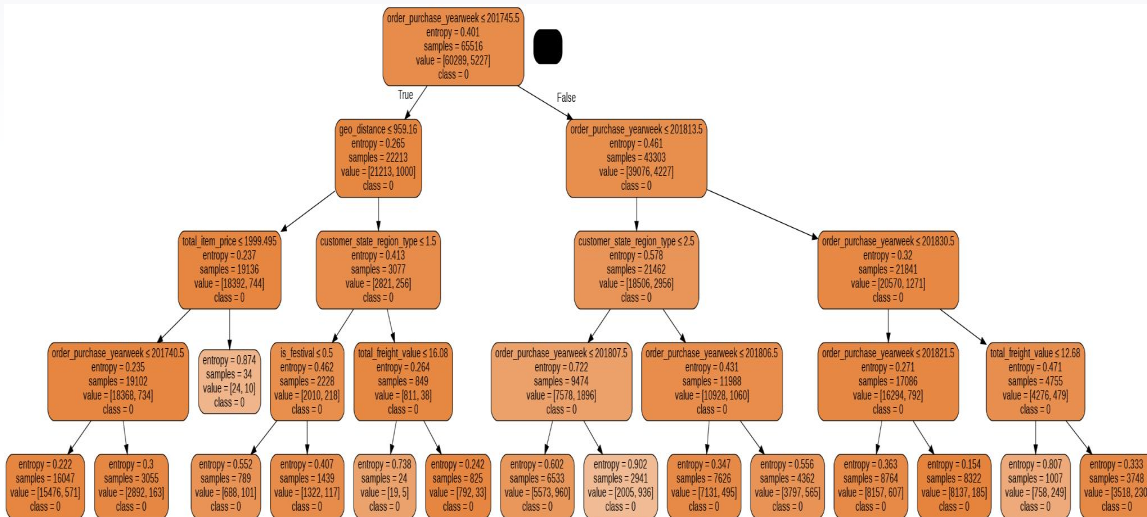
	RF 測試集 預測準確率	CAT 測試集 預測準確率
非延遲訂單	0.94	0.93
延遲訂單	0.33	0.32

特徵工程



特徵選擇

降低維度：刪減變項，參考決策樹關鍵因子



	RF 測試集 預測準確率	CAT 測試集 預測準確率
非延遲訂單	0.93	0.92
延遲訂單	0.30	0.31

Model調校

特徵工程: 集群後的 dataset

Logistic_analytics_v4

嘗試新增變數集群

延遲訂單準確率: 30% 上下

刪去過多變數甚至會降低準確率

5

4

特徵工程: 節慶與雨季後的 dataset

Logistic_analytics_v3

新增/刪減變數

(是否為雨季、是否為節慶月)

延遲訂單準確率: 30% 上下

6

特徵工程: 降維度後的 dataset

Logistic_analytics_v5

使用 GridSearchCV 來找出模型

超參數的最佳組合

延遲訂單準確率: 30% +

提高模型運算速度

模型適配結果

	X:15 原始model	X:13 資料預處理	X:8 降低維度	X:6 降低維度
模型	模型測試準確率	模型測試準確率	模型測試準確率	模型測試準確率
Random forests	91%	88.84%	87.61%	87.44%
CatBOOST	91%	87.75%	86.53%	86.03%

結論與建議



使用Random forests

訂單預測模型準確度高達 **87.4%**

並將**延遲訂單**準確度自**0.09**提升至**0.3**



因Olist Kaggle數據集限制模型變項建立、集群分析等，故考量合理性與可行性來調整模型後，準確率提升幅度較小。



服務補救

提早傳簡訊告知可能延遲，並準備小禮物給消費者，降低消費這負面觀感



事先考量商品備貨，提早安排倉儲、物流量，準時將商品送到消費者手中



拉長行銷戰線，提供每日不同的商品折扣活動，讓消費者不需等到活動當日下單，也同樣享有折扣，達到物流疏散作用

評價組

預測商品為「好評」與「負評」

黃怡家、簡智弘



研究動機

- ★ 根據《Forbes》報導, 94% 的消費者會避開有負面評價的公司
- ★ 對於收到 1-1.5 星評價的企業, 研究顯示他們的營收會比一般企業少33%
- ★ 流失客戶的百分比會隨著負評數的增加而上升:
 - ⇒ 1 個負面評論會帶走 22% 的潛在客戶
 - ⇒ 3 個負面評論會流失 59.2% 的潛在客戶
 - ⇒ 超過 4 個負面評論會流失 70% 的潛在客戶



研究目的

- ▶ 預測商品的評價結果為「好評」或「負評」
- ▶ 做為電商平台篩選新進賣家所販售商品的依據
- ▶ 維護電商平台的聲譽，進而提高買家的購買意願



新 4.2 折

秋 日 秋 新 件 OVERSIZE 大 元 學 工 女 大

\$180 - \$227 2 件 9.5 折

\$428 - \$540

★★★★☆ 3.4 | 已售出 12

2 件 9.5 折 查看全部 >

Miss Li

Miss Li

Miss Li

較長備貨 (出貨天數 8 天)

聊聊 加入購物車 直接購買

全部 (5) 附上評論 (2) 附上照片/影片 (2)

★★★★★ (3) ★★★★★ (0) ★★★ (0) ★★ (0) ★ (2)

s****n

☆☆☆☆☆

規格: 黑色, 2XL

圖文不符，字根本不一樣，還掉漆。照片上的字是正常的但是來的根本不一樣，賣家跟我說寄的是新版本。那你們照片不更新的嗎，字都不一樣。== 看得懂英文的人根本不會穿出去。買了快布回家。根本不會穿== 品質也很薄，就真的是一層很薄的布==

糟糕的商品品質 糟糕的CP值

SD

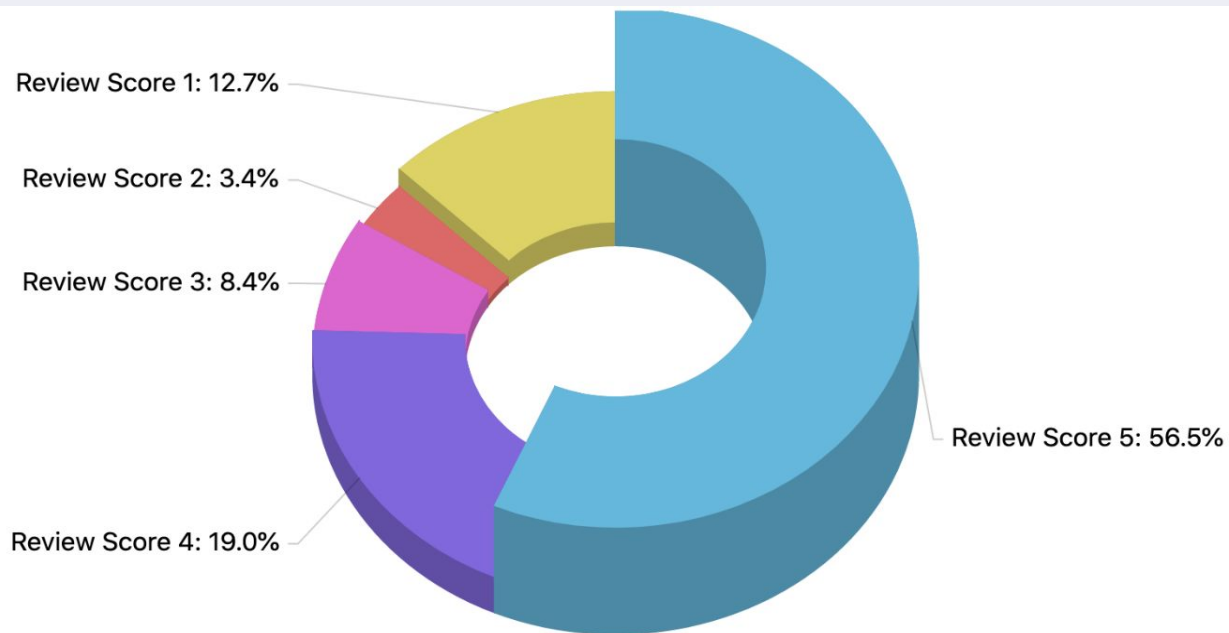
SDEAR YOUTH

2021-12-22 18:02

分析流程



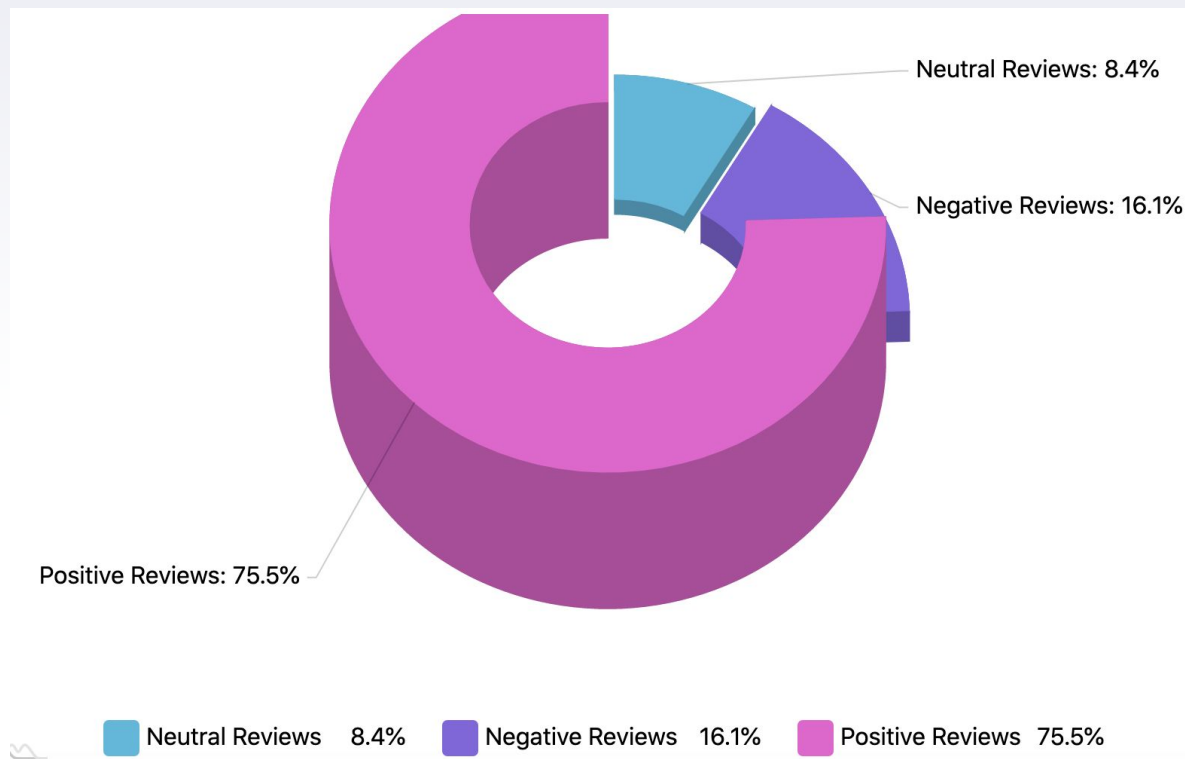
評價分數分佈圖



Review Score 5 56.5% Review Score 4 19.0% Review Score 3 8.4% Review Score 2 3.4% Review Score 1 12.7%

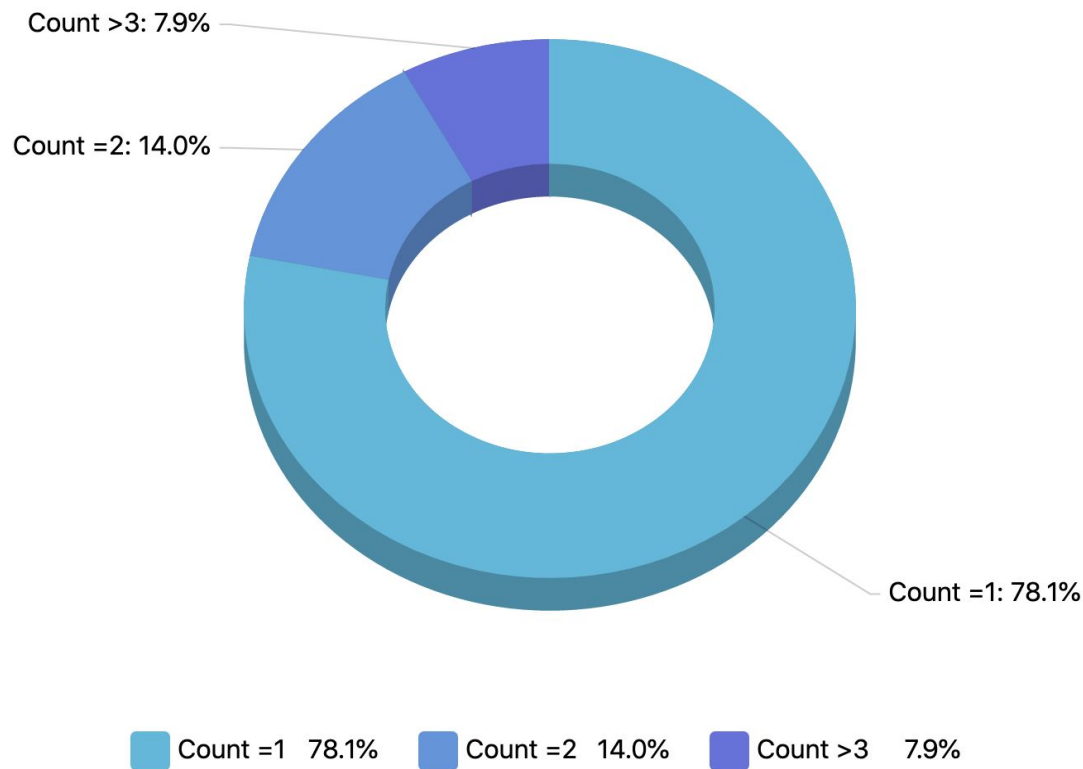
評價分數:1~5顆星星

「好評」和「負評」分佈圖



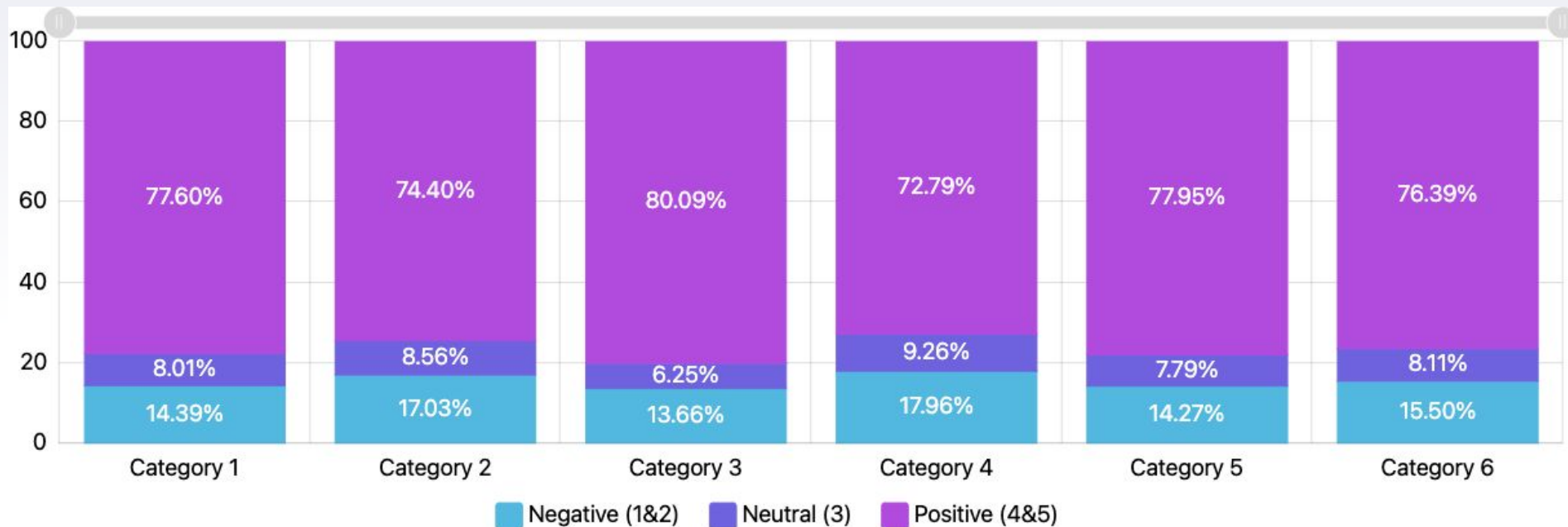
「好評」:4~5顆星星 ;「負評」:1~2顆星星

買家購買的商品數量分佈圖



Item_Count: 買家一筆訂單的商品購買數量

買家購買的商品種類&評價好壞之比例



Category 1 ⇒ Fashion & Accessory

Category 2 ⇒ Electronic Device & Home Appliance

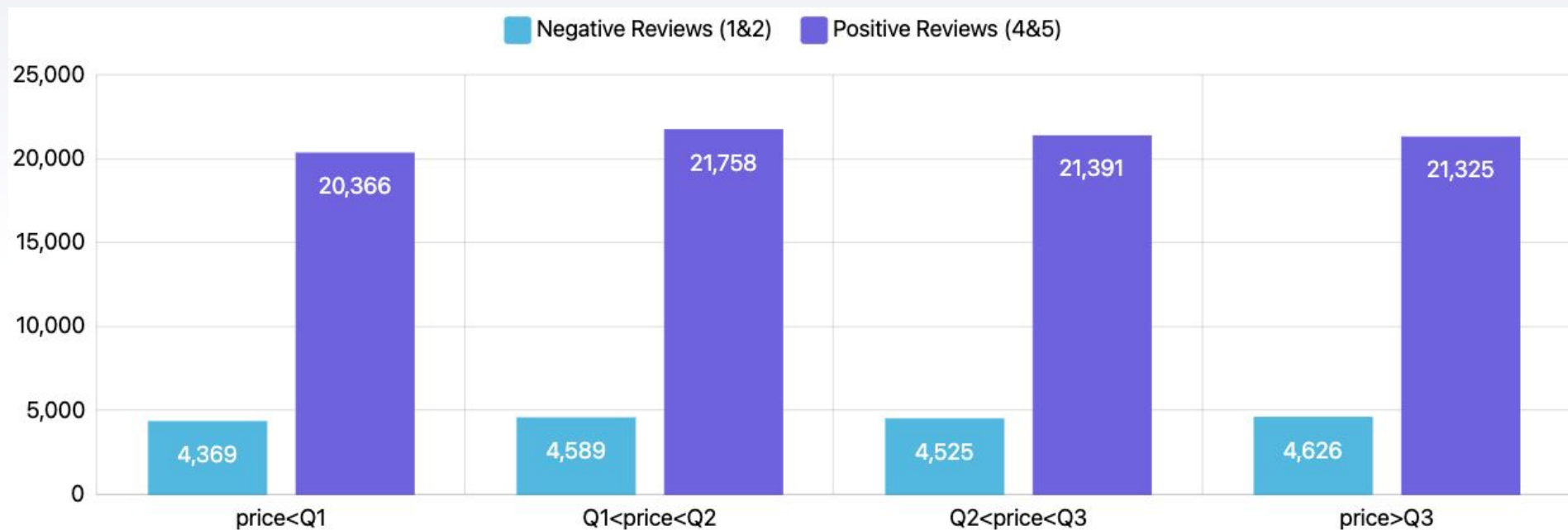
Category 3 ⇒ Art & Music & Book

Category 4 ⇒ Office Furniture & Home Decoration

Category 5 ⇒ Baby Goods & Food & Health Supplement

Category 6 ⇒ Tools & Others

買家購買商品的價格&評價好壞之分佈



由於商品價格懸殊過大 (0.85 ~ 6735), 故採用四分位數來劃分出 4 大區間
Q1=39.9; Q2=74.9; Q3=134.9

變數選擇

product_volume
product_weight
product_height
product_length
product_width

seller_state
seller_state_region_type
geo_distance



product_category_name_english
self_defined_product_category

item_price

approved_waiting_hrs (order_approved_at - order_purchase_timestamp)
seller_to_logistics_hrs (order_delivered_carrier - approved_at)
total_shipping_hrs (order_delivered_customer_date - order_approved_at)

從原始資料取出與評價相關的欄位

- 共 22 欄位
- 共 79,852 筆資料

資料處理過程

- 對數轉換: 使成為類常態分佈
- 資料編碼 (0.2~0.8): 使所有欄位權重一致

- 嘗試多種抽樣方法及比例
 - (1) Oversampling
 - (2) Bootstrapping
- 抽樣完畢重新打亂資料排序



- 刪除空值
- 刪除不合理負數

- 離群值處理
- 類別轉數值: Label Coding
- 數值轉類別: One-Hot Coding

- 將名目型態欄位重新歸納
- 刪除無效資料欄位

模型適配

T/F 7:3 (44010/18861筆)

	X 數量:12 原始model 無抽樣	X 數量:12 分6大類&4大類 Bootstrapping	X 數量:12 分6大類&4大類 Oversampling	X 數量:13 新增評價相關 集群欄位
模型	測試準確率	測試準確率	測試準確率	測試準確率
DNN	77.3%	84.8%	88.7%	88.3%
Random Forests	82.0%	83.3.0%	89.0%	89.0%
XGBOOST	79.8%	83.1%	89.1%	88.5%
LightGBM	81.3%	82.5%	80.9%	89.1%

T/F 7:1 (65798/9255筆)

T/F 1:1 (各18576筆)

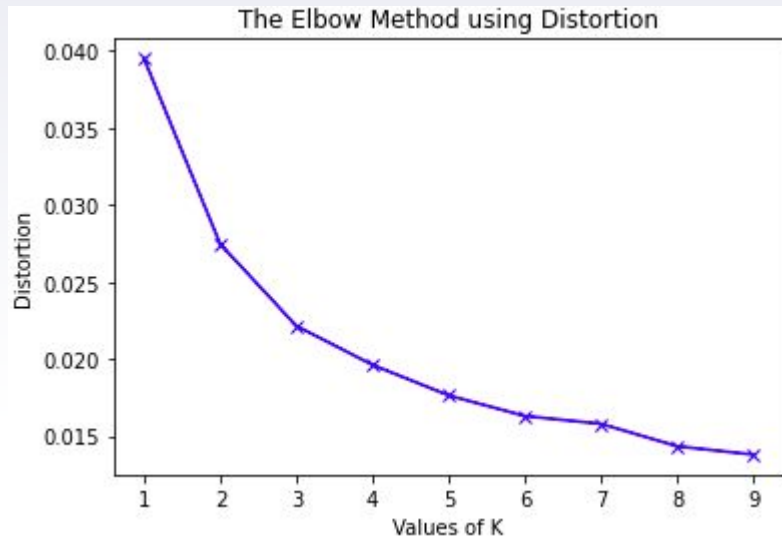
特徵工程

用 Elbow Method 做集群, 新增 速度 (y2) 欄位

- 距離 geo_distance
- 運送時間 total_shipping_hrs

=> 準度提升 8% (從 0.81 變成 0.89)

試降維度, 但準確率不增反減, 故保留



刪除欄位	ln_geo_distance	ln_total_shipping_hrs	ln_geo_distance & ln_total_shipping_hrs
準確度	0.88	0.87	0.86

模型預測最終結果

LightGBM	precision	recall	f1-score	support
0: 負評	0.57 (↑ 45%)	0.39	0.46	2980
1: 好評	0.92 (↑ 4.3%)	0.96	0.94	21788
accuracy			0.89	24768

原始資料集的好 / 負評比率 0.877 : 0.123

- 隨機猜中 0 (負評) 的可能性為 12%
- 隨機猜中 1 (好評) 的可能性為 88%

經 LightGBM 模型調教:

- 預測 0 (負評) 的準確率上升 45%
- 預測 1 (好評) 的準確率上升 4.3%

降維度-特徵篩選

比較「評價分數」(self_defined_review_score) 與其他 X 欄位的相關係數

- 將相關係數取絕對值 < 0.1 的欄位刪除
- 保留 ln_geo_distance (準度僅微幅下降)
- 將原 13 個欄位減少至 5 個欄位
 1. seller_state_region_type
 2. ln_seller_to_logistics_hrs
 3. ln_total_shipping_hrs
 4. ln_geo_distance
 5. y2 (集群欄位)

欄位名稱	相關係數
self_defined_product_category	0.008191
ln_item_price	-0.022346
ln_product_length_cm	-0.012646
ln_product_height_cm	-0.006426
ln_product_width_cm	-0.00866
ln_product_weight_cm	-0.017473
ln_product_volume	-0.012412
ln_approved_waiting_hrs	-0.011089
ln_geo_distance	-0.03091

模型預測最終結果

X 數量:13 → X 數量:5

LightGBM	precision	recall	f1-score	support
0: 負評	0.57 (↑ 45%)	0.39	0.46	2980
1: 好評	0.92 (↑ 4.3%)	0.96	0.94	21788
accuracy			0.89	24768

原始資料集的好 / 負評比率 0.877 : 0.123

- 隨機猜中 0 (負評) 的可能性為 12%
- 隨機猜中 1 (好評) 的可能性為 88%

經 LightGBM 模型調教:

- 預測 0 (負評) 的準確率上升 45%
- 預測 1 (好評) 的準確率上升 4.3%

結論與建議

電商平台內部人員

商務開發部門在招商時

⇒ 制定評價指標

(e.g. 負評比例需維持在幾個百分比以下)

賣家未能通過評價指標之門檻

⇒ 請營運部門針對問題提出改善建議

賣家



制定嚴格的訂單批准與出貨天數，降低買家給負評的風險



請賣家重新定位預販售之商品 (依售價、規格、內容調整)，或改善接單速率與出貨流程



請賣家提供預販售商品的過往銷售額 (參考資料) / 其他販售渠道的商品聲量

總結：
將成果設計成完整的營運系統

銷售決策

電商平台
營運管理系統

物流延遲預測

評價好壞預測

THANKS FOR YOUR ATTENTION

