

數位發展部數位產業署

AI 產業實戰應用人才淬煉計畫

111年度解題實證成果報告書

解題團隊： 不知名

題目名稱： 來店用餐的現場等候時間預測-以2018年之 POS 資料，輔以 AI 數據分析技術預測2018年預測來店用餐的現場等候時間

出題單位： 鼎泰豐小吃店股份有限公司

題目類別：
☐ 電腦視覺
☐ 自然語言
☒ 數據分析

中 華 民 國 一 百 一 十 一 年 十 月 三 日

※申請團隊保證申請文件所列資料及附件均屬正確※

※若有偽造不實者或侵權行為，申請團隊須負完全之法律責任※

目 錄

表目錄	III
圖目錄	IV
AIGO 人才解題成果報告摘要表	1
解題執行內容與成果說明	3
一、 計畫背景與目的	3
(一) 題目背景	3
(二) 構想說明	4
二、 分析架構與方法	5
(一) 流程與架構	5
(二) 資料源說明	7
(三) 預測與評估方法	9
(四) 預測環境	12
三、 實作與分析	14
(一) 資料檢視	14
(二) 衍生變數	17
(三) 探索式資料分析	20
(四) 資料預處理	26
(五) 預測實作與結果	26
(六) 變數重要性	30

四、	成果效益與完成之工作	35
五、	商業應用價值與創新亮點.....	38
六、	結論與建議	39
(一)	結論	39
(二)	題目限制	39
七、	交付項目	44

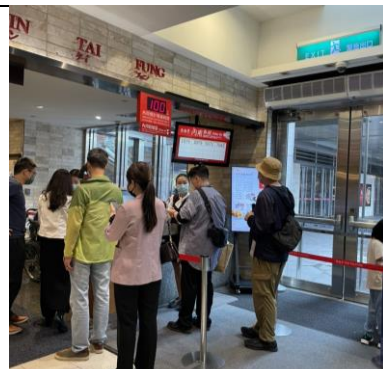
表目錄

表1.1	痛點說明	3
表1.2	解題構想摘要.....	4
表2.1	外部開放數據介紹	8
表2.2	模型概述	9
表2.3	演算法說明	10
表2.4	準確度範例說明	12
表2.5	硬體環境.....	12
表2.6	軟體環境.....	12
表3.1	預想可能之衍生變數範例清單	18
表3.2	數據檢視洞察.....	20
表3.3	各演算法預測效度	30
表3.4	變數重要性定義	31
表4.1	模型結果	35
表4.2	工作項目與時程規畫表.....	36

圖目錄

圖2.1	分析流程	5
圖2.2	分析架構	6
圖2.3	資料來源	7
圖2.4	SHAP 示意圖	11
圖3.1	衍生變數開發概念示意圖	19
圖3.2	建模資料切分	27
圖3.3	TABNET 模型初版效度	28
圖3.4	異常資料散佈圖	29
圖3.5	整體模型架構	29
圖3.6	變數重要性(SHAP VALUE)	31
圖3.7	SHAP 指標	32
圖3.8	「現場排隊人數(未點餐)」與排隊時間貢獻度散佈圖	33
圖3.9	「現場排隊組數(未點餐)」與排隊時間貢獻度散佈圖	34
圖5.1	商轉可行性規劃	38

AIGO 人才解題成果報告摘要表

團隊名稱		不知名			
題目名稱		來店用餐的現場等候時間預測-以2018年之 POS 資料，輔以 AI 數據分析技術預測2018年預測來店用餐的現場等候時間			
出題單位		鼎泰豐小吃店股份有限公司			
題目類別		<input type="checkbox"/> 電腦視覺 <input type="checkbox"/> 自然語言 <input checked="" type="checkbox"/> 數據分析			
解題期程		自111年6月15日至10月3日止			
聯絡人		姓名	王政雲	電話	0912-015822
		Email	cloudy8222@gmail.com		
團隊任務配置：AI 演算法、資料處理、前後端、PM, etc (參考但不限於)					
No.	團隊角色	姓名/ 公司名稱	經歷專長/ 公司簡介	分工說明	
1	隊長	王政雲 國泰世華銀行	✓ 自然語言處理 ✓ 機器/深度學習 ✓ 統計分析	聯繫窗口、分析技術研究、模型建置	
2	組員	葉靜縈 新光人壽	✓ 機器學習技術 ✓ 統計分析	聯繫窗口、變數開發、模型建置	
3	組員	吳重億 研華科技	✓ 影像辨識技術 ✓ 網頁設計	變數開發、分析技術研究、模型建置	
4	組員	盧勁瑋 元大期貨	✓ ETL 數據工程 ✓ 視覺化報表設計 ✓ 數據分析技術	變數開發、分析環境建置、資料彙整	
5	組員	林宇桐 安侯建業	✓ 機器學習技術 ✓ 數位轉型經驗	EDA、分析技術研究	
6	組員	鍾馨瑩 意藍資訊	✓ 文字探勘技術 ✓ 網路輿情分析	EDA、簡報製作	
7	組員	黃婷 全家便利商店	✓ 數據探勘 ✓ 統計分析	EDA、成果報告	
計畫執行過程照片與說明 (至少四張)		<div></div> <div></div>			

	<p>說明：進度線上會議，討論會議資料。</p>		<p>說明：實地場勘，假日中午用餐人數眾多，需等候約 100 分鐘。</p>																					
																								
	<p>說明：模型結果討論與洞察，以利後續模型優化。</p>		<p>說明：模型優化、實證成果討論</p>																					
解題內容摘要	<p>民以食為天，餐廳聚集各式美食、廚師、消費者與服務人員，使其成為最誘人、最龐大、也最生氣勃勃的產業，消費者從選擇餐廳開始即展開一連串旅程，從入座、選擇餐點、用餐到結帳，每個環節累積了大量數據，現場等候時間、餐點製作時間、用餐時間、服務人員狀況、消費者滿意度等。</p> <p>其中<u>現場等候時間</u>將是各餐飲業門市首要面臨的挑戰，當門市呈現營運忙碌、滿桌的情況下，需要依店內的桌位使用狀況，進行下一桌可使用的等候時間的預估，而過往常常需仰賴資深服務人員對門市長期服務經驗進行判斷，若能藉由 AI 數據分析技術輔助下，預估出當下需等候的時間，則能適當地提醒現場排隊的消費者，也能通知相關服務人員進行服務準備和安排，不僅能有效地為各門市提升消費者滿意度、降低消費者流失與客怨次數，並為門市與公司達到營運目標。</p>																							
解題成果簡述	量化成果	<p>本解題團隊模型準確率已達到 AIGO 競賽標準：</p> <table><tr><th>項目</th><th>說明</th><th>準確度目標</th><th>結果</th></tr><tr><td>競賽標準</td><td>±誤差15分鐘內</td><td>70%以上</td><td>達成</td></tr><tr><td>進階期望</td><td>±誤差5分鐘內</td><td>80%以上</td><td>未達成</td></tr></table> <p>檢視預測誤差±5分鐘、±10分鐘、±15分鐘時之效度，準確度詳細如下：</p> <table><tr><th>項目</th><th>±誤差5分鐘內</th><th>±誤差10分鐘內</th><th>±誤差15分鐘內</th></tr><tr><td>準確度</td><td>33%</td><td>58%</td><td>73%</td></tr></table>			項目	說明	準確度目標	結果	競賽標準	±誤差15分鐘內	70%以上	達成	進階期望	±誤差5分鐘內	80%以上	未達成	項目	±誤差5分鐘內	±誤差10分鐘內	±誤差15分鐘內	準確度	33%	58%	73%
項目	說明	準確度目標	結果																					
競賽標準	±誤差15分鐘內	70%以上	達成																					
進階期望	±誤差5分鐘內	80%以上	未達成																					
項目	±誤差5分鐘內	±誤差10分鐘內	±誤差15分鐘內																					
準確度	33%	58%	73%																					
	質化成果	<p>提供顧客準確的等候時間，除可增加顧客體驗，讓顧客掌握需等候多久做好自己的時程安排；並可減少員工勞務量，如顧客客訴等。</p>																						
成果電子檔	https://github.com/wty81213/AIGO/tree/master																							
備註	<p>交付項目詳細資訊提供於上述 github 連結中</p>																							

解題執行內容與成果說明

一、計畫背景與目的

(一) 題目背景

民以食為天，餐廳聚集各式美食、廚師、消費者與服務人員，使其成為最誘人、最龐大、也最生氣勃勃的產業，消費者從選擇餐廳開始即展開一連串旅程，從入座、選擇餐點、用餐到結帳，每個環節累積了大量數據，現場等候時間、餐點製作時間、用餐時間、服務人員狀況、消費者滿意度等。

其中現場等候時間將是各餐飲業門市首要面臨的挑戰，當門市呈現營運忙碌、滿桌的情況下，需要依店內的桌位使用狀況，進行下一桌可使用的等候時間的預估，而過往常常需仰賴資深服務人員對門市長期服務經驗進行判斷，因此就企業與客戶而言，以往常存在以下痛點，如表1-1所示。

表1.1 痛點說明

角度	痛點
企業	<ul style="list-style-type: none">資深人員常養成不易，需要經過長期且專業培訓較難有系統性的評估準則，不同的專業人員可能存在評估結果的差異
客戶	<ul style="list-style-type: none">預估等候時間的精準度，及精準度的變異程度，影響客戶對數值的信賴度，及客戶評估能否作為等待期間行程規劃的依據

根據上表，若能藉由 AI 數據分析技術輔助下，預估出當下需等候的時間，則能適當地提醒現場排隊的消費者，也能通知相關服務人員進行服務準備和安排，不僅能有效地為各門市提升消費者滿意度、降低消費者流失與客怨次數，並為門市與公司達到營運目標。

(二) 構想說明

針對前述背景與痛點，本解題團隊預計以下方表1-2的摘要解題構想，進行議題排解與現況優化。

表1.2 解題構想摘要

No.	階段	作業項目概要
1	資料蒐集	<ul style="list-style-type: none">除企業內部數據外，根據等候時間預測任務，亦規劃結合豐富的多項外部資料，以利提高模型效度。
2	資料處理	<ul style="list-style-type: none">根據資料內容大量發想且延伸可能相關特徵，並進一步檢視，如時間序相關因子等，藉以提高模型精準度。
3	模型建立	<ul style="list-style-type: none">根據目前熱門的機器學習演算法，建立多個候選模型，有效學習顯著的特徵因子嘗試拆分子模型進行預測以提升模型效度
4	模型解析	<ul style="list-style-type: none">透過模型解析，找到重要的特徵因子，輔助服務人員可以進行相關決策
5	分析報告	<ul style="list-style-type: none">目標模型預測準確率達7成以上，以利後續落地應用產出及交付相關資訊，後續可再依需求討論調整

本解題團隊預期，預測模型對門市的幫助與建置目的如下，此外，亦可將此模式作為整體產業優化此項相關議題的參考：

- 提升消費者等候時間準確率：目標準確率七成以上、誤差在提案單位期望範圍內為內且盡可能降低，以增加消費者滿意度並降低消費者因此流失的可能。
- 找出影響等候時間的關鍵因子，作為各門市經營團隊之參考依據。

就過往經驗而言，資料維度的豐富度與乾淨度對於預測準確度至關重要，因此本解題團隊除嘗試多元且業界或競賽中常用的演算法與預處理方式外，亦同步盡可能豐富資料源與衍生變數，初步盤點可引入外部數據達上述九項，作為優化模型效度的依據。

二、 分析架構與方法

(一) 流程與架構

關於預測客戶的來店用餐時間的任務，本次提案的分析流程可分為五大部分，如圖2-1所示，分別為資料收集、資料處理、模型建置、模型解析與分析報告，以利解決出題方的痛點，後續逐項進行說明。



圖2.1 分析流程

分析架構如圖2.2所示。

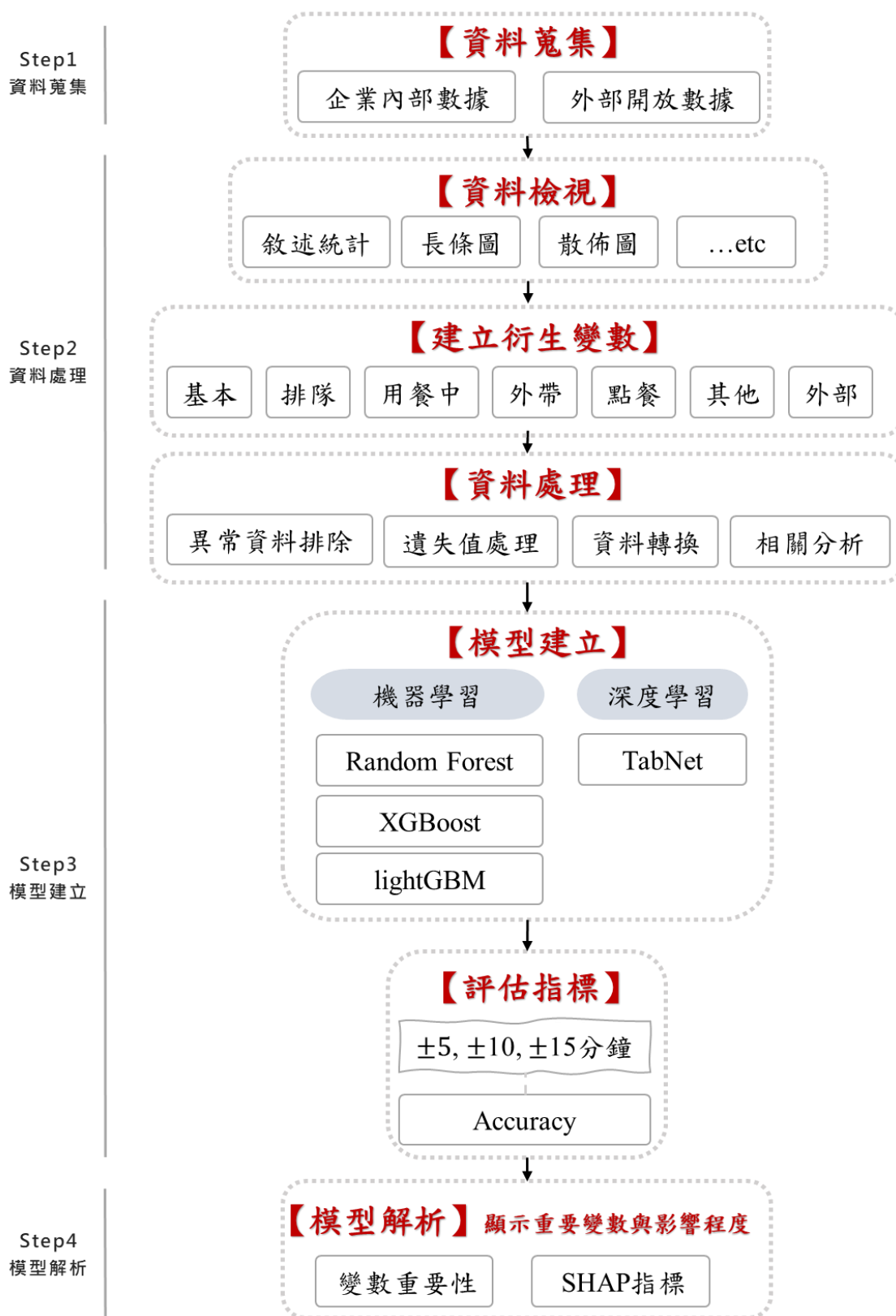


圖2.2 分析架構

(二) 資料源說明

本解題團隊將彙整各方數據，包含企業內部蒐集資訊，與外部開放資料（Open Data）進行整合，從更廣泛且豐富的面向，分析影響客戶等候時間的原因，以更精準化預測效度，主要資料來源分為門店資料、節日與人流資料、市場與氣候資料三大類別的資訊，相關如圖2.3所示。

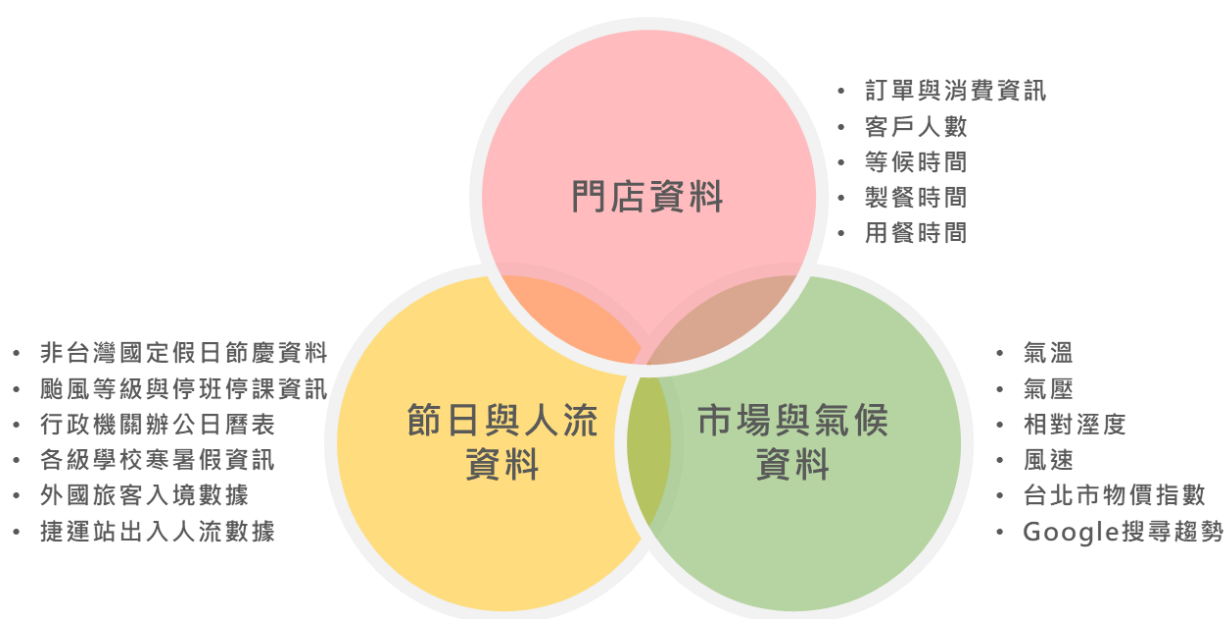


圖2.3 資料來源

● 企業內部數據

企業內部數據包含內用、外帶、排隊入場及點餐資料，期間為2016~2018年每個營業日各時段紀錄。

- **內用資料(OrderInside)**：記錄消費者內用相關資訊，如訂單編號、點餐時間、離席時間等共22個欄位，資料期間為2016~2018年。
- **外帶資料(OrderOutside)**：記錄消費者外帶相關資訊，如訂單編號、訂單類型等共7個欄位，資料期間為2016~2018年。
- **排隊資料(Queue、Linein)**：記錄消費者排隊相關資訊，如訂單編

號、排隊模式、排隊編號等共9個欄位，資料期間為2016~2018年。

- **點餐資料(Order_Achievement)**：記錄消費者點菜相關資訊，如訂單編號、點菜流水號、產品編號、點餐數量等共16個欄位，資料期間為2016~2018年。

● 外部開放數據

本解題團隊蒐集多種外部開放數據包含人流、氣象、節假日、網路搜尋量、消費者物價指數資料等如表2.1所示。

由於資料量大且繁複，故利用爬蟲相關技術(也稱網路爬蟲，web crawler)，自動化爬取相關數據，達到減省時間與重複性作業。

表2.1 外部開放數據介紹

類別	資料名稱	來源
人流相關	外國旅客入境數據	https://stat.taiwan.net.tw/inboundSearch
	捷運站出入人流數據	https://data.gov.tw/dataset/128506
氣象相關	颱風停班停課資訊	https://dop.gov.taipei/cp.aspx?n=72E3AB3DA4700EF6
	氣溫	1. https://stat.motc.gov.tw/mocdb/stmain.jsp?sys=100&funid=b8101
	降雨量	2. https://stat.motc.gov.tw/mocdb/stmain.jsp?sys=100&funid=a8101
		3. https://e-service.cwb.gov.tw/HistoryDataQuery/index.jsp 4. https://opendata.cwb.gov.tw/dataset/climate/C-B0026-002 5. https://opendata.cwb.gov.tw/index
	颱風警報註記	https://rdc28.cwb.gov.tw/TDB/public/warning_typhoon_list/
節假日相關	非台灣國定假日節慶資料	中國/日本/美國行事曆
	行政機關辦公日曆表	https://www.dgpa.gov.tw/information?uid=30&pid=9811
	各級學校寒暑假資訊	https://data.gov.tw/dataset/6231 https://www.nmes.tp.edu.tw/SchCalendar/105/105-holidays-all.pdf https://www.nmes.tp.edu.tw/SchCalendar/106/106-holidays-all.pdf https://www.nmes.tp.edu.tw/SchCalendar/107/107-holidays-all.pdf 大專院校之行事曆以台灣大學為主

類別	資料名稱	來源
網路搜尋量 相關	鼎泰豐 GoogleTrends	https://trends.google.com.tw/trends/yis/2021/TW/
消費者物價 指數資料	台北市消費者物價指數(月)	https://data.gov.tw/dataset/131822 https://data.gov.tw/dataset/145782

使用外部開放資料將會遇到三個主要問題，(1) 爬蟲過程中若資料提供方網站改版、鎖住 IP，則需修改爬蟲程式；(2) 外部資料無法確保提供穩定性，如本次蒐集過程中，盡可能收集2016~2018年完整資料，但如捷運人流，資料官方無提供2016年(含)以前資料；(3) 資料提供因顆粒度不同，如台北市消費者物價指數是以月方式提供等，相對日資料提供資訊較少，仍需進行調整。

(三) 預測與評估方法

● 演算法

在整體過程中，不知名團隊嘗試採用坊間比賽或一般企業內部常使用的演算法，共計5個，詳細如下說明。



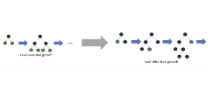
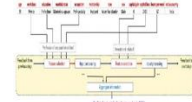
本提案為進行等候時間的數值型預測，將依前述彙整之內外部資料與衍生變數，規劃採用競賽與業界常用之演算法，其中採用集成學習（Ensemble Learning）之方式，由多個不同維度的弱模型同時預測目標，進而提升預測精準度，達到最小化預測結果與實際結果的誤差之目標，如表2-2所示，而採用之各演算法特性與說明如表2-3所示。

表2.2 模型概述

項目	說明
問題種類	迴歸問題

預測結果	連續型/數值
利用資料方式	使用數學函數組合不同的特徵(features)，預測出一個連續函數作為結果的輸出
模型演算法	Random Forest, XGBoost, LightGBM, TabNet
訓練目標	最小化預測結果與實際結果的誤差

表2.3 演算法說明

模型	Random Forest	XGBoost	LightGBM	TabNet
特性	Bagging：產生多棵樹，由投票對結果產生影響。	Boosting：新的樹針對舊的樹進行補強與優化。	優化 XGBoost 缺點，有更快的訓練效率、準確度，並支持平行運算，處理大規模數據。	深度學習模擬基於樹的決策過程，並將資訊由上層傳下層。
示意圖				

除前述提及之模型架構外，於資料檢視與建模過程亦嘗試拆分子模型，進行各別預測之可能，以確認有無藉此提升模型預測效度之機會，例如拆分是否為週末子模型、依照現況等候組數高低拆分子模型等。

● 模型解析

隨著應用場景多元，不單單只是追求預測的精準度，會想要解析模型內部的演算過程，了解那些特徵對於預測結果來說是有顯著性的影響，讓我們的使用者有更多的資訊進行操作，提供更多元的服務，以本次的分析題目為例，若知道哪些因子影響顧客的等候時間，就可以進行優化，讓顧客的等候時間可以縮短，提供更良好的客戶服務體驗。

另外隨著數據分析技術發展，漸漸機器學習模型不再是黑盒子，而可

以進一步探索與解析，獲得關聯性特徵並做重要性排序，以利後續使用者有足夠資訊，進行後續相關應用，本提案提出最常用模型解析的技術為「Shapley Additive exPlanations(SHAP)」。

SHAP 是基於賽局理論 (cooperative game theory) 提出來解決方案，其中 SHAP 值 (SHapley Additive exPlanations) 分解預測以顯示每個特徵的影響。遊戲理論中使用的一種技術，用於確定協作遊戲中每個玩家對其成功的貢獻程度。換句話說，每個 SHAP 值測量我們模型中每個特徵對每個預測的正面或負面貢獻的程度，如圖2-4所示。



圖2.4 SHAP 示意圖

此相關技術應用可運用本提案所提供模型方法，可以進一步了解特徵重要性，且針對每一筆預測，也能進一步解析且提供哪些因子會影響我們預測結果，以利有詳細的模型資訊，提供使用者進一步的相關運用。

● 評估指標

模型評估利用三種時間指標衡量模型準確度，準確度(Accuracy)公式如下，範例說明如表2.4所示。

$$\frac{(|\text{實際等候時間} - \text{預測等候時間}| < \text{【時間指標】}) \text{ 筆數}}{\text{總資料筆數}} \times 100\%$$

表2.4 準確度範例說明

#	預測等候時間	實際等候時間	預測落差時間 (實際等候時間-預測等候時間)	時間指標(分鐘)		
				15	10	5
case1	13	15	2	✓	✓	✓
case2	17	29	12	✓	✗	✗
case3	20	23	3	✓	✓	✓
case4	30	8	22	✗	✗	✗
case5	14	20	6	✓	✓	✗
準確度				80%	60%	40%

(四) 預測環境

- 本次解題之硬體設備如表2.5所示。

表2.5 硬體環境

硬體環境	
中央處理器	2.81GHz Intel Core i7
記憶體	16GB
硬碟	1TB

- 軟體環境

本次解題之軟體設備如表2.6軟體環境所示。

表2.6 軟體環境

軟體環境	
作業系統	Windows 10 專業版
程式開發套件	os、numpy、pandas、pyarrow、tqdm、

	sklearn、lightgbm、pytorch_tabnet、 matplotlib、plotly
程式語言	Python 3、R、T-SQL

三、 實作與分析

在預測模型建立之前，需先了解資料並確保資料品質，進而衍生相關變數，本解題團隊經團隊共同發想與討論，內外部資料利用特徵工程共衍生出459個變數，進而進行探索性分析，掌握變數間的關聯性，最後資料準備妥當後，即可開始訓練模型程序。

(一) 資料檢視

- 企業內部數據

在此僅列出重要調整項目進行說明。

資料表	問題	說明	與鼎泰豐討論後對應作法
其他	提供的資料是哪一間分店	各店鋪所在地有自己的商圈、顧客屬定，有自己的商圈、顧客特質等因素。	鼎泰豐101店 台北市市府路45號 B1
內用 (OrderInside)	Nation_Code、 Adult_Count、 Kid_Count、Team_Flag 與 Unhandy_Flag 變數含 NA 值	用餐顧客國別、大人、小孩人數等不應有空值。	此為交易時系統異常，建議刪掉。
	用餐時間數值範圍介於0	時間太長不合乎常理，經討論後排除用餐	排除626訂單編號

資料表	問題	說明	與鼎泰豐討論後對應作法																																																		
	分~900分鐘	時間大於180分鐘。 <table><tr><th>用餐狀況</th><th>客戶數</th><th>客戶數_累加</th><th>客戶佔比</th><th>客戶累加佔比</th></tr><tr><td>用餐15分內</td><td>6256</td><td>6256</td><td>0.009149369</td><td>0.009149369</td></tr><tr><td>用餐15分至30內</td><td>124832</td><td>131088</td><td>0.182566182</td><td>0.191715551</td></tr><tr><td>用餐30分至45內</td><td>287249</td><td>418337</td><td>0.420100239</td><td>0.61181579</td></tr><tr><td>用餐45分至60內</td><td>170363</td><td>588700</td><td>0.249155043</td><td>0.860970833</td></tr><tr><td>用餐60分至1小時半內</td><td>81379</td><td>670079</td><td>0.119016384</td><td>0.979987218</td></tr><tr><td>用餐1小時半至2小時內</td><td>10400</td><td>680479</td><td>0.015209948</td><td>0.995197166</td></tr><tr><td>用餐2小時半至3小時內</td><td>2969</td><td>683448</td><td>0.004342148</td><td>0.999539314</td></tr><tr><td>用餐3小時半至4小時內</td><td>243</td><td>683691</td><td>0.000355386</td><td>0.9998947</td></tr><tr><td>用餐4小時半至15小時內</td><td>72</td><td>683763</td><td>0.0001053</td><td>1</td></tr></table>	用餐狀況	客戶數	客戶數_累加	客戶佔比	客戶累加佔比	用餐15分內	6256	6256	0.009149369	0.009149369	用餐15分至30內	124832	131088	0.182566182	0.191715551	用餐30分至45內	287249	418337	0.420100239	0.61181579	用餐45分至60內	170363	588700	0.249155043	0.860970833	用餐60分至1小時半內	81379	670079	0.119016384	0.979987218	用餐1小時半至2小時內	10400	680479	0.015209948	0.995197166	用餐2小時半至3小時內	2969	683448	0.004342148	0.999539314	用餐3小時半至4小時內	243	683691	0.000355386	0.9998947	用餐4小時半至15小時內	72	683763	0.0001053	1	
用餐狀況	客戶數	客戶數_累加	客戶佔比	客戶累加佔比																																																	
用餐15分內	6256	6256	0.009149369	0.009149369																																																	
用餐15分至30內	124832	131088	0.182566182	0.191715551																																																	
用餐30分至45內	287249	418337	0.420100239	0.61181579																																																	
用餐45分至60內	170363	588700	0.249155043	0.860970833																																																	
用餐60分至1小時半內	81379	670079	0.119016384	0.979987218																																																	
用餐1小時半至2小時內	10400	680479	0.015209948	0.995197166																																																	
用餐2小時半至3小時內	2969	683448	0.004342148	0.999539314																																																	
用餐3小時半至4小時內	243	683691	0.000355386	0.9998947																																																	
用餐4小時半至15小時內	72	683763	0.0001053	1																																																	
排隊 (Queue, Linein)	訂單編號在 Queue 但不在 OrderInside，共1490筆。	顧客有排隊但沒有內用，非此分析範圍。	排除1490筆資料																																																		
	訂單編號出現在 queue 與 orderinside 但沒有排隊時間	顧客有排隊但沒有內用，非分析範圍。	排除3筆訂單編號																																																		
	預約排隊原時間序流水號 Reserve_No 保留前置字母為 M、N 的資料，排掉 A	A 為系統測試資料，非分析範圍。	排除2,420筆訂單編號																																																		
	Linein 排隊資料部分日期缺漏	因系統異常導致營業日未紀錄資料	後續開發衍生變數時，已前一週同星期之資料補值																																																		
	等候時間介於0分~600分鐘	一般正常願意等候時間不可能高達10小時，經討論後，排除等候時間大於180分鐘。	排除97筆訂單編號																																																		

資料表	問題	說明	與鼎泰豐討論後對應作法																																																		
		<table><tr><th>等待狀況</th><th>客戶數</th><th>客戶數_累加</th><th>客戶佔比</th><th>客戶累加佔比</th></tr><tr><td>等待15分內</td><td>1779</td><td>1779</td><td>0.034821586</td><td>0.034821586</td></tr><tr><td>等待15分至30內</td><td>12848</td><td>14627</td><td>0.251482707</td><td>0.286304293</td></tr><tr><td>等待30分至45內</td><td>13424</td><td>28051</td><td>0.262757149</td><td>0.549061442</td></tr><tr><td>等待45分至60內</td><td>9272</td><td>37323</td><td>0.181487209</td><td>0.73054865</td></tr><tr><td>等待60分至1小時半內</td><td>10031</td><td>47354</td><td>0.196343636</td><td>0.926892286</td></tr><tr><td>等待1小時半至2小時內</td><td>2880</td><td>50234</td><td>0.056372213</td><td>0.983264499</td></tr><tr><td>等待2小時半至3小時內</td><td>765</td><td>50999</td><td>0.014973869</td><td>0.998238368</td></tr><tr><td>等待3小時半至4小時內</td><td>52</td><td>51051</td><td>0.001017832</td><td>0.9992562</td></tr><tr><td>等待4小時半至10小時內</td><td>38</td><td>51089</td><td>0.0007438</td><td>1</td></tr></table>	等待狀況	客戶數	客戶數_累加	客戶佔比	客戶累加佔比	等待15分內	1779	1779	0.034821586	0.034821586	等待15分至30內	12848	14627	0.251482707	0.286304293	等待30分至45內	13424	28051	0.262757149	0.549061442	等待45分至60內	9272	37323	0.181487209	0.73054865	等待60分至1小時半內	10031	47354	0.196343636	0.926892286	等待1小時半至2小時內	2880	50234	0.056372213	0.983264499	等待2小時半至3小時內	765	50999	0.014973869	0.998238368	等待3小時半至4小時內	52	51051	0.001017832	0.9992562	等待4小時半至10小時內	38	51089	0.0007438	1	
	等待狀況	客戶數	客戶數_累加	客戶佔比	客戶累加佔比																																																
等待15分內	1779	1779	0.034821586	0.034821586																																																	
等待15分至30內	12848	14627	0.251482707	0.286304293																																																	
等待30分至45內	13424	28051	0.262757149	0.549061442																																																	
等待45分至60內	9272	37323	0.181487209	0.73054865																																																	
等待60分至1小時半內	10031	47354	0.196343636	0.926892286																																																	
等待1小時半至2小時內	2880	50234	0.056372213	0.983264499																																																	
等待2小時半至3小時內	765	50999	0.014973869	0.998238368																																																	
等待3小時半至4小時內	52	51051	0.001017832	0.9992562																																																	
等待4小時半至10小時內	38	51089	0.0007438	1																																																	
	未入店用餐者無入店時間，在計算排隊相關的衍生變數時，較難以判斷	在產生排隊相關衍生變數時，需統計該時間有哪些客戶仍在等餐，但資料未記錄排隊狀態之轉換時間，僅記錄最後是重複取號、內用轉外帶等狀態，較難以回推各時間點之排隊狀態，若入店時間仍以空值呈現，會造成排除該些資料，造成引用到未來結果的概念，上線時無法以同概念進行變數開發，而若入店時間補入當日最後營業時間，會造成愈晚排隊者，等候人數愈多，與真實情況亦不相符。	若無入店用餐之資訊，則入店時間補上開始排隊時間加上180分鐘，在該期間視為客戶等餐中，若超過則視為已無等餐需求																																																		
點餐 (Order_Achievement)	產品編號為41、42、43、44、53	該產品編號為員工餐或特殊品等，非分析範圍。	共排除25,262筆訂單編號																																																		

- 外部開放數據：因 Open data 提供前已經過清整或彙整，故無問題。

(二) 衍生變數

本解題團隊運用特徵工程技術進行各項衍生變數產出，方法如下：

1. Component encoding：利用資料的時間相關性，給予對應時間相關的特徵，例如：年、月、周、星期，或是時、分、秒。

2. Characteristics in time series：

- *Target/Feature Lag N period*：取先前時間點(Lag)的單一數值當作特徵。例如：前30天的資料點 (Lag30)。
- *Target/Feature Lag N periods Aggregate*：取先前一段時間點(Window)的數值，經過統計的計算 (平均值、最大值、標準差) 後當作特徵。例如：過往七天來店人數的平均趨勢。
- *Target/Feature Lag N periods Interaction*：不同時間點資料彼此的變化。例如：前兩天營業額的變化 (Lag2-Lag1)

3. Dummy variables：

加入與時間可以連結，或是無法用時間數值表示的外部資訊。

4. Relevant time series data：

其他相關性高的資料也可以加入特徵生成，或有互補性或替代性的消長。

5. Algorithm predictions：

利用其它演算法的預測結果當作特徵。

衍生共產出459個，因數量較多，故詳細請參考附件，表3.1列舉部分重要衍生變數。

表3.1 預想可能之衍生變數範例清單

類別	變數 總個數	衍生變數範例清單
基本 資訊	5 (1%)	<ul style="list-style-type: none"> • A06_queue_month(用餐月份) • A04_queue_weekday(用餐星期) • ...
排隊 資訊	152 (33%)	<ul style="list-style-type: none"> • B273_Sameday_Queue_NoIntime_3BeginningCount (當天至排隊當下，已排隊但尚無入店時間的排隊號碼3開頭的組數) • B271_Sameday_Queue_NoIntime_Count(當天至排隊當下，已排隊但尚無入店時間的組數) • ...
用餐 中	131 (28%)	<ul style="list-style-type: none"> • A11_flag_Nation_3_4 (本次用餐是否有大陸+港澳人亞洲人) • C013_ING_Adult_NUMBER (用餐中的客戶大人數) • ...
外帶 資訊	26 (6%)	<ul style="list-style-type: none"> • B254_Yesterday_Queue_Takeaway_Percent(前一天排隊資料，最後改外帶的比例) • B213_Outside15min_Serial_Number(近15分鐘外帶組數) • ...
點餐 資訊	22 (5%)	<ul style="list-style-type: none"> • C11_unique_order_sum(用餐中的組數已點餐不重複品項數總和) • C12_max_time(用餐中的組數入席到第一道上菜的最久時間(分鐘)) • ...
時間 資訊	47 (10%)	<ul style="list-style-type: none"> • B15_mean_of_waiting_time_for_the_same_cnt_and_month_1_aggfun_mean(近一個月相同等候組數的平均等候時間(分鐘)) • B16_mean_of_waiting_time_on_last_week_and_60_min_aggfun_max(上週同時段(60分鐘)開始等候者的最長等候時間(含不需排隊者；分鐘)) • ...
外部 資訊	76	<ul style="list-style-type: none"> • E1016_ForeignerResidence_Americas(前一個月外國旅客居住地為美洲的人次)

	(17%)	<ul style="list-style-type: none"> • E111_CPI_TTL(前一月份之台北市物價總指數) • E114_CPI_Food(前一月份之台北市物價總指數_食物類指數) • E072_Temperature(前一天同時段測站氣溫) • E073_RH(前一天同時段測站相對濕度) • E0908_GoogleTrends_grobalch7Day_sum(前一週全球以中文 Google 搜尋鼎泰豐的趨勢總和)
--	-------	---

以用餐中類別其中之用餐中已點餐品項數的衍生變數為例，如圖3.1所示，假設被預測的客戶開始排隊時，共有三組客人在店內用餐，從歷史資料中，可檢視共計該三組客人共計點餐六次，但因變數皆須回推客戶抽號碼牌開始排隊的當下情況，因此在統計客戶已點餐的相關資訊時，僅會選取圖中打勾之點餐資料，打叉者視為未來資料，不會納入統計，而相關之概念同步套用至各項開發的變數之中。

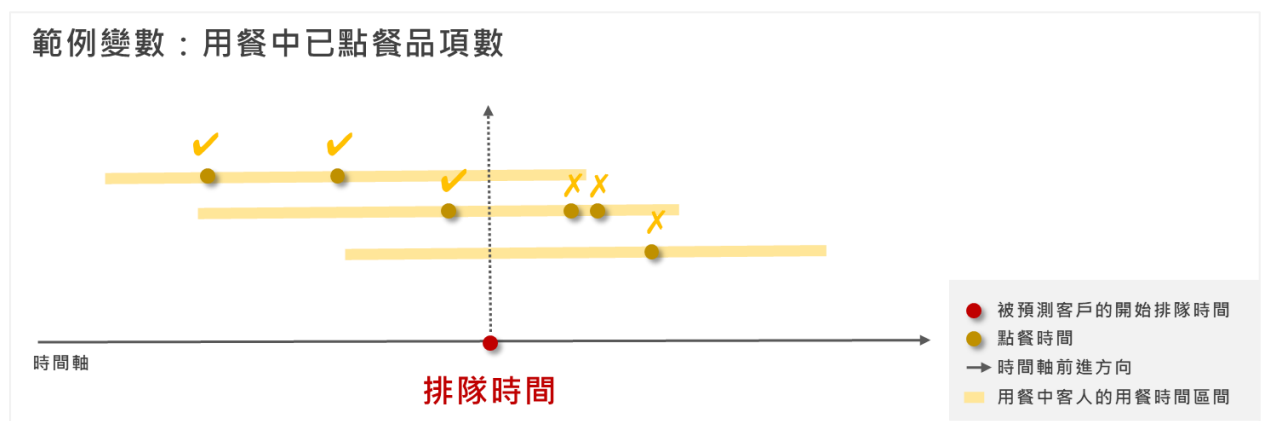


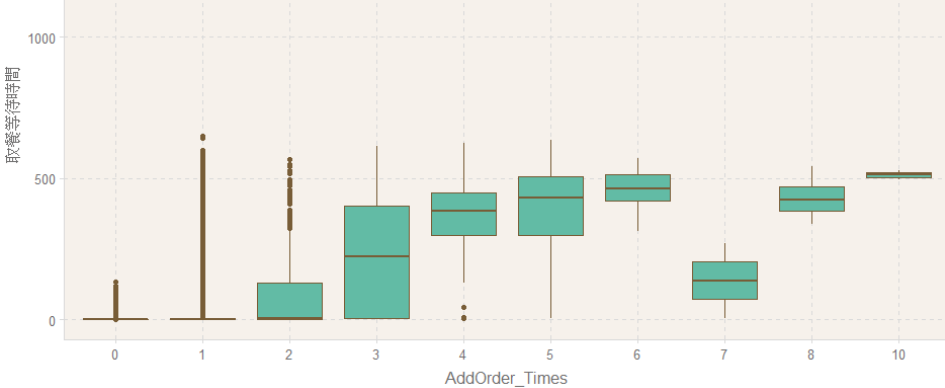
圖3.1 衍生變數開發概念示意圖

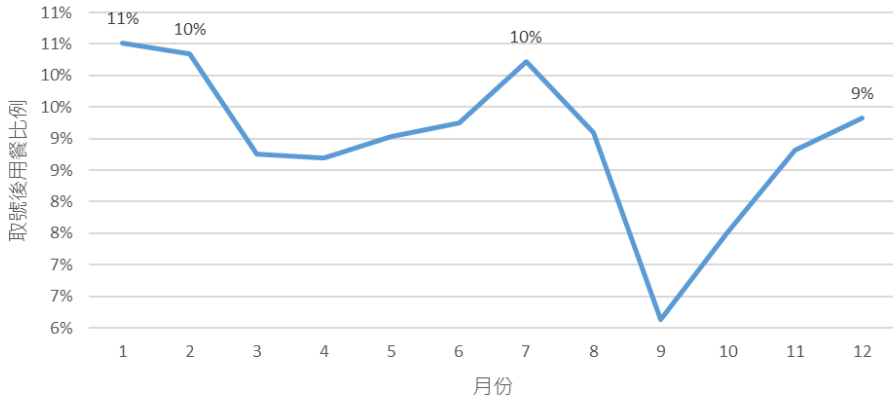
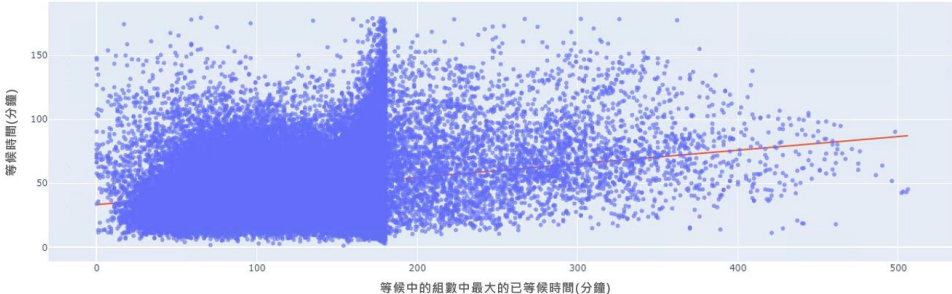
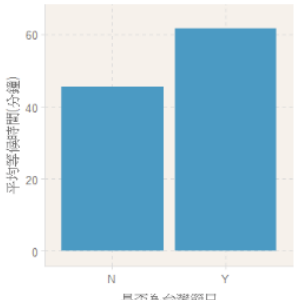
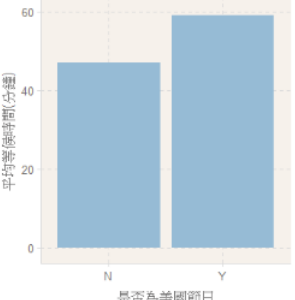
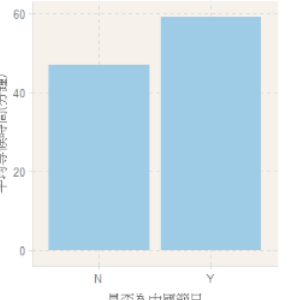
(三) 探索式資料分析

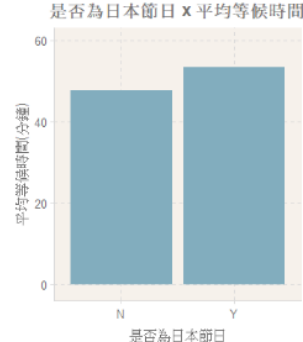
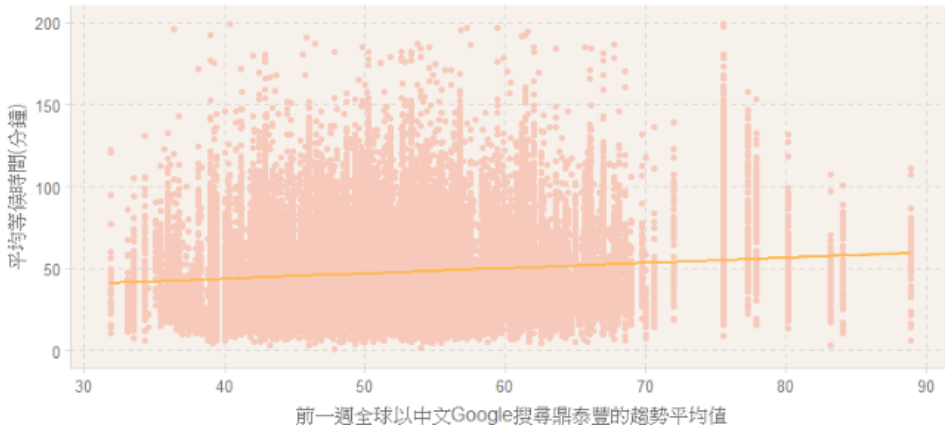
探索式資料分析(EDA, Exploratory Data Analysis)利用視覺化或基本統計數值，對資料進行初步認識，以利後續我們對資料進行複雜或嚴謹分析。可以達到三項特點：(1) 確認資料資訊、結構與特點；(2) 確認資料有無異常；(3)分析各變數間、變數與目標變數間關聯性，找到有趣、重要變數。

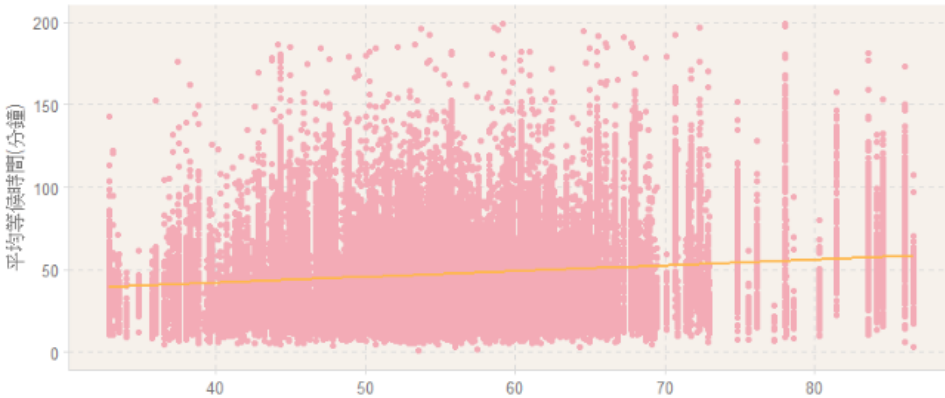
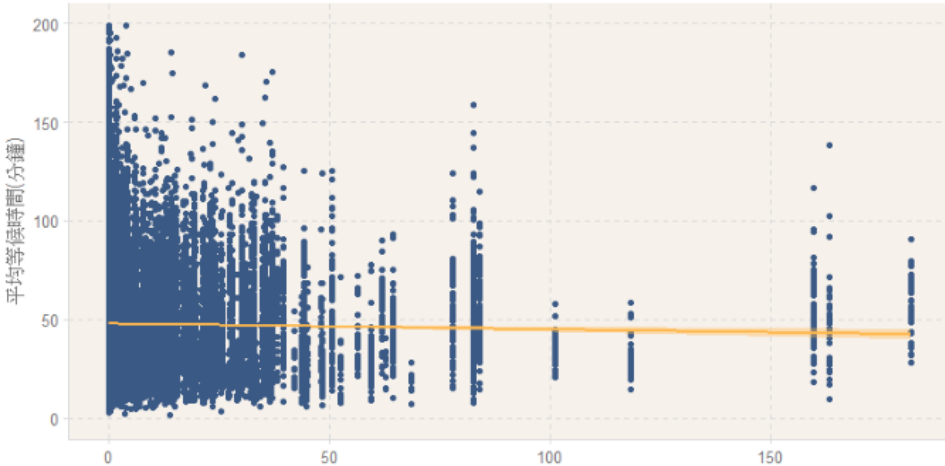
本解題團隊解析所有內外部資料、衍生變數，並與目標變數進行探索，因數量眾多且非所有變數都有發現有趣的地方，以下挑選重要發現進行說明。

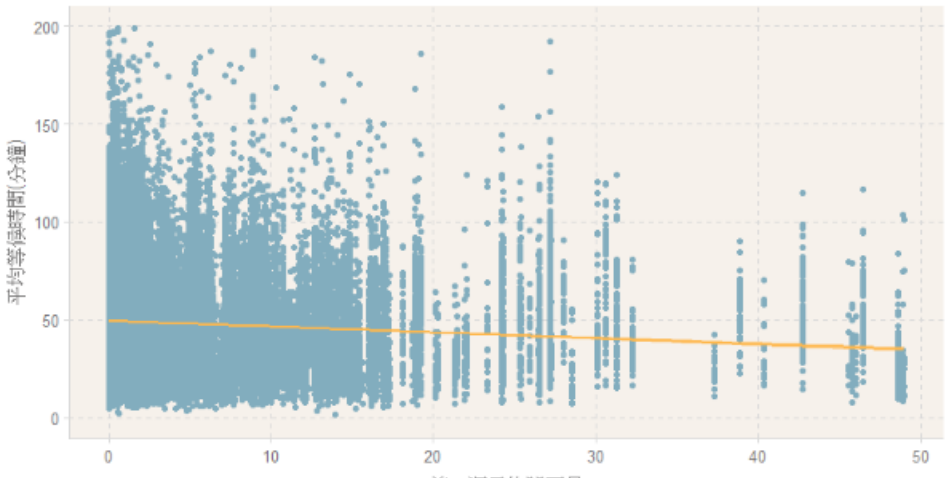
表3.2 數據檢視洞察

資料	說明	視覺化呈現
企業內部數據-點餐	加點次數越多，用餐時間就會越久趨勢。	

資料	說明	視覺化呈現
企業內部數據	春節與暑假期間取號後用餐比例較高	
企業內部數據	由於排隊資料未記錄棄餐的客戶排隊狀態改變的時間，因此經過前述補值後，出現許多資料在該變數皆為180的情形	
外部開放數據-節假日相關	遇到各國節慶假期時，排隊時間平均等候時間較非假期長。	<div><div><p>是否為台灣節日 x 平均等候時間</p></div><div><p>是否為美國節日 x 平均等候時間</p></div><div><p>是否為中國節日 x 平均等候時間</p></div></div>

資料	說明	視覺化呈現																
		<p>是否為日本節日 x 平均等候時間</p>  <table><thead><tr><th>是否為日本節日</th><th>平均等候時間(分鐘)</th></tr></thead><tbody><tr><td>N</td><td>~48</td></tr><tr><td>Y</td><td>~53</td></tr></tbody></table>	是否為日本節日	平均等候時間(分鐘)	N	~48	Y	~53										
是否為日本節日	平均等候時間(分鐘)																	
N	~48																	
Y	~53																	
外部開放數據- 網路搜尋量相關	網路搜尋量越高，排隊時間越長。	<p>前一週全球以中文Google搜尋鼎泰豐的趨勢平均值</p>  <table><thead><tr><th>前一週全球以中文Google搜尋鼎泰豐的趨勢平均值</th><th>平均等候時間(分鐘)</th></tr></thead><tbody><tr><td>30</td><td>~45</td></tr><tr><td>40</td><td>~48</td></tr><tr><td>50</td><td>~50</td></tr><tr><td>60</td><td>~52</td></tr><tr><td>70</td><td>~55</td></tr><tr><td>80</td><td>~58</td></tr><tr><td>90</td><td>~60</td></tr></tbody></table>	前一週全球以中文Google搜尋鼎泰豐的趨勢平均值	平均等候時間(分鐘)	30	~45	40	~48	50	~50	60	~52	70	~55	80	~58	90	~60
前一週全球以中文Google搜尋鼎泰豐的趨勢平均值	平均等候時間(分鐘)																	
30	~45																	
40	~48																	
50	~50																	
60	~52																	
70	~55																	
80	~58																	
90	~60																	

資料	說明	視覺化呈現
		<p>前一週全球以英文Google搜尋鼎泰豐的趨勢平均值</p>  <p>前一週全球以英文Google搜尋鼎泰豐的趨勢平均值</p>
外部開放數據- 氣象相關	降雨量越大，排隊時間越短	<p>前一天降雨量</p>  <p>前一天降雨量</p>

資料	說明	視覺化呈現
		<p data-bbox="1406 247 1608 279">前一週平均降雨量</p>  <p data-bbox="1003 422 1037 582">平均等候時間(分鐘)</p> <p data-bbox="1429 758 1608 782">前一週平均降雨量</p> <p>The figure is a scatter plot with a light beige background and a dashed grid. The x-axis is labeled '前一週平均降雨量' (Average rainfall from the previous week) and ranges from 0 to 50 with major ticks every 10 units. The y-axis is labeled '平均等候時間(分鐘)' (Average waiting time in minutes) and ranges from 0 to 200 with major ticks every 50 units. The plot contains numerous blue data points, mostly concentrated between 0 and 30 on the x-axis and 0 to 150 on the y-axis. A solid orange trend line is drawn across the plot, starting at approximately (0, 50) and ending at approximately (50, 35), showing a slight downward trend.</p>

資料	說明	視覺化呈現										
外部開放數據- 氣象相關	中度颱風的平均排隊時間最長 註：H：強烈颱風；M：中度颱風；L：輕度 颱風；N：無颱風	<p>颱風等級 x 平均等候時間</p> <table><tr><th>颱風等級</th><th>平均等候時間(分鐘)</th></tr><tr><td>H</td><td>40</td></tr><tr><td>L</td><td>40</td></tr><tr><td>M</td><td>54</td></tr><tr><td>N</td><td>48</td></tr></table>	颱風等級	平均等候時間(分鐘)	H	40	L	40	M	54	N	48
颱風等級	平均等候時間(分鐘)											
H	40											
L	40											
M	54											
N	48											

(四) 資料預處理

根據資料屬性進行對應的資料預處理。

● 遺失值補值

1. 類別型變數：因無法利用其他類別進行補值，且演算法碩有空值也無法使用，故將遺失值作為新的類別。
2. 連續型變數：根據變數定義以平均值或者0進行補值。

● 資料型態調整

1. 連續型變數：因各變數單位不同，如人數、組數等，故將連續型變數標準化(Normalization)。
2. 類別型變數
 - One-Hot Encoding：將類別型變數各類別轉成二元資料。
 - Target Encoding：對類別型欄位做 groupby，取 wait_Time 的值，再分別做 mean, mode, min, max, median。
 - Count Encoding：將各個類別型欄位，做每個類別的計數，當成此類別的值。
 - Percent Encoding：對每個類別作百分位數處理，分別取 Q25、Q50、Q75作為此類別的值。

(五) 預測實作與結果

1. 建模說明

在建模前，為評估模型成效，並且不能看解答來解釋，因此須將所有資料進行切分為訓練期間、驗證期間、測試期間。

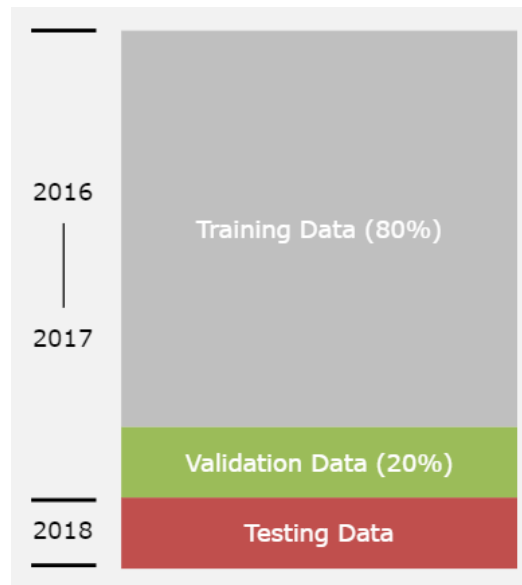


圖3.2 建模資料切分

- 訓練期間為2016-01-01~2017/07/31，訓練資料(Training Data)作為訓練模型用，通常抽全部資料約80%筆數。
- 驗證期間為2017-08-01~2017-12-31，驗證資料(Validation Data)從客觀角度評估目前模型訓練狀況，判斷模型是否過擬合(Overfitting)，以及用來調整模型參數，可得到一些評估的指標(ex: Accuracy, F1-score)，此結果非最後評估模型好壞的指標。
- 測試期間為2018/01/01~2018/12/31，測試資料(Testing Data)用來評估最後模型表現。

2. 基準效度

訓練模型開始，會先將直接套用模型，確認基準版本模型效度，因使用模型眾多，接下來以 TabNet 方法進行說明。

圖3.3中 x 軸為等候時間實際值，y 軸為等候時間預測值，模型相關為77%。可看到預測結果準確度高，但仔細看右下方有低估現象，推估模型受到等候時間較長的客戶影響，導致模型有低估的情況，故將進行異常偵測，排除異常值，讓資料減少影響。



圖3.3 TabNet 模型初版效度

3. 架構與優化

因有部分的客戶在等待過程中，離開其他地方逛街導致過號，資料則會呈現此客戶等待時間較久，且資料面無法排除這些客戶，則採用異常偵測方法(ex: isolationforest)，將異常資料進行排除，如圖3.4所示。

散佈圖上方紅色區塊，表示取號碼牌時當前面客戶較少，仍需要等候較長的時間，目前此部分推測為過號的客戶。



圖3.4 異常資料散佈圖

因此整體模型架構(圖3.5)上，訓練資料會先進行異常偵測，並排除異常值後，再進到不同模型學習，避免模型有低估的情況。接著將測試資料放入得到多個模型結果，可透過 Ensemble Learning 的概念，將所有模型依照模型的 Performance 進行加權平均，減少資料的變異，得到更精準的模型預測。

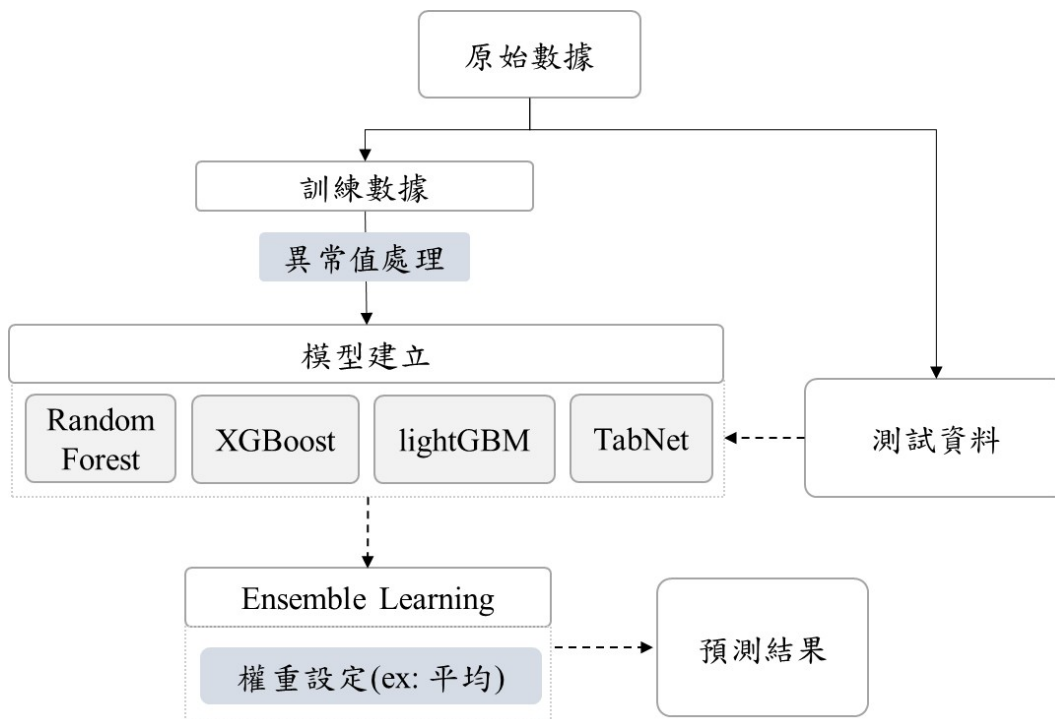


圖3.5 整體模型架構

本解題團隊模型準確率已達到 AIGO 競賽標準，等候時間預測誤差 ± 15 分鐘之準確度70%以上，但未達到鼎泰豐進階期望標準：預測值與實際值誤差落在 ± 5 分鐘之準確度80 %以上，三個時間尺度預測結果如下表。

表3.3 各演算法預測效度

訓練/驗證/ 測試	XGBoost	LightGBM	TabNet	Final
± 5 分鐘	45% / 36% / 30%	53% / 45% / 33%	36% / 31% / 32%	46% / 34% / 33%
± 10 分鐘	74% / 62% / 53%	80% / 71% / 57%	62% / 56% / 55%	73% / 58% / 58%
± 15 分鐘	88% / 78% / 69%	90% / 84% / 72%	79% / 74% / 71%	84% / 73% / 73%

(六) 變數重要性

SHAP (SHapley Additive exPlanations) 將模型的預測解釋分析成每個變數的貢獻，計算每個特徵的 SHAP value，來衡量特徵對預測的貢獻度。

● 變數重要性

圖3.6為各變數平均貢獻度，發現重要性最高的變數為現場排隊人數 (Z03_WAITTIME_cnt)，其次為該組客人人數(Sum of 444 other)。

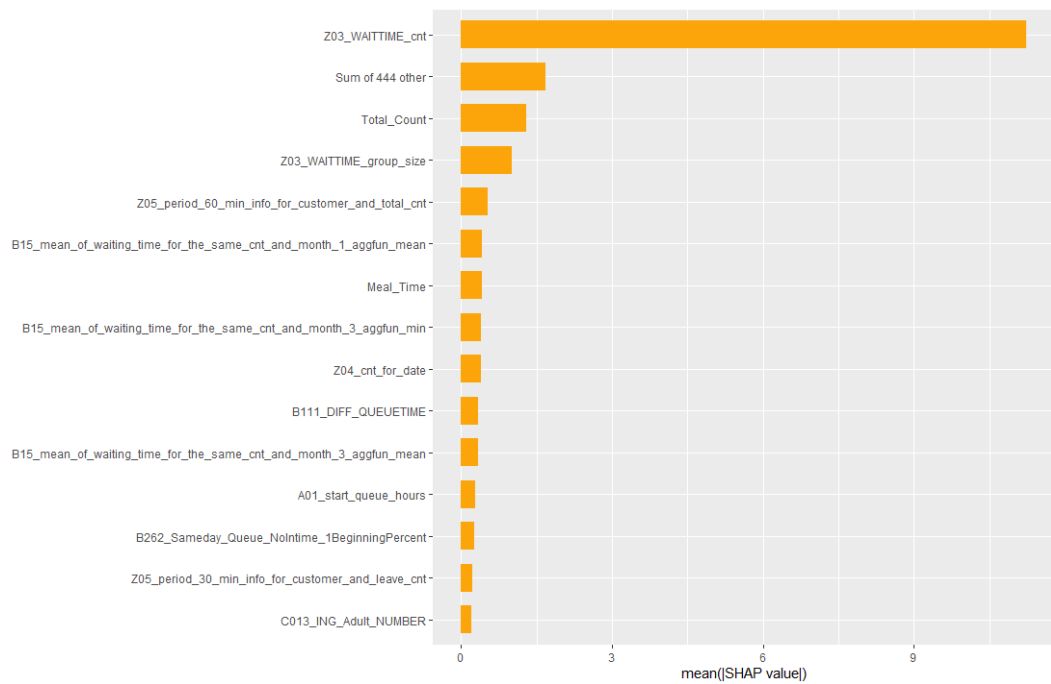


圖3.6 變數重要性(SHAP value)

表3.4 變數重要性定義

重要變數英文名稱	重要變數中文名稱
Z03_WAITTIME_cnt	從 linein 計算排隊等待中人數
Total_Count	總人數
Z03_WAITTIME_group_size	從 linein 計算排隊等待中組數
Z05_period_60_min_info_for_customer_and_total_cnt	前 60 分鐘的來店人數
B15_mean_of_waiting_time_for_the_same_cnt_and_month_1_aggrun_mean	近一個月相同等候組數的平均等候時間(分鐘)
Meal_Time	用餐時間
B15_mean_of_waiting_time_for_the_same_cnt_and_month_3_aggrun_min	近三個月相同等候組數的最短等候時間(分鐘)
Z04_cnt_for_date	當天來店人數
B111_DIFF_QUEUE TIME	前一組開始排隊距離本次開始排隊的時間(分鐘) - 使用 OrderInside 計算
B15_mean_of_waiting_time_for_the_same_cnt_and_month_3_aggrun_mean	近三個月相同等候組數的平均等候時間(分鐘)
A01_start_queue_hours	開始排隊時段(小時)
B262_Sameday_Queue_NoIntime_1BeginningPercent	當天至排隊當下，已排隊但尚無入店時間的排隊號碼 1 開頭的比例
Z05_period_30_min_info_for_customer_and_leave_cnt	前 30 分鐘的棄單人數

C013_ING_Adult_NUMBER	用餐中的客戶大人數
-----------------------	-----------

● 模型結果解釋：SHAP 指標

運用 additive model 與賽局理論，量化出 SHAP 指標來解釋變數對排隊時間的貢獻。

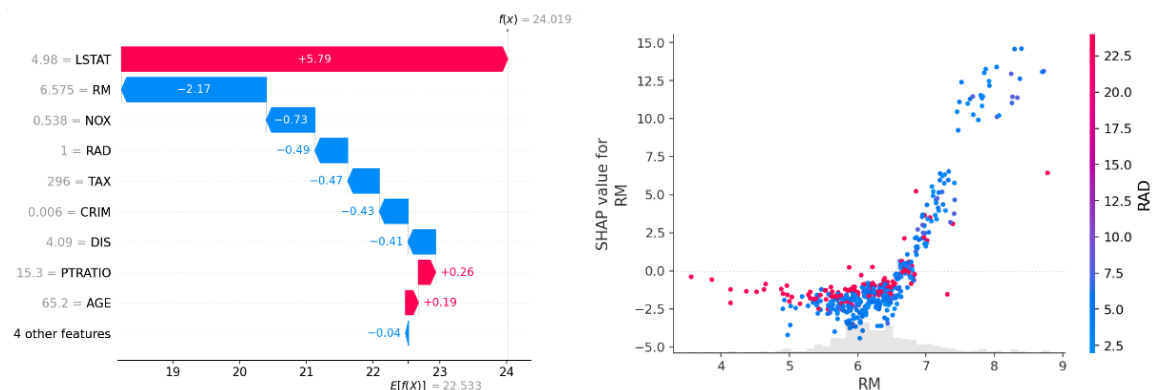


圖3.7 SHAP 指標

● 「現場排隊人數(未點餐)」高重要變數與排隊時間貢獻觀察

將高重要變數「現場排隊人數(未點餐)」與目標變數「排隊時間」貢獻度繪製散佈圖，發現兩變數間的相關性很高。

在圖左下角發現貢獻度低無法提供良好的預測結果，故加入「距離前一組預點餐的時間」(顏色越淺代表時間越長)增加分析維度進行探索。部分客人來的時候排隊人數不多，但他們距離前一組點餐的時間異常的高，推測是在商場逛完其他品牌後才回來用餐導致。

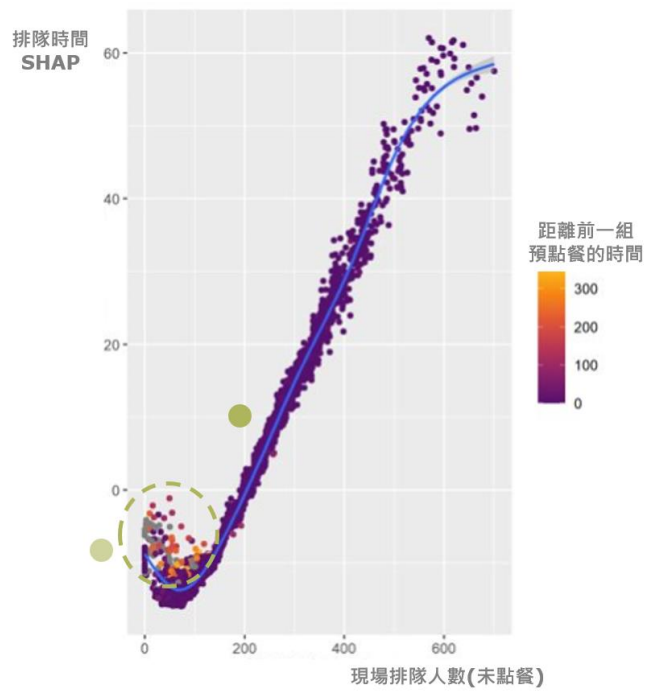


圖3.8 「現場排隊人數(未點餐)」與排隊時間貢獻度散佈圖

● 「現場排隊組數(未點餐)」高重要變數與排隊時間貢獻觀察

除排隊人數影響外，現場排隊組數也會影響排隊等候時間，從圖3.9發現客人來的時候前面排隊組數並不多，但他們的排隊時間特別高，可能是因為在商場逛完其他品牌後才回來用餐。

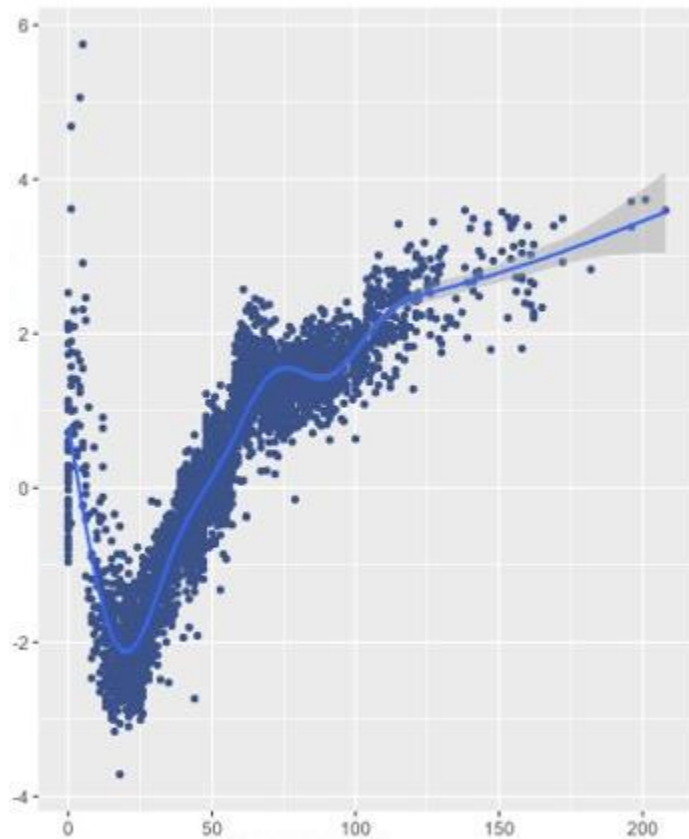


圖3.9 「現場排隊組數(未點餐)」與排隊時間貢獻度散佈圖

因此，可發現客人已知需排隊情況下，猜測會先取號，再去附近商場逛街，而因為鼎泰豐為現場叫號入內用餐，且過號之客戶在後續再經短暫等待仍可入店用餐，故針對這些離開去逛街的客人就容易過號，導致他們排隊時間長，影響模型預測。

四、 成果效益與完成之工作

本解題團隊模型準確率已達到 AIGO 競賽標準，等候時間預測誤差 ± 15 分鐘之準確率70%以上，但未達到鼎泰豐進階期望標準預測誤差 ± 5 分之準確率80 %以上，三個時間尺度預測結果如表4.1。

表4.1 模型結果

準確率	± 5 分鐘	± 10 分鐘	± 15 分鐘
訓練組	46%	73%	84%
驗證組	34%	58%	73%
測試組	33%	58%	73%

模型結果除提供顧客準確的等候時間，增加顧客體驗，讓顧客掌握需等候多久做好自己的時程安排；並可減少員工勞務量，如顧客客訴等。

表4.2為本解題團隊工作細項，過程中與出題方電話、信件、會議溝通數次，使團隊更加了解資料，以利使用更加乾淨資料。

表4.2 工作項目與時程規畫表

類別	項目	起	迄	工作天數	出題方協助
0.前置作業	簽訂 MOU 合約與完成核銷	2022/6/21	2022/6/30	10	✓
1.資料收集	內部資料驗證與欄位定義確認	2022/7/1	2022/7/8	8	✓
	外部資料待辦事項	2022/6/21	2022/6/30	10	
	DB 環境建置與整備	2022/7/14	2022/7/21	8	
2.資料處理	門市與外部資料 EDA	2022/7/28	2022/8/4	8	✓
	資料面問題收集與確認	2022/7/28	2022/8/4	8	✓
	衍生變數發想、定義	2022/6/28	2022/8/4	38	
	衍生變數開發	2022/8/4	2022/8/18	15	
	衍生變數 EDA	2022/8/11	2022/8/25	15	
	*實體會議	2022/8/16	2022/8/16	1	✓
	相關衍生變數會議討論	2022/8/29	2022/9/1	4	✓
	統整資料處理相關程式	2022/9/1	2022/9/15	15	
3.模型建立	前置技術方法研究	2022/7/28	2022/8/4	8	

類別	項目	起	迄	工作天數	出題方協助
	討論模型訓練過程及相關評估指標	2022/8/4	2022/8/11	8	
	各別模型建置與評估 baseline	2022/9/1	2022/9/8	8	
	各別模型優化	2022/8/25	2022/9/8	15	
	Ensemble Learning	2022/9/8	2022/9/22	15	
4.模型解析	模型解釋性與變數重要性	2022/9/8	2022/9/22	15	
	*實體會議	2022/9/15	2022/9/15	1	✓
	統整訓練模型相關程式	2022/9/22	2022/9/29	8	
5.分析報告	報告書 word 撰寫	2022/9/1	2022/10/2	32	
	簡報 PPT	2022/9/1	2022/10/2	32	
	將程式上傳 github	2022/9/29	2022/10/3	5	

五、商業應用價值與創新亮點

本計畫目標為建置等候時間模型，可協助出題方

1. 整合內外部數據：過程中以不同面向提供出題方深度了解內部資料，發現資料記錄完整重要性、系統紀錄錯誤等問題，藉此調整系統資料存取內容，為用餐旅程提供完整、正確數據，落實資料治理，以利未來模型優化等。
2. 衍生各式變數：團隊依據數據技巧、個人經驗提供多面向變數定義，並進行開發，共衍生出459個變數。
3. 重要影響變數：初步找出影響等候時間因子，讓出題方可思考在店鋪營運端如何調整工作流程(SOP)，如針對客人先抽取號碼牌後，去其他地方逛街情況，若已到號可提供簡訊通知等服務。

資料完整性、正確性、改善既有營運 SOP，將有益模型完美落地。



圖5.1 商轉可行性規劃

六、 結論與建議

(一) 結論

模型建置結果準確率為73%，已達到 AIGO 競賽標準：等候時間預測誤差 ± 15 分鐘之準確度70%以上，未能達到出題方進階期望之標準：預測誤差 ± 5 分鐘之準確度80%以上，歸納以下重要影響等候時間之因子，提供出題單位作為營運或現場資深人員評估人力或提供客戶等候時間之參考：

重要程度	因子說明
高	現場排隊人數（未點餐）
	該組客人人數
	現場排隊組數（未點餐）
中	過去1小時，來排隊的人數
	過去1、3個月，相同排隊組數的平均等候時間
	距離前一組預點餐的時間
	過去半小時，棄單的人數
	過去1週，相同時段的平均等候時間

而經過數據檢視與分析，歸納出以下未達進階期望之可能原因為資料品質議題、資料紀錄是否足夠議題、預測標的適切性之議題，由於預測準確度與資料面之相關議題存在非常大的關聯性，每個資料對應之預測標的皆有一個預測效度之瓶頸，在有限時間下盡可能提升預測效度後，我們可知現況資料應用此項預測之瓶頸約落於何處，提供出題單位作為是否調整內部策略或是否進行優化之評估。

(二) 題目限制

經過數據檢視與分析，歸納出以下未達進階期望之可能原因，為本次題目之限制，未來可根據相關優化作為提升效度之方向之一：

- 資料品質議題：
 - (1). 當系統異常時，可能會有屬於營業日但卻缺少資料的情形
 - (2). 不同系統之間的資料能否直接互相能否串接，及串接後數據對應結果是否可以合乎邏輯
 - (3). 時間記錄正確性或確認差異之原因，例如出餐時間早於訂餐時間等
 - (4). 是否有特定情境時的客戶會有部分資料表缺少紀錄之情形
 - (5). 是否有隨時間推進，因企業經資料紀錄方式或系統優化，資料記錄方式、定義或存放欄位不同的議題
- 資料記錄是否足夠議題：
 - (1). 進行等候中相關衍生變數開發時，以歷史資料面檢視，發現有棄餐的客戶（即無入店用餐時間之客戶），屬於正常現象，但因衍生變數之產生需回推過往每個當下之情況進行產製，需得知客戶的等候狀態變更之時間，現況資料尚未進行相關紀錄，較難推得當時各排隊資料的等待狀態知該客戶是否仍等餐中，因此經與出題單位，僅能以統一補上開始排隊時間後之180分鐘，期間作為該客戶等候時間判斷是否仍在等餐，或引用歷史棄餐率進行當時等餐人數之推估，但該些作法皆可能對預測結果造成干擾
 - (2). 等候客戶被叫號之時間資訊為一個重要的資料，因在2016至2018年期間尚未紀錄該項資料，因此在本次模型訓練過程未引用到相關資訊進行衍生變數之開發
- 預測標的適切性之議題：

(1). 本次的預測標的為客戶到店用餐與開始排隊時間之間的等候時間，且因出題單位非常重視客戶服務，採過號仍可到店用餐之模式提供給客戶最好的體驗，因此該需預測之等候時間包含客戶抽取號碼牌後離開至其他地方進行活動之時間，此段因人而異，在資料面上較難蒐集到各客戶之行程規劃，也造成極大之預測干擾因素

(三) 未來建議

系統面調整系統資料存取內容，為用餐旅程提供完整、正確數據，落實資料治理，以利未來模型優化等。營運面可思考店鋪營運端可如何調整工作流程(SOP)，減少異常資料產生。

根據前述限制之可能原因，未來建議可以進行以下相關因應或調整，而在調整後及經過一段歷史資料之累積後，可再次檢視是否有顯著準確率之提升：

- 資料品質議題：未來企業內部可自製或借助第三方顧問服務之力，啟動資料品質議題之相關專案，針對現有資料問題進行盤點，梳理顧客用餐旅程、資料正確性驗證、建立資料治理制度，及進行相關因應之調整
- 資料記錄是否足夠議題：盤點確認顧客用餐過程中，是否有重要資訊在系統中未被記錄到，及確認適切之資料記錄方式，例如排隊狀態變更時間為目前缺少之資料，若欲優化此段，未來資料留存方式建議之作法為新增一個欄位紀錄等候狀態變更時間，使未來能回推個時間點之實際排隊狀態，而若狀態變更可能為多次，則建議新增

一張後續有鍵值可串接之資料表，以歷程方式記錄每一個狀態變更之時間，以利未來分析與利用，同步亦可盤點內部其他之資料留存是否有相似議題，亦可以相同概念進行優化

- 預測標的適切性之議題：由於在資料面上較難蒐集到各客戶之行程規劃，且就客戶角度而言，應較希望得知最早可入店用餐之時間，其後皆可入店用餐，因此建議未來待資料蒐集及累積到位後，將預測標的調整為抽取號碼牌至第一次叫號之時間，作為客戶之實際等候時間

除上述外，未來若進行相關預測之優化，亦可再引入更豐富的外部數據，例如以下：

- 本次原先預計引用之捷運人流外部數據，後來實際未進行使用，原因為訓練期間中較早資料與後續年份資料格式不相符，較難互相對應，未來以較新之期間做為訓練之資料源，可避免該些情況
- 若有目前較難蒐集過往歷史資訊亦可確認是否同步於企業內進行蒐集與累積，例如未來天氣預報等，可爬蟲蒐集於企業內部留存，避免未來如需使用時，外部已無法取得過往歷史資料之議題
- 外部新聞報導等可能與預測標的之等候時間有相關聯，本次僅先引用 Google 搜尋趨勢，未來如有需求可進一步引入外部新聞數據等

而未來若模型欲上線使用，以下議題可能需進一步討論或留意：

- 需評估上線之方式，應於何處呈現預測結果，系統介接之方式可能為需進一步討論之議題，例如排程定期提供或 API 有需求時提供、變數開發和執行預測分別於何系統等
- 需評估上線時使用之資料和訓練模型時資料之分布與資料被記錄的

方式是否有差異，如有差異亦可能造成上線後預測效度之下降

- 如訓練模型之衍生變數，與實際上線應用時顯著變數之開發程式或撰寫邏輯不相同時，因兩者變數需相同，測試資料之效度才能具參考性的議題，因此若程式或撰寫邏輯不相同，後續會有兩者變數一致性比對之議題

希望本次提供之回饋，有助於出題單位現況在門店的等候時間預測或營運策略分析，或對於未來進行相關進階應用之規劃方向參考。

七、 交付項目

#	項目	檔案名稱	提供格式
1	衍生變數清單	1. 衍生數據 schema_AIGO_20221002.xlsx	excel
2	衍生變數 EDA	1. 衍生變數 EDA 衍生變數 EDA_1.docx 衍生變數 EDA_2.docx 衍生變數 EDA-3.xlsx	word excel
3	整體說明簡報/文件 (含影響等候時間 關鍵因子重要性排 行榜)	1. 說明文件 AIGO/report/預測來店用餐的現場 等候時間_不知名團隊_v3.docx 2. 說明簡報 AIGO/report/預測來店用餐的現場 等候時間_不知名團隊_v2.pptx	ppt word
4	執行資料預測的程 式	1. LightGBM Model AIGO/model/lgb/ main.py 2. TabNet Model AIGO/model/tabnet/main.py 3. XGBoost Model AIGO/model/xgb /main.ipynb	py ipynb
5	獎補金核銷表	1. 參考本文件附件一	
6	與出題單位會談之 會議紀錄或簡報	1. AIGO/meeting_record/0816會議 /20220816 會議記錄_提供版.docx 2. AIGO/meeting_record/0915會議 /0915_meeting_record.docx	word

附件一、獎補金核銷表

自中華民國111年07月至111年10月

【請以解題獎補助金30萬元規模編列】

項目	內容細項	關聯性說明	單價 (含稅)	數量	小計 (含稅)
人事費	執行本題目分析之解題人員 相關薪資	參與隊員薪資	40,000	7	\$280,000
雜支	比賽所需雜支	郵資、運費、文件印製、書籍採購	500	4	\$2,000
其他項目	場地費用	比賽討論所需租借場地費用	900	20	\$18,000
總計(含稅)					\$300,000