手搖飲料店設點選址的分析研究 —以台北市為例

組別 第十組

組員

資管碩一 R09725059 吳昀蔚

資管碩一 R09725060 黃瀚陞

資管碩二 R08725004 吳詩萍

資管碩二 R08725011 沈家圻

指導老師 李家岩 老師

壹、 研究背景與動機

一、研究動機

近年手搖飲料店如雨後春筍般興起,十年內成長近九千家,卻也有許多短期「爆 紅」的店家在崛起後短短幾年內便一一收店,除了品牌本身的形象與商品之外,我們亦 好奇:店面的選擇是否亦對其有所影響?而現存相對成功的連鎖飲料店在選址上,又考 慮了什麼關鍵的因素呢?

為了解決我們以上的問題,我們參考了兩篇過去相關的選址研究,並以台北市手搖飲料店為例探討,希望可以了解飲料店在成功選址的關鍵因素,並提供未來欲開設手搖飲料店的創業者做為參考。

二、文獻回顧

1. 結合政府開放資料集於建構開業選址決策支援系統—以店址當地消費能力與鄰近產業特性觀點(杜逸寧等,2018)

以新北市新莊區為例,蒐集該區之交通、地理、人口、各行業店面等資料,考量商圈環境並加入空間自相關分析,利用 KNN、Naive Bayes、LDA、決策數、隨機森林等模型,建立該區的開業選址決策支援系統,提供房東和創業者該地段的租屋、適合開設店面種類等資訊。

2. 地理資訊系統及資料探勘技術在連鎖咖啡店設點之分析與研究(劉奕宏,2011)

針對台北市連鎖咖啡店的選址分析,利用地區的人口、經濟、空間(與其他建築之距離等)等資料,進行變數間相關係數之統計分析,以找出其成功選址之關鍵因素。同時運用資料探勘的分類技術,建構成功選址之分類模型,預測出租店面開設咖啡店的成功機率,最後經由地理資訊系統比對結果。

本小組參考上述兩篇研究的在選址方面的變數選擇以及模型運用,最終根據我們對手搖飲料店的了解調整變數的選擇、挑選樣本並建立假設。

三、問題定義

分析手搖飲料店成功選址的關鍵因素,並預測某地址開設手搖飲料店的成功機率。

貳、 研究方法

一、研究架構

1. 建立假設

針對我們認為與「手搖飲料店選址成功與否」相關的因素,建立以下假設,並 根據以上假設來進行資料的收集以及建立模型。

- a. 選址成功與否與該地點的人口特性有關,例如:性別分布、人口密度、年齡 分佈、教育程度等等。
- b. 選址成功與否與該地點的經濟活動有關,例如:平均租金、購買力指數、零售飽和指數等等。
- C. 選址成功與否與該地點附近有多少重要地標有關,重要地標包含:捷運站、 購物中心、學校、景點等等。

2. 資料收集

依據我們在前段的假設收集需要的資料,收集完後將上述資料根據行政區或里 做合併,作為我們的訓練模型資料集,資料的細項如下:

- a. 台北市飲料店的地址,作為選址成功的資料,隨機選取其他地點,作為選址 失敗的資料。
- b. 台北市各里的人口特性資料。
- c. 台北市各區的購買力指數與零售飽和指數資料。
- d. 台北市各路段的租金行情。
- e. 每個地點附近有多少捷運站、購物中心、學校、景點。

3. 變數篩選

由於我們的研究主題屬於分類問題,不適合使用相關係數當作篩選變數的依據,因此我們利用 Logistic Regression 來做變數的篩選。將變數分別代入 Logistic Regression 的模型中做預測,篩選出 p-value 小於 0.001 的變數作為我們未來放進模型的變數。

4. 模型分析

在此次報告當中,我們將用以下四種方法分別建立分類模型,並調整參數後, 進行模型之間的比較:

- a. Decision Tree
- b. Random Forest
- c. Gradient Boosting
- d. Logistic Regression

5. 模型評估

上述四種模型分別做 10-fold 的 Cross Validation 後,再依據以下五個指標各自的平均來做比較:

- a. Accruacy
- b. Precision

- c. Recall
- d. F1-score
- e. Auc

6. 結論

找出適合用來預測「手搖飲料店選址成功與否」的變數以及模型,並分析的涵義,以及統整未來改善的方向。

二、研究方法

「手搖飲料店選址成功與否」屬於分類問題(共「成功」或「失敗」兩個類別), 因此我們選擇四種適合處理分類問題的方法。

1. Decision Tree

Decision Tree 比較不會受資料分布的影響,而模型的結果簡單易懂,解釋性 也很高,不論回歸或分類問題都很適合。

2. Random Forest

Random Forest 是基於 Decision Tree 的一種 Ensemble Method,會隨機抽取樣本以及變數來建立決策樹,最後結合多棵決策樹的投票結果來進行分類,藉此優化預測的準確率。

3. Gradient Boosting

Gradient Boosting 是一種 Boosting 的方法,每次建立新的模型會參考上一個模型,讓模型的損失函式沿著梯度下降,進而提升模型的效能。

4. Logistic Regression

Logistic Regression 與線性迴歸類似,差別在於線性迴歸的 Y 是連續變數,而 Logistic Regression 的 Y 是類別變數,計算事件成功機率與事件失敗機率的比值,作為勝算(Odds),觀察 X 變數對勝算的影響。

參、 資料收集與結果分析

一、資料收集

1. 成功選址樣本(飲料店)

本小組利用 2020/12/26 網路溫度計網站上手搖飲料排行榜得到排名前五名的飲料店品牌迷客夏、五十嵐、清心福全、麻古茶坊與春水堂,並且透過這五個品牌的官網取得在台北市的分店資訊,作為選址成功的樣本,表 1 為每個品牌在台北市的分店數量統計。

表 1 五個品牌在台北市分店數量統計

飲料店品牌	台北市分店數量		
迷客夏	32		
五十嵐	75		
清心福全	80		
麻古茶坊	19		
春水堂	12		
總計	218		

2. 失敗選址樣本

除了成功選址的樣本外(以下簡稱成功點),亦需要挑選選址失敗的樣本(以下簡稱失敗點)作為模型訓練的樣本資料。

在失敗點的選擇上,由於無法找到明確的飲料店經營失敗的地點,因此改從成功點的反面來推敲。我們假設如果一個地點附近都沒有上述五大連鎖飲料店的分店,那麼就認定此地點應該是屬於選址失敗的樣本。

實際做法上,我們使用了地理資訊系統工具 QGIS 先在台北市行政區底圖上點出前段所提的成功點,並以各成功點為中心,繪製出 300 公尺環域(圖1中橘色點為成功點,紫色半透明圓則是 300 公尺環域範圍)。接著再隨機選取 300 公尺環域以外的地方作為失敗點(圖1中的綠色點)。

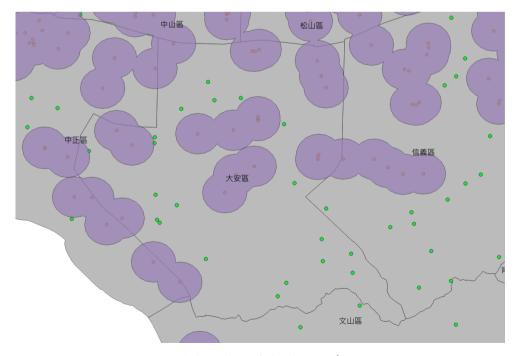


圖 1 選址樣本選取示意

另外,我們移除了位於山地或是河川地形上的失敗點,因為這些地方本身就無法開設店面,選址決策時亦不會將這些地方納入考量範圍,故不將位於山地或河川上的失敗點放入模型當中。最後,共篩選出了206筆失敗點以及218筆成功點作為模型訓練的資料,圖2為各樣本點在全台北市的分佈。由表2及圖1可以看到樣本類別相當平均,避免類別樣本數不平均的情況。。

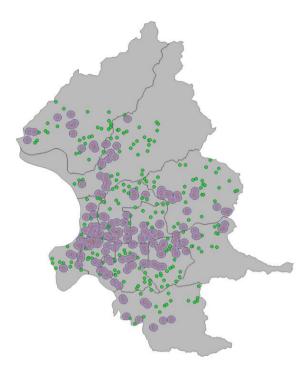


圖2 全臺北市選址樣本分布

表 2 樣本類別數量統計

樣本類別	數量
成功	218
失敗	206
總計	424

3. 變數

在變數的部分,我們分成三大方向,分別是人口統計、經濟活動與一定距離 內重要地標數量,下表列出三個方向的細項:

a. 人口統計資料

利用台北市民政局官網上面的資料取得台北市以里為單位的人口統計相關資料,以下列出變數的細項:

i. 人口數

ii. 性别分布:包含男性比例、女性比例

- iii. 年齡分布:包含平均年齡、20~59歲的人數與比例、以五歲為一個區間,列出從0到100歲以及100歲以上,一共26個變數。
 - iv. 人口密度:利用人口數與面積相除。
 - v. 教育程度:包含博士、碩士、大專、高中職、國中、國小、自修、不 識字等一共八個變數。
- vi. 所得中位數

b. 經濟活動指標

- i. 租金行情:利用「過去三年台北市店舖的實價登錄租金」,從店面租 賃交易行情的資料表找出與樣本點最靠近的三筆資料進行平均。
- ii. 購買力指數:從台北市家庭收支調查結果取得資料。

$$BPI_i = \omega Y_i + \mu R_i + (1 - \omega - \mu) F_i$$

購買力指數可以來評估一個區域是否有市場潛力,如果此指數高 則代表此區域有較強的購買力,選址的成功機率上也有可能越高。

公式中, ω 、 μ 是需要根據條件調整的權重值,在本份報告中分別選擇為 0.5 和 0.3。分區 i 的家庭收入佔整個區域家庭收入的百分比是以 Y_i 來表示,而分區 i 的某類消費支出(在本份報告中是取台北市收支調查中的「食品及非酒精飲料」分類)佔全區此分類消費支出則是以 R_i 表示,而分區 i 的家庭戶數在整個區域中家庭戶戶的佔比則是以 F_i 表示(劉奕宏,2011)。

將權重值乘上各個變數之後,即可得到一個分區的購買力指數 BPI_i,為此分區的購買力指數佔全區購買力指數的百分比。而在本份報告當中,分區的單位是台北市的行政區(共12個)。

iii. 零售飽和指標:從台北市政府民政局、台北市政府家庭收支調查與飲料店家數等三項資料來源計算而得。

$$IRS_i = \frac{H_i \times RE_i}{RF_i}$$

一個地區的市場飽和程度也會影響選址是否成功,通常認為,如果一個地區的市場已經飽和,那將不利於此地區的選址,而零售飽和 指標的值越高,則代表此市場的飽和程度較小、市場潛力較大。

公式當中,一個分區 i 的零售飽和指數 IRS_i 為分區 i 的家庭數 H_i 乘以區域 i 中每個家庭對某類商品(在本份報告中是取台北市收支調查中的「食品及非酒精飲料」分類)的消費支出 RE_i ,接著再除以區域 i 某類商品的營業面積 RF_i (劉奕宏,2011)。

C. 一定距離內重要地標數量變數

利用 google api 計算兩點之間的走路距離,找出每個樣本地址跟捷運站、百貨公司、景點與學校之間的距離,並且找出在距離內的數量有多少。

i. 捷運站:

從台北大眾捷運的公司官網找到每個台北捷運站的地址,並且利用 google api 計算每個捷運站跟樣本點之間透過行走所需要的距離, 下列為利用計算出來的距離所找出的變數。

- 1. 最近的捷運站距離
- 2. 距離五百公尺之內的捷運站數量
- 3. 距離一公里之內的捷運站數量

ii. 購物中心:

從維基百科臺北購物中心列表找到臺北市所有的購物中心,並透過 google map 搜尋得到地址、google api 計算各購物中心跟樣本點間透過行走的所需距離,下列為計算出的變數。

- 1. 距離五百公尺之內的購物中心數量
- 2. 距離一公里之內的購物中心數量

iii. 學校:

從台北市教育局的官網上找到臺北市各級學校名錄與學校地址, 並且透過 google api 計算每個學校跟樣本點之間透過行走所需要的距離,下列為計算出的變數。

- 1. 距離五百公尺之內的學校數量
- 2. 距離一公里之內的學校數量
- 3. 距離一·五公里之內的學校數量

iv. 景點:

從交通部觀光局的官網、台北旅遊網以及 AsiaYo 上找到臺北市的 景點,並且透過 google api 計算每個景點跟樣本點之間透過行走所需 要的距離,下列為計算出的變數。

- 1. 距離一公里之內的景點數量
- 2. 距離兩公里之內的景點數量
- 3. 距離三公里之內的景點數量
- 4. 距離四公里之內的景點數量
- 5. 距離五公里之內的景點數量

綜合以上,加總人口統計、經濟活動與一定距離內重要地標數量等三大類,本 次報告共有53個變數,各類變數數量如表3。

表 3 變數類別數量統計

變數類別	數量
人口統計	37
經濟活動	3
距離	13
總計	53

二、分析

1. 變數選擇

本小組利用 statsmodels.api.Logit()進行 Logistic Regression 去看每個變 數的 p-value,找出小於 0.001 的變數,並且在一定距離內重要地標數量變數上 面,每個重要地標只選一個做為代表(例如:跟捷運站相關的變數只選一個,跟百 貨公司相關的變數只選一個),表4為選出的變數與其對應的 p-value 和係數。 表 4 重要變數的 p-value

選擇的變數 p-value 係數 人口密度 1.09785e-08 9.6273e-06 0.00035 -0.0012租金 距離一公里內捷運站數量 4.49638e-10 0.1480 距離一公里內購物中心數量 4. 45426e-07 0.2578 距離一公里內學校數量 3. 52671e-12 0.3401 4.68443e-09 -0.0272

2. 模型參數挑選

本小組在 Decision Tree、Random Forest 與 Gradient Boosting 這三個模型 中,進行最大深度 max depth 的參數挑選,可選擇的範圍是1到15,以下為三個模 型的選擇結果:

a. Decision Tree

距離兩公里內景點數量

在 Decision Tree 的分類模型當中,有最高準確度得最大深度 max depth 為 3, 圖 3 為最大深度為 1 到 15 的模型準確度折線圖:

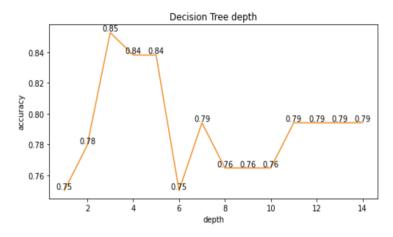


圖 3 Decision Tree 參數 max_depth=1 到 15 的 accuracy

b. Random Forest

在 Random Forest 的分類模型當中,有最高準確度得最大深度 max_depth 為 4、5 與 6,取最小的 4 作為後續模型訓練時的最大深度,圖 4 為最大深度為 1 到 15 的模型準確度折線圖:

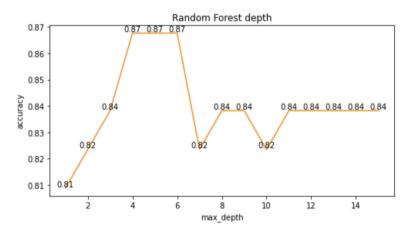


圖 4 Random Forest 參數 max_depth=1 到 15 的 accuracy

c. Gradient Boosting

在 Gradient Boosting 的分類模型當中,有最高準確度得最大深度 max_depth 為 1 與 2 ,取最小的 1 作為後續模型訓練時的最大深度,圖 5 為最大深度為 1 到 15 的模型準確度折線圖:

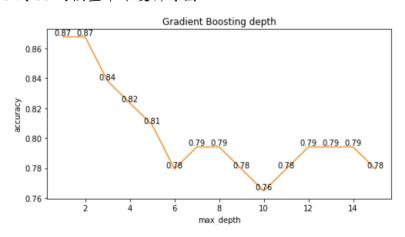


圖 5 Gradient Boosting 參數 max_depth=1 到 15 的 accuracy

3. 模型的選擇

本小組將資料進行 10-fold 的 cross validation,計算出 accuracy、precision、recall、f1-score 與 roc_auc 的平均值,並且利用上述指標的平均值進行模型的選擇,下圖是這五個指標在四個模型的數值圖:

a. Accuracy:以Random Forest 的 accuracy 最高,約0.816。

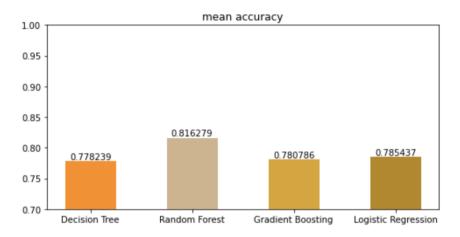


圖 6 四種模型經過交叉驗證之後的平均 accuracy

b. Precision:以Logistic Regression 最高,約0.804。

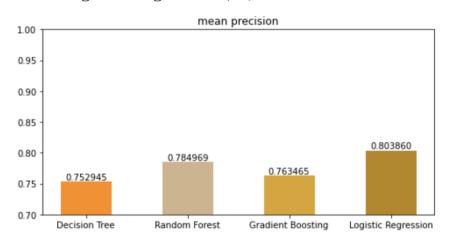


圖 7 四種模型經過交叉驗證之後的平均 precision

c. Recall:以Random Forest的accuracy最高,約0.895。

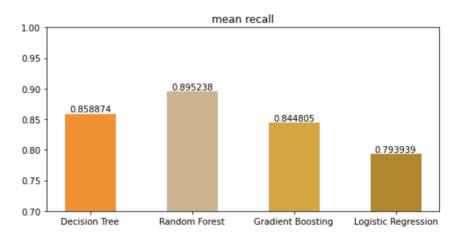


圖 8 四種模型經過交叉驗證之後的平均 recall

d. fl-score:以Random Forest的accuracv最高,約0.834。

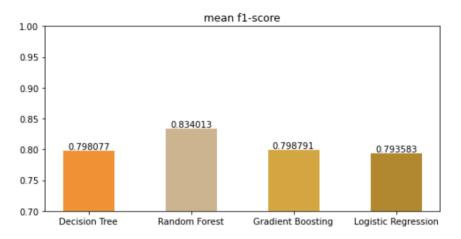


圖 9 四種模型經過交叉驗證之後的平均 fl score

e. roc auc:以Random Forest的accuracy最高,約0.883。

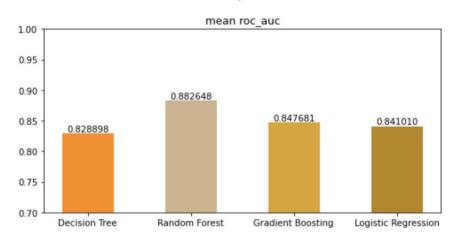


圖 10 四種模型經過交叉驗證之後的平均 auc

綜合上述五種指標,四種模型之中以 Random Forest 的模型表現最好,因此在本次報告中選用 Random Forest 作為最佳模型。另外,四個模型的 accuracy 差異不大,故從 Logistic Regression 的表現與另外三種非線性的模型很接近來看,可以推測被選到的變數跟選址是否成功應該有線性關係。

表 5 四種模型經過交叉驗證之後的各項指標數值

模型/指標	accuracy	precision	recall	f1-score	roc_auc
Decision Tree	0. 778	0. 753	0.859	0. 798	0.829
Random Forest	0.816	0. 785	0.895	0.834	0.882
Gradient Boosting	0. 781	0. 763	0.845	0. 799	0.848
Logistic Regression	0. 785	0.804	0. 794	0. 793	0. 841

三、解釋與管理層面意涵

在變數的選擇方面,本小組發現會影響到飲料店的選址是成功與否的因素是人口密度、租金、距離一公里內的捷運站數量、距離一公里內的購物中心數量、距離一公里內的學校數量與距離兩公里內的景點數量。

其中人口密度部分呼應到我們提出的假設 a 人口特性會對飲料店選址是否成功有影響;租金部分呼應到假設 b 經濟活動指標會影響到選址是否成功;最後跟一定距離內重要地標數量變數則是呼應到假設 c 選址成功與否與該地點附近有多少重要地標有關。

此外,這六個變數的重要程度也不相同,而變數的重要性跟 p-value 有關,將 p-value 由小到大排序的話是距離一公里內學校數量、距離一公里內捷運站數量、距離兩公里內景點數量、人口密度、距離一公里內購物中心數量與租金,可以發現說一定距離內重要地標數量的影響最重要,再來是人口統計最後是經濟活動。

透過決策樹圖也可以發現距離的影響較大,接著是人口特性,最後是經濟活動。同時也能看出當以下幾點滿足時,選址成功的機率越高:

- 1. 距離一公里內的學校數量多
- 2. 人口密度高
- 3. 距離兩公里內景點數量多
- 4. 距離一公里內購物中心多

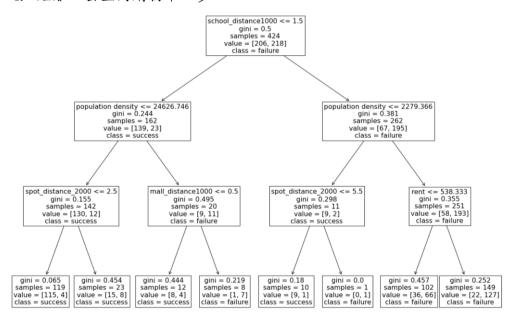


圖 11 Decision Tree 的分支圖

最後在管理層面的意涵方面,由模型結果可以得知「與重要地標的距離」(如學校、景點等)是決定選址是否成功最重要的因素,推測結果之所以如此是因為重要地標 附近往往會帶來人潮與消費,進而使得選址的成功機率增加,此結論與過去我們所認知 飲料店往往開在熱鬧或車水馬龍的地方的概念相符。 然而,若換個角度思考,可以得知當前手搖飲料店的市場都受限於「距離」因素,導致服務的範圍普遍限於店面附近,是非常強調地域性的行業。不過,隨著這幾年外送平台的興起,餐飲外送打破了許多地域的限制,對於使用外送平台叫餐的消費者與餐飲業者來說,覓食或是服務的範圍都比起過往拓展了許多,手搖飲料店亦然。因此,若是手搖飲料店可以將店面地址選擇在地段較為冷門但租金便宜許多的地點(如離商圈中心較遠或是易達性較低的巷子內等),但同時結合餐飲外送平台,將省下的租金反映在飲料的售價或是品質的提升上,也許會是另外一種新型態的商業模式。而從本次報告的分析看起來,此種商業經營模式還沒有被各大熱門連鎖飲料店所採用,或許是非常有潛力的新嘗試。

肆、 結論

一、系統實作

由於我們希望此次報告的結果可以運用在真實場域,因此決定以本章節所訓練出來模型為基礎,實作一個飲料店開店選址時的決策輔助系統,圖 12 為此系統書面。

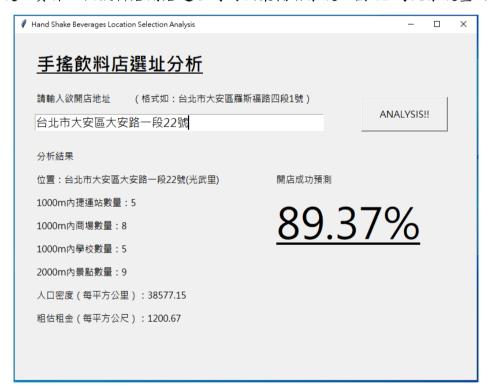


圖 12 選址輔助系統

使用方式上,使用者可直接輸入欲開店的地址,系統就會根據我們整理的背景資料即時計算出此地址的 1000 公尺內捷運站數量、1000 公尺內商場數量、1000 公尺內學校數量、2000 公尺內景點數量、人口密度以及粗估租金。接著再將這些變數以前述表現最好的隨機森林模型(深度為4)預測並顯示此地點的開店成功機率,供決策者作為開店選址時的參考。

二、研究成果

1. 選址成功的關鍵因素:

根據分析結果,選址成功的主要因素包含:距離一公里內的學校數量多、人口密度高、距離兩公里內景點數量多、距離一公里內購物中心多,顯示位址的空間因素影響較大,並且可根據上述因素推斷成功選址的主要原因可能為處於人潮較多的「熱鬧地段」。

2. 預測率:

根據四種模型分別經過 10-fold 的 Cross Validation 的交叉驗證後的結果顯示預測率約在八成左右,並且從 Logistic Regression 與其他三者之差距不大來看,顯示所選變數與成功選址之間可能存在線性關係。

三、未來展望

1. 樣本選擇

a. 成功樣本:

本次僅考慮連鎖的因素,將連鎖視為成功的象徵,然而並未考慮各分店 的經營狀況,故未來的研究在成功樣本的選擇可以考慮更多經營的因素做為 判別成功與否的考量。

b. 失敗樣本:

本次研究雖已根據地形排除過於偏遠的地區,但仍選到部分商家稀少的 地址,未來選取時可再根據當地飲料點設點情形、各區成功飲料店分布等作 為失敗樣本選擇之考量。

c. 增加總樣本數

2. 變數選擇

- a. 空間變數部分的可以擴大範圍並增加區間數,找出影響最大的範圍。此外,由於近日外送服務盛行,以打破近距離的店家優勢,未來變數選擇譯音將外送之因素考慮進去。
- b. 本次經濟活動指標之變數以區為單位,未來若縮小尺度也許能有更精準的預 測或不同發現。

伍、 參考文獻

- 杜逸寧、徐維澤、黃祥晉、許明楷、林鈺翔、洪健傑(2018)。結合政府開放資料集於 建構開業選址決策支援系統—以店址當地消費能力與鄰近產業特性觀點。科儀新知, 215期,P4-21。
- 2. 劉奕宏(2011)。地理資訊系統及資料探勘技術在連鎖咖啡店設點之分析與研究。政治 大學資訊科學學系學位論文。
- 3. Jamleecute。Decision Tree 決策樹 | CART, Conditional Inference Tree, RandomForest。

https://www.jamleecute.com/decision-tree-cart-%E6%B1%BA%E7%AD%96%E6%A8%B9/。

- 4. Chung-Yi。ML 入門(十七)隨機森林(Random Forest)。
 https://medium.com/chung-yi/m1%E5%85%A5%E9%96%80-%E5%8D%81%E4%B8%83%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-6afc24871857。
- 5. swensun。機器學習之 GBDT(簡單理解)。https://iter01.com/66707.html。
- 6. 永析統計及論文諮詢顧問。羅吉斯迴歸分析(Logistic regression, logit model)-統計說明與 SPSS 操作 https://www.yongxi-stat.com/logistic-regression/。
- 7. 網路溫度計一飲料排行榜。 https://dailyview.tw/top100/topic/35?volumn=1&page=1。
- 8. 迷客夏。https://www.fonfood.com/store/293295/branch。
- 9. 五十嵐。http://501an.com/web/pointsout.asp 。
- 10. 清心福全。

https://www.chingshin.tw/store.php?city=5&area=&keyword=&page=1#change_point

- 11. 麻古茶坊。http://www.maculife.com.tw/store list.asp 。
- 12. 春水堂。https://chunshuitang.com.tw/location.php?loc=l。
- 13. 台北市政府民政局—臺北市每月各里人口數及戶數。
 https://ca.gov.taipei/News_Content.aspx?n=8693DC9620A1AABF&sms=D19E9582624D8
 3CB&s=6F385E21D02AAFD5。
- 14. 台北市政府民政局—臺北市每年人口數依性別及年齡分。
 https://ca.gov.taipei/News_Content.aspx?n=8693DC9620A1AABF&sms=D19E9582624D8
 3CB&s=78DC4B104D9D374E。

- 15. 台北市政府民政局—臺北市各行政區 15 歲以上現住人口數按年齡及教育程度分。 https://ca.gov.taipei/News_Content.aspx?n=8693DC9620A1AABF&sms=D19E9582624D8 3CB&s=49EE949EED38EF73。
- 16. 財政部財政資訊中心。https://data.gov.tw/dataset/103066。
- 17. 內政部不動產成交案件實際資訊資料供應系統。 https://plvr.land.moi.gov.tw/DownloadOpenData。
- 18. 台北市政府主計處。

https://dbas.gov.taipei/News.aspx?n=783F3F5F4E6C457E&sms=CF3C6E6E7767448F。

- 19. 北捷。https://www.metro.taipei/cp.aspx?n=91974F2B13D997F1。
- 20. 交通部觀光局的官網。https://www.taiwan.net.tw/ml.aspx?sNo=0001090。
- 21. 台北旅遊網。https://www.travel.taipei/zh-tw/must-visit/attractions-top25。
- 22. AsiaYo https://asiayo.com/event/taipei_fit.html •
- 23. 台北市各級學校。

https://www.doe.gov.taipei/News_Content.aspx?n=0D95D51944239B9E&s=F9C764C609 2F40B1

- 24. 大專院校。https://ulist.moe.gov.tw/Query/Area/N1。
- 25. 維基百科-台北市購物中心列表。

https://zh.wikipedia.org/wiki/%E8%87%BA%E5%8C%97%E5%B8%82%E8%B3%BC%E7%89%A9% E4%B8%AD%E5%BF%83%E5%88%97%E8%A1%A8。