為什麼要做這個專案？

這個專案想解決什麼問題？

數據如何蒐集、數據如何清理？

如何呈現數據，來回答專案中的問題？

第一步：想好為什麼要做這個作品集？

第二步：這個專案碰到的問題是什麼？

第三步：資料蒐集、資料清洗

第四步：數據分析及視覺化

第五步：簡報，準備你的故事

(Why) 為什麼要做這個專案？→ 第一步：跟你的背景、跟你要申請的公司有什麼關係？

(What) 這個專案想解決什麼問題？→ 第二步：核心問題是什麼？

(How) 數據如何蒐集、數據如何清理？第三步：為什麼要蒐集這些數據？如何做數據清理？

(Output) 如何呈現數據，來回答專案中的問題？第四步：你做出的結果是什麼？

(Conclusion & Insight) 結論？為這個問題下你的總結，不一定要對，但故事的結尾一定要跟前面呼應，要通順！

**為什麼要做這個專案？**

前公司的工作內容是透過各項特徵(天氣、氣溫、進場購買人數、到貨量……等)來判斷水果的底價，主要是以人工經驗法則來決定，所以會因為裁決者的個人習慣有所偏誤，故想以較科學的方式操作該工作內容。

**這個專案想解決什麼問題？**

透過公司的每日交易資料找出制定火龍果價格的模型，並且利用客群分析找出舊客群中那些是火龍果購買頻率較高的客群，以及其購買價位在哪，再發掘可以推銷的新客群有哪些，以作為日後推銷火龍果的客群名單。

**數據如何蒐集、數據如何清理？**

該資料為公司網站的刊登資料，可下載為xls檔案進行存取，利用pandas讀取xls檔案並以dataframe型態操作資料，最後將清整後的資料利用pymysql儲存在mysql資料庫的table，以方便日後資料的讀取。

**資料分析：**

**欄位說明：**

dt日期 sc供應代號 ft水果品種 sp規格 gd價格等級 ct銷售數量 nw淨重(公斤) up單價(每公斤) tc傳票種類 br(購買人) sn(拍賣人) it(拍賣順序)

**資料清整：**

1.去除傳票種類5的交易，因為5為農藥超標的報廢類別。

2.發現811(白肉火龍果)的欄位是有空值得(原因811並非每日都有到貨)，所以空值部分以0作填補。

3.整理交易日期的格式(原始顯示方式：0111-01-01)轉換成(2021-01-01)

4.從<https://e-service.cwb.gov.tw/HistoryDataQuery/>天氣觀測網站下載CSV檔，加入溫度和降雨的特徵

5.新增星期欄位

6.新增欄位為判斷是否為節慶或是假期

7.新增欄位判斷是否為初一或十五

8.新增欄位為當日進場購買的總人數

9.新增欄位紅龍果和白龍果分別的到貨總重量

10.新增欄位每日平均價格

11.新增欄位全場的水果總重量

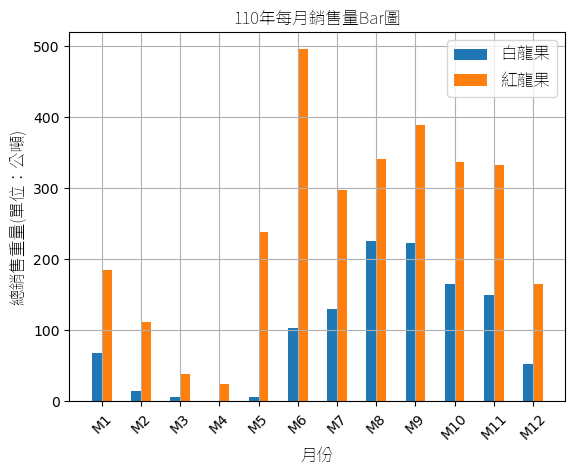
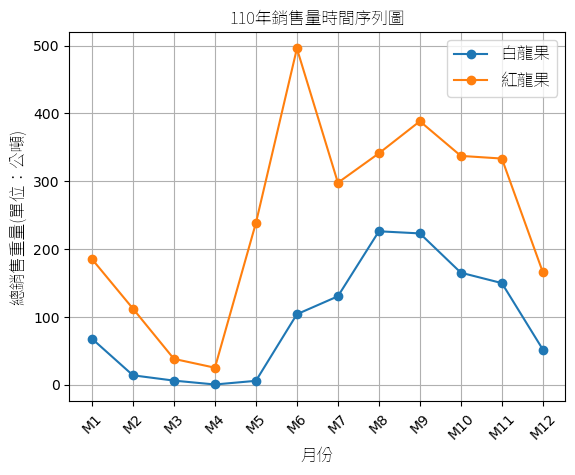
12.將拍賣員、火龍果、月份欄位都進行單熱編碼

13.最後得到的特徵欄位有[811(白龍果), 812(紅龍果), tp(溫度), rh(是否下雨), wd(星期), fs(是否遇到節慶), lu\_dt\_fs(是否為初一或十五), br(進場人數), total\_nw(紅龍果或白龍果的重量), all\_total\_nw(全場水果總重量), 【008, 018, 032, 039, 049, 051, 059, 061, 063, 066, 068, 069, 083,

096, 099, 107, 595, 707】(拍賣員編號), 【1, 2, 3, 4, 5, 6, 7, 8, 9,

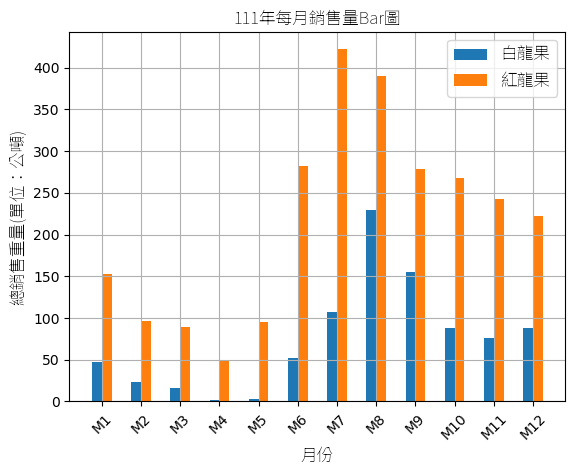
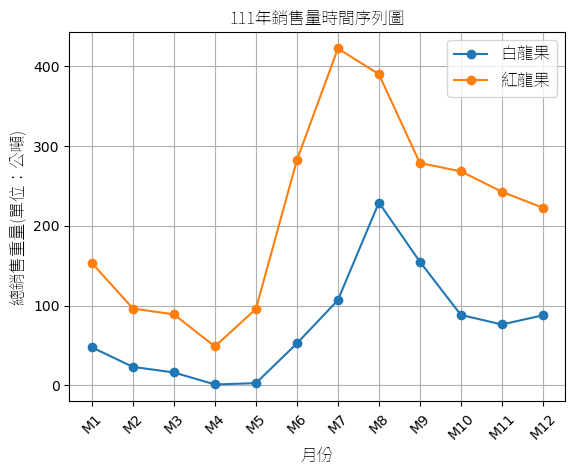
10, 11, 12】(月份)]

**觀察火龍果的銷售量：**

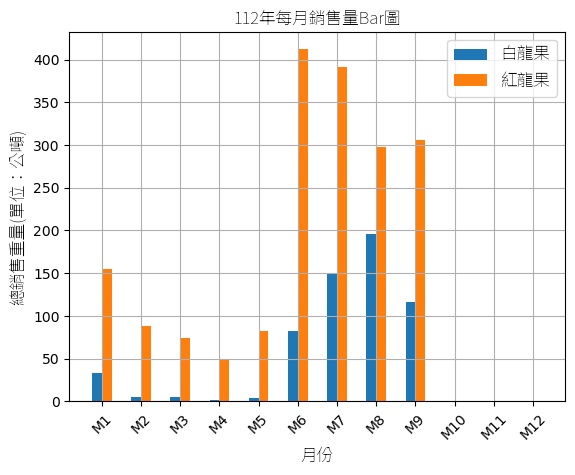
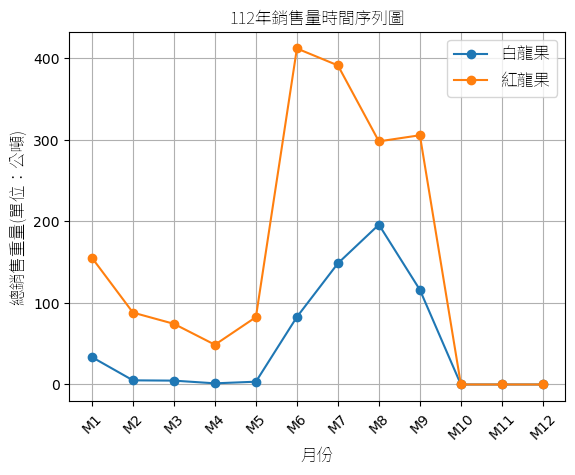
** **

首先北農為當日進貨當日銷售完畢，所以銷售量也代表到貨量。

在110年度中火龍果的淡季落在1月~4月、12月，尤其是3、4月的量是最低的，而產期落在5月~11月這段區間，又以紅龍果比白龍果的到貨量還多。

** **

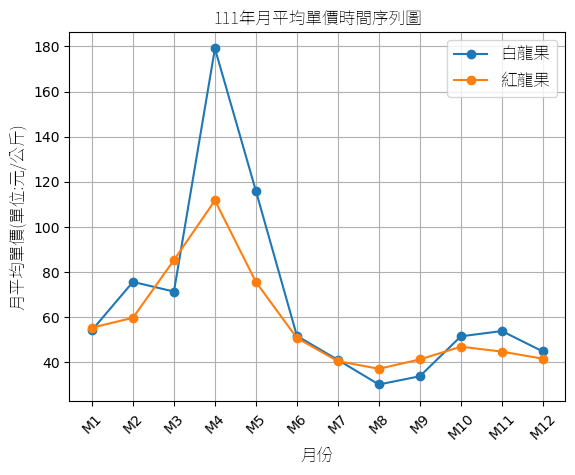
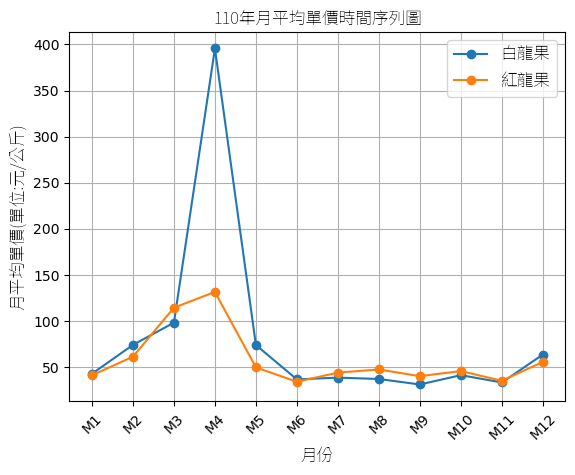
在111年度中火龍果的淡季落在1月~5月，在4月又是最低銷售量，而產期落在6月~12月，一樣是以紅龍果到貨量較多。

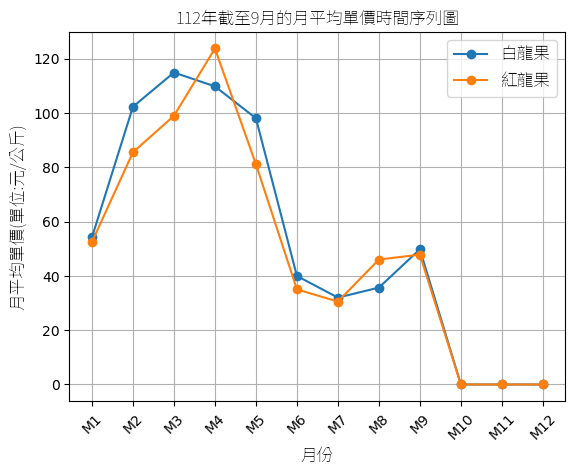
** **

在112年度中火龍果的淡季同樣落在1月~5月，並且同樣在4月也是最低銷售量，而產期以目前統計到9月來看，也是6月開始產量變多，並且一樣也是以紅龍果到貨量較多。

綜合上述圖表可以發現紅龍果的到貨量都比白龍果來的多，並且火龍果產季大致上落在6月~12月這段區間，故這段期間較容易面臨火龍果因產量太多造成價格崩盤的情況，並且在4月份是火龍果銷售量最少的時候，故容易會有供不應求的情形。

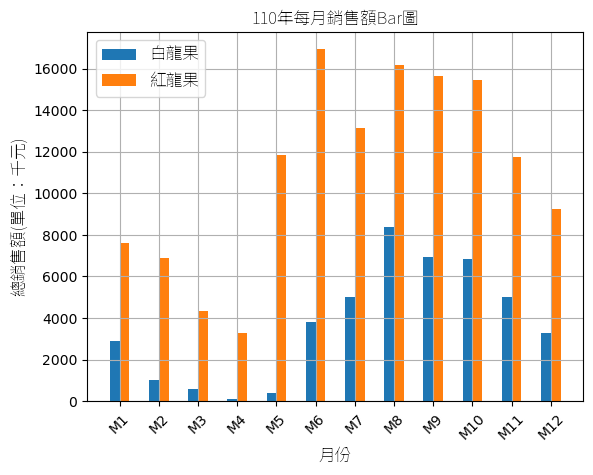
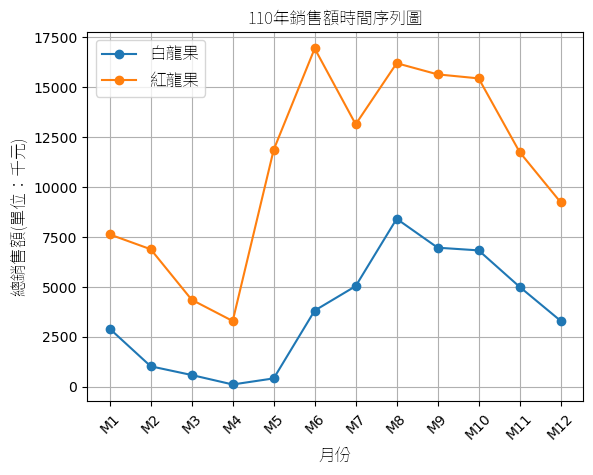
**觀察火龍果的月平均單價：**

****

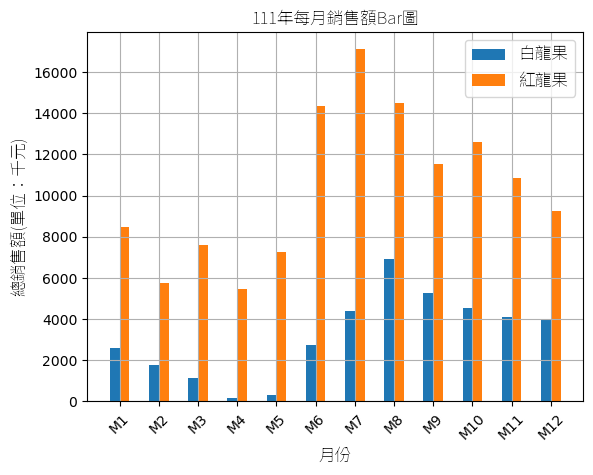
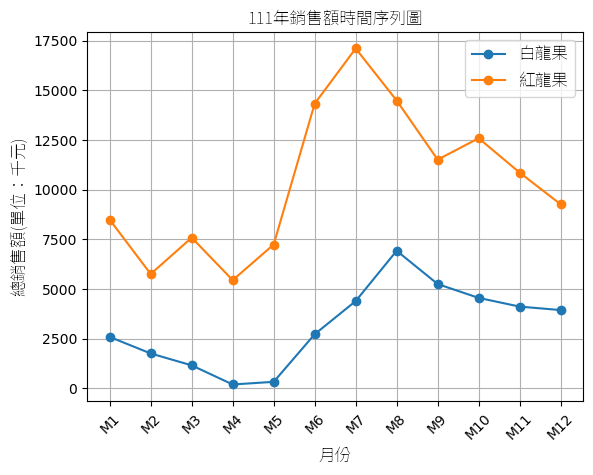
****

火龍果的月平均單價在每年4月份均為最高峰時期(因為每年4月份是火龍果量最少的時候，供給比需求還少)，而在6月份~隔年1月份的月平均單價都是處於較低的(因為這段時期是火龍果量較多的時候，供給比需求還多)，並且起漲點落在2月，再從5月開始進入相對的低點。

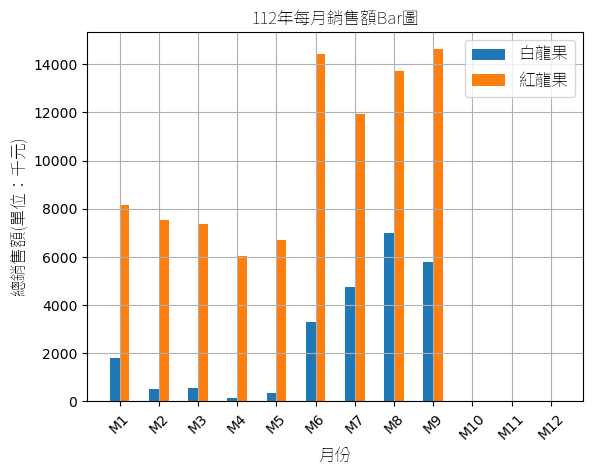
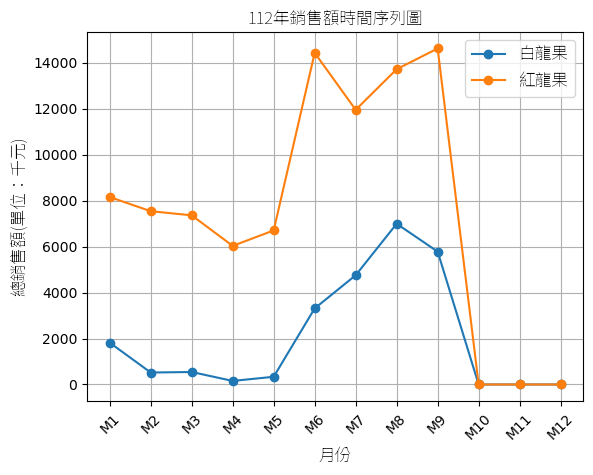
**觀察火龍果的銷售額：**

** **

110年度火龍果在6月~12月雖然單價是較低的，但是因為量較多的關係，銷售額也相對是較多的，而1月~5月雖然單價較高，尤其是4月份，但是因為量較少的關係，造成銷售額也相對較低，尤其是4月份的銷售額也是最低的，圖形分布與110年度銷售量的分布圖類似。

** **

在111年火龍果的銷售額分佈基本上與110年度的差不多，並且也與111年度銷售量的分布圖類似。

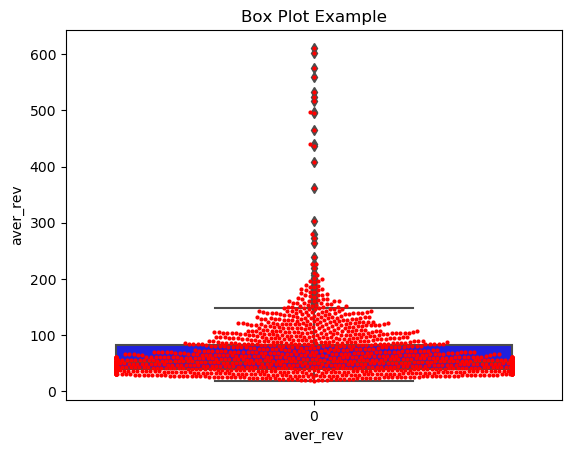
** **

在112年火龍果的銷售額(目前來說)同樣在6月~9月是相對較多的，1月~5相對較低，也與112年度銷售量的分布圖類似。

綜合上述圖表可以發現紅龍果的銷售額也比白龍果多，並且火龍果的銷售額分佈圖是跟銷售量的分佈圖類似的，都是在6月~12月相對較多，在4月份是最低的。

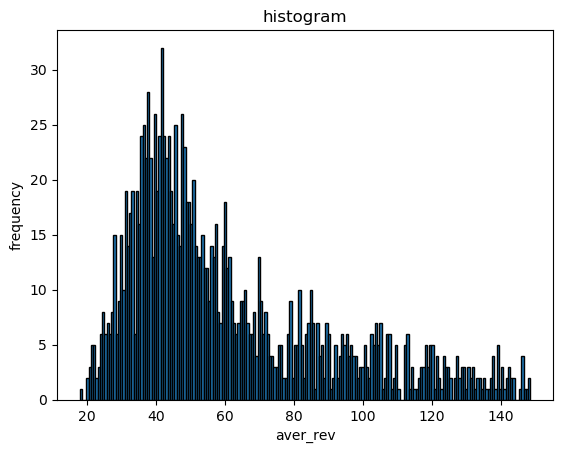
**每日平均價格預測：**

使用XGBoost regression模型(原因：XGBoost在預測上因為有內建正則化和提前停止訓練，所以比較不容易過度擬合，並且模型會透過計算gain information選擇特徵的重要程度，所以也可以了解哪些特徵較為重要)



離群值:小於-25.3元，大於148.7元

離群值的資料筆數：71筆，剃除後剩下資料總筆數：1460筆



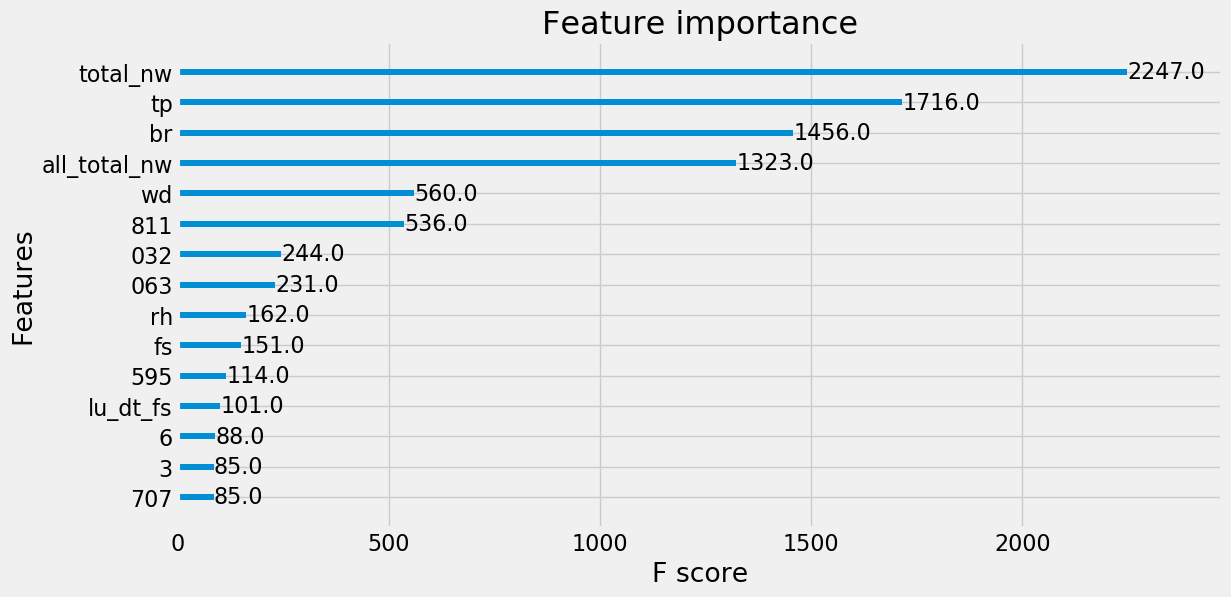
因為資料的分佈是呈現右偏，所以選擇使用rmsle當作評估指標，rmsle為均方對數誤差，使用原因是因為rmsle會將數據經過對數轉換，計算出來的誤差會較均勻，比較不會讓模型受到異常值的干擾。

透過gridsearchcv找出最佳超參數組合為：

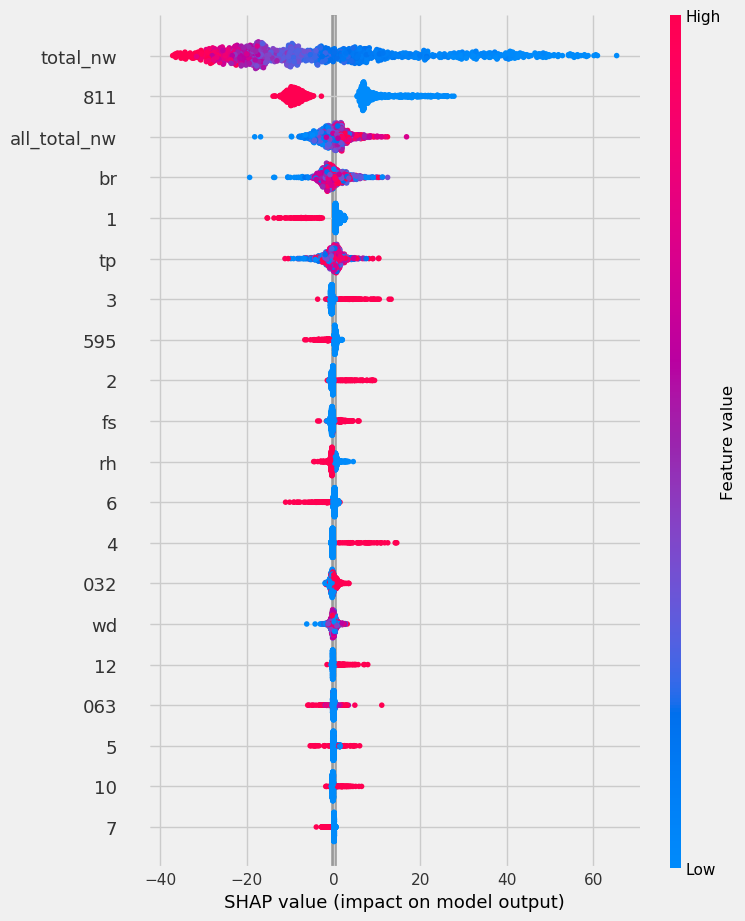
{'learning\_rate': 0.05, 'max\_depth': 5, 'n\_estimators': 500}

其中驗證集的R2為0.80210(模型可解釋性佔比) MSLE為0.19392

R2蠻接近1並且MSLE也蠻接近0，模型訓練出來的指標都蠻好的。

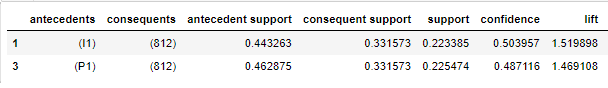


可以看到前幾名重要的特徵為total\_nw(火龍果的總重量)、tp(溫度)、br(全場人數)、all\_total\_nw(全場水果的總重量)、wd(星期)



並且從SHAP圖也可以看到total\_nw(火龍果總重量)對預測價格的貢獻度最大，並且該特徵的值越小對預測正向貢獻程度就越大，也就是火龍果的量越小，對預測價格的正向影響就越大，而tp(溫度)、wd(星期)值略小對時，對預測價格的負向影響略大，而br對預測價格較無明顯的正向或負向影響，其中比較意外的是all\_total\_nw(全場的水果總重量)的值越大，竟然對預測價格是正向的影響。

**尋找可能會購買火龍果的潛在客群：**

****

從支持度可以看到全場的所有交易中有1/5的支持度I1(木瓜)和812(紅龍果)會同時被購買，也有1/5的支持度P1(芭樂)和812(紅龍果)會同時被購買，而在已知木瓜或是芭樂被購買後，同時購買火龍果的信心度都是接近1/2，並且提升度也都有超過1，代表火龍果和木瓜是正相關；火龍果和芭樂也是正相關。

故可以嘗試在火龍果盛產時期(6月~12月這段區間)遇到爆量滯銷時，可以嘗試優先推銷以下客戶['0753','2101','0124','2158','0508','0994','0190',

'0101','2068','0178','0186','2018','2166','0968','0172','0932','0538','2149','0191','0128','0170','0159','0717','0960','0972','0916',

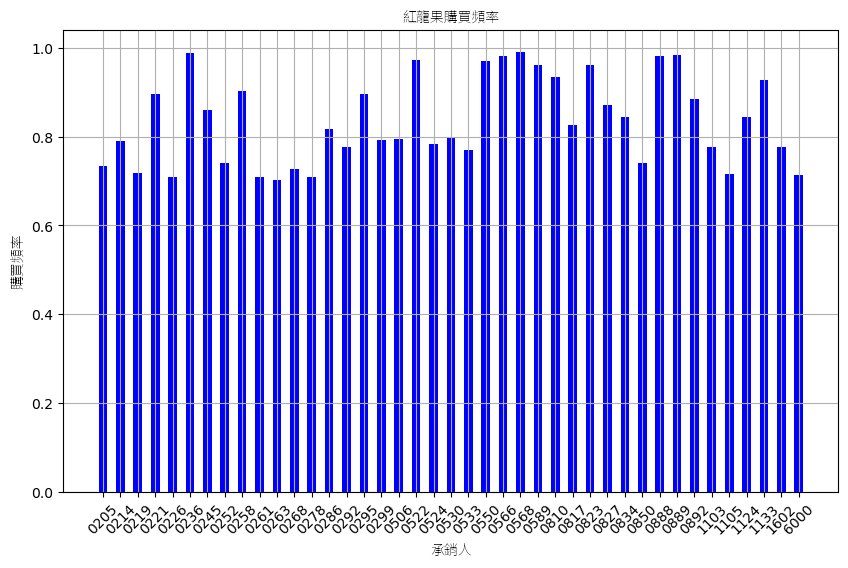
'0229','0105','0788','0757','0738','2062','2161','0772','2066','0132','0931','2053','2008','0996','0871','0134','2078','2060','0872',

'2165','2155','0171','0139','0944','0956','0576','0959','0585','2505','0167','2508','0598','0720','8888','0718','0125','0108','2017',

'0733','2162']。

**尋找舊客群中購買頻率較高的客人：**

透過了解舊有承銷人的購買頻率，可以在火龍果滯銷且購買頻率較高的舊客人還未購買時推銷其採買，成功機會會比較高。

****

可以發現['0205','0214','0219','0221','0226','0236','0245','0252',

'0258','0261','0263','0268','0278','0286','0292','0295','0299','0506','0522','0524','0530','0533','0550','0566','0568','0589','0810',

'0817','0823','0827','0834','0850','0888','0889','0892','1103','1105','1124','1133','1602','6000']這些承銷人較常購買紅龍果，頻率達0.7以上。

而這幾位承銷人的購買價位分別為：

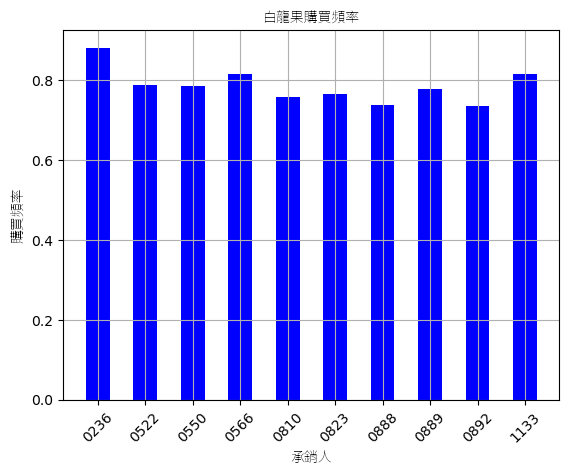
高價位：['0530','0278','0589','0827','0834','0550','0533','0214',

'0522','0295','0892','0823','0286','0263','1105','0258','1124','1133','0221','0219','0810']

中價位：['1103','0889','0888','0850','1602','0205','0817','0568',

'0566','0524','0506','0299','0292','0268','0261','0245','0236','0226','6000']

低價位：['0252']



['0236','0522','0550','0566','0810','0823','0888','0889','0892',

'1133']這些承銷人較常購買白龍果，頻率達0.7以上。

而這幾位承銷人的購買價位分別為：

高價位：['0522','0550','0810','0823','1133']

中價位：['0236','0566','0888','0889','0892']

低價位：[]