參賽內容

競賽平台：kaggle

競賽名稱：Predict Future Sales

解決方案：

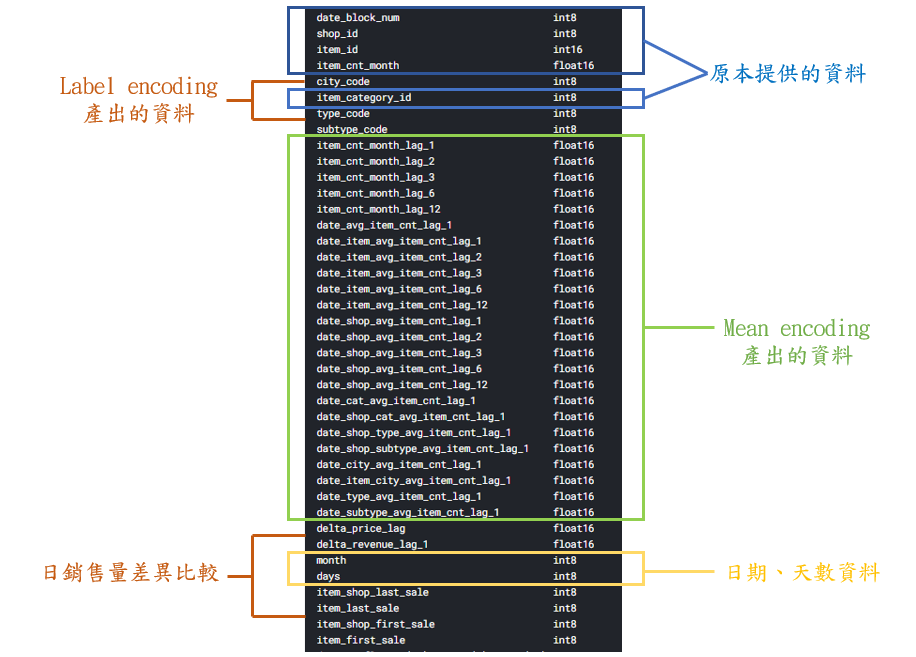
* 資料前處理

1. 觀察資料數值分佈，去除離峰值。
2. 觀察資料欄位。若發現NA，則根據商店、商品以及日銷售量的欄位取平均值來填補。
3. 將店名相同但商店id不同的資料統一。
4. 根據店家名稱，分析商店所在城市。
5. 根據商品項目，分析主要類型以及次要類型。
6. 對城市、商品主要類型和次要類型做label encoding。
7. 根據商店和商品資料計算商品的月銷售量，並將月銷售量小於0的欄位補0，大於20的欄位補20。(比賽有特別說明月銷售量只會介於0到20之間)
8. 透過mean encoding的方法增加資料欄位。針對日銷售量、商品id、商店id、商品項目id、商品主要類型、商品次要類型的欄位以不同方式組合做mean encoding。
9. 計算商品日銷售額和同商品平均銷售額之間的差異。
10. 計算商店日銷售額和同商店平均銷售額之間的差異。
11. 計算每間商店、每樣商品和前一次售出日日銷售量之間的差異。
12. 計算每樣商品和前一次售出日日銷售量之間的差異。
13. 計算每間商店、每樣商品和售出最少日日銷售量之間的差異。
14. 計算每樣商品和售出最少日日銷售量之間的差異。
15. 增加月份以及當月天數的欄位。

總體來說，資料前處理部分有去除離峰值、填補空缺值、針對有字串的欄位做label encoding、透過mean encoding的方法增加新欄位、以不同方法計算日銷售額的差異以及增加日期天數的欄位。

除此之外，將34個月的資料內容切割成前32個月訓練集，第33個月驗證集和第34個月測試集。由於前3個月在產生新欄位時，欄位內容無可追朔性，所以刪除前3個月的資料。

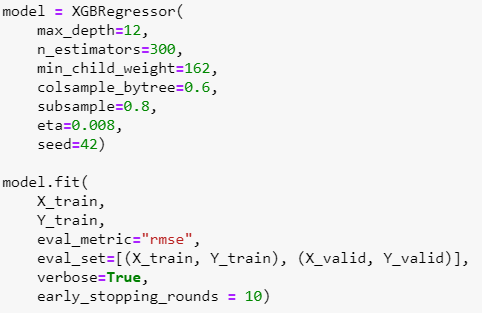
下圖為所有輸入模型訓練的資料，總共40個欄位。



* 模型運算

經過多種不同模型的測試，我們最終選擇Xgboost作為最終的訓練模型。而模型參數的選擇上透過Optuna框架選擇最佳的參數組合。

下圖為我們最終參數的選擇。



參賽名次:

於12726名對手中排名第1576名，成績為0.89647。

