資料科學 期末報告
第十組

組員:

資科三 鄭宛薰 108703003
資科三 林藝潔 108703005
資科三 郭沛澐 108703006
統計三 江宏繹 108304016
企管四 劉柏毅 106305022

dataset

- 五個table
  - 較舊的交易紀錄
  - 較新的交易紀錄
  - 商品種類
  - train (201917個卡號)
  - test (123623個卡號)

train (201917個卡號)

train.csv	
Columns	**Description**
card_id	Unique card identifier
first_active_month	'YYYY-MM', month of first purchase
feature_1	Anonymized card categorical feature
feature_2	Anonymized card categorical feature
feature_3	Anonymized card categorical feature
target	Loyalty numerical score calculated 2 months after historical and evaluation period

含card id、target

較舊的交易紀錄

historical_transactions.csv	
Columns	**Description**
card_id	Card identifier
month_lag	month lag to reference date
purchase_date	Purchase date
authorized_flag	Y' if approved, 'N' if denied
category_3	anonymized category
installments	number of installments of purchase
category_1	anonymized category
merchant_category_id	Merchant category identifier (anonymized)
subsector_id	Merchant category group identifier (anonymized)
merchant_id	Merchant identifier (anonymized)
purchase_amount	Normalized purchase amount
city_id	City identifier (anonymized)
state_id	State identifier (anonymized)
category_2	anonymized category

- 29m筆交易紀錄
  - 325540張卡
  - 1Jan17-1Mar18
  - 共含326k種商品
  - card id
  - merchant id

較新的交易紀錄

new_merchant_period.csv	
Columns	**Description**
card_id	Card identifier
month_lag	month lag to reference date
purchase_date	Purchase date
authorized_flag	Y' if approved, 'N' if denied
category_3	anonymized category
installments	number of installments of purchase
category_1	anonymized category
merchant_category_id	Merchant category identifier (anonymized)
subsector_id	Merchant category group identifier (anonymized)
merchant_id	Merchant identifier (anonymized)
purchase_amount	Normalized purchase amount
city_id	City identifier (anonymized)
state_id	State identifier (anonymized)
category_2	anonymized category

- 1.96m筆交易紀錄
  - 290001張卡
  - 1Mar17-1May18
  - 共含226k種商品
  - card id
  - merchant id

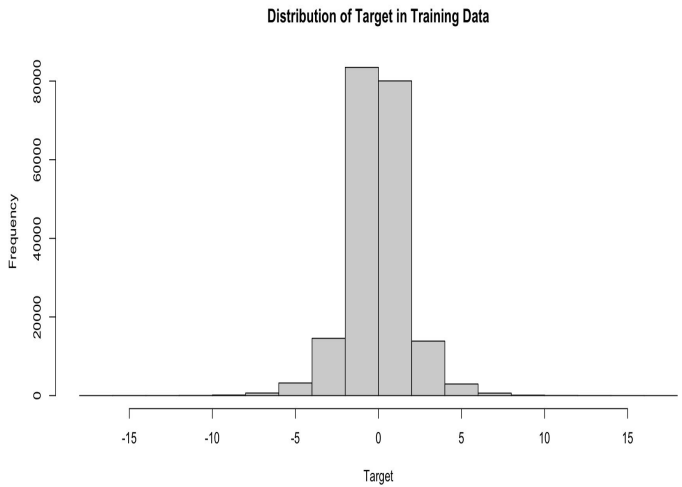
商品種類

merchants.csv

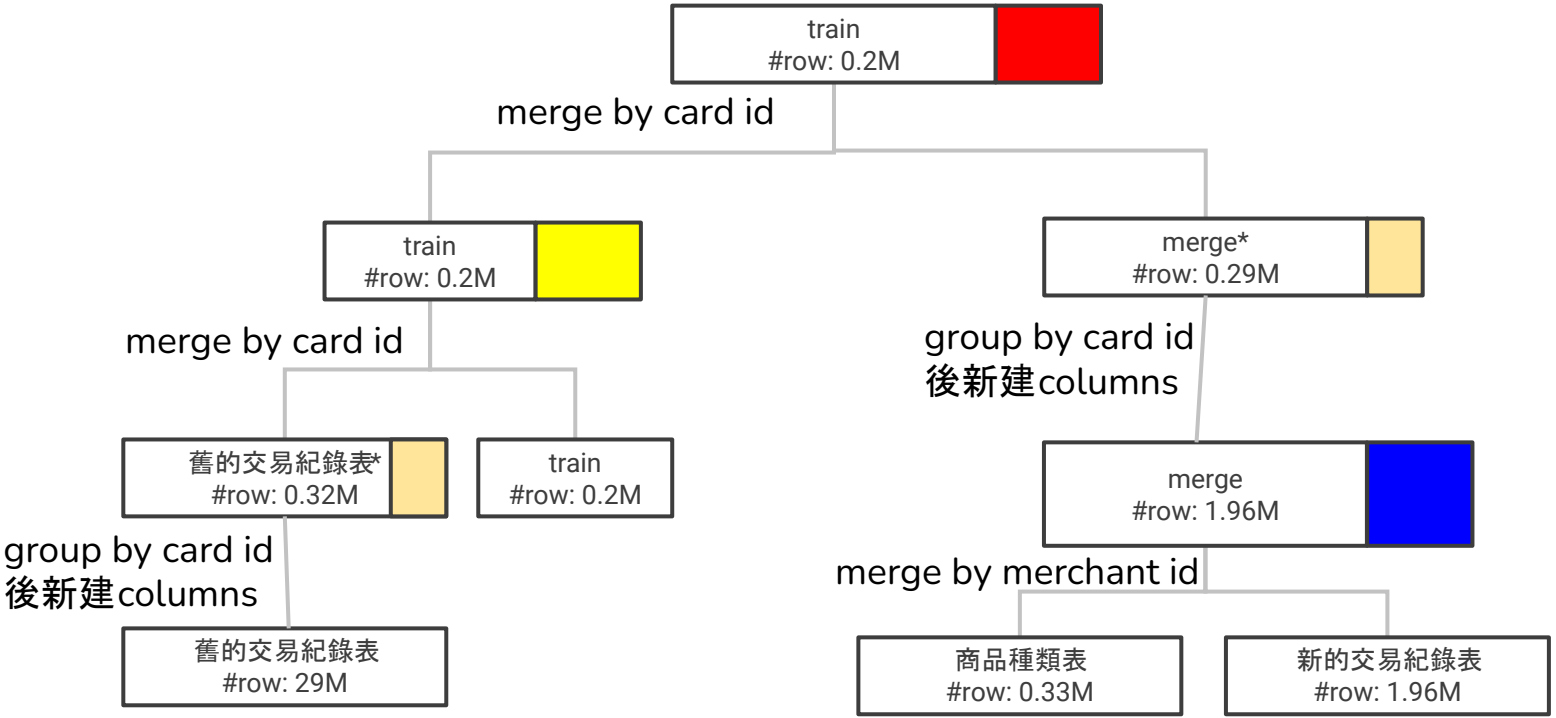
Columns	Description
merchant_id	Unique merchant identifier
merchant_group_id	Merchant group (anonymized)
merchant_category_id	Unique identifier for merchant category (anonymized)
subsector_id	Merchant category group (anonymized)
numerical_1	anonymized measure
numerical_2	anonymized measure
category_1	anonymized category
most_recent_sales_range	Range of revenue (monetary units) in last active month --> A > B > C > D > E
most_recent_purchases_range	Range of quantity of transactions in last active month --> A > B > C > D > E
avg_sales_lag3	Monthly average of revenue in last 3 months divided by revenue in last active month
avg_purchases_lag3	Monthly average of transactions in last 3 months divided by transactions in last active month
active_months_lag3	Quantity of active months within last 3 months
avg_sales_lag6	Monthly average of revenue in last 6 months divided by revenue in last active month
avg_purchases_lag6	Monthly average of transactions in last 6 months divided by transactions in last active month
active_months_lag6	Quantity of active months within last 6 months
avg_sales_lag12	Monthly average of revenue in last 12 months divided by revenue in last active month
avg_purchases_lag12	Monthly average of transactions in last 12 months divided by transactions in last active month
active_months_lag12	Quantity of active months within last 12 months
category_4	anonymized category
city_id	City identifier (anonymized)
state_id	State identifier (anonymized)
category_2	anonymized category

- 334633種商品 merchant id

預測值(target)的分布



- 我們發現train資料集裡面的target column的分布大多介於0上下，因此取以e為底的指數後exp{target}其大多會呈現介於1上下，因此推斷其為某兩個數相除之比例，而那兩個數就由特徵工程創造出來的新特徵去發掘!


train
#row: 0.2M
merge by card id
train
#row: 0.2M
merge*
#row: 0.29M
merge by card id
group by card id
後新建columns
舊的交易紀錄表*
#row: 0.32M
train
#row: 0.2M
merge
#row: 1.96M
group by card id
後新建columns
merge by merchant id
舊的交易紀錄表
#row: 29M
商品種類表
#row: 0.33M
新的交易紀錄表
#row: 1.96M

一些新建的columns

建了84個columns

舊的交易紀錄-日期相關
舊的交易紀錄-消費金額相關
舊的交易紀錄-其他

新的交易紀錄-日期相關
新的交易紀錄-消費金額相關
新的交易紀錄-其他

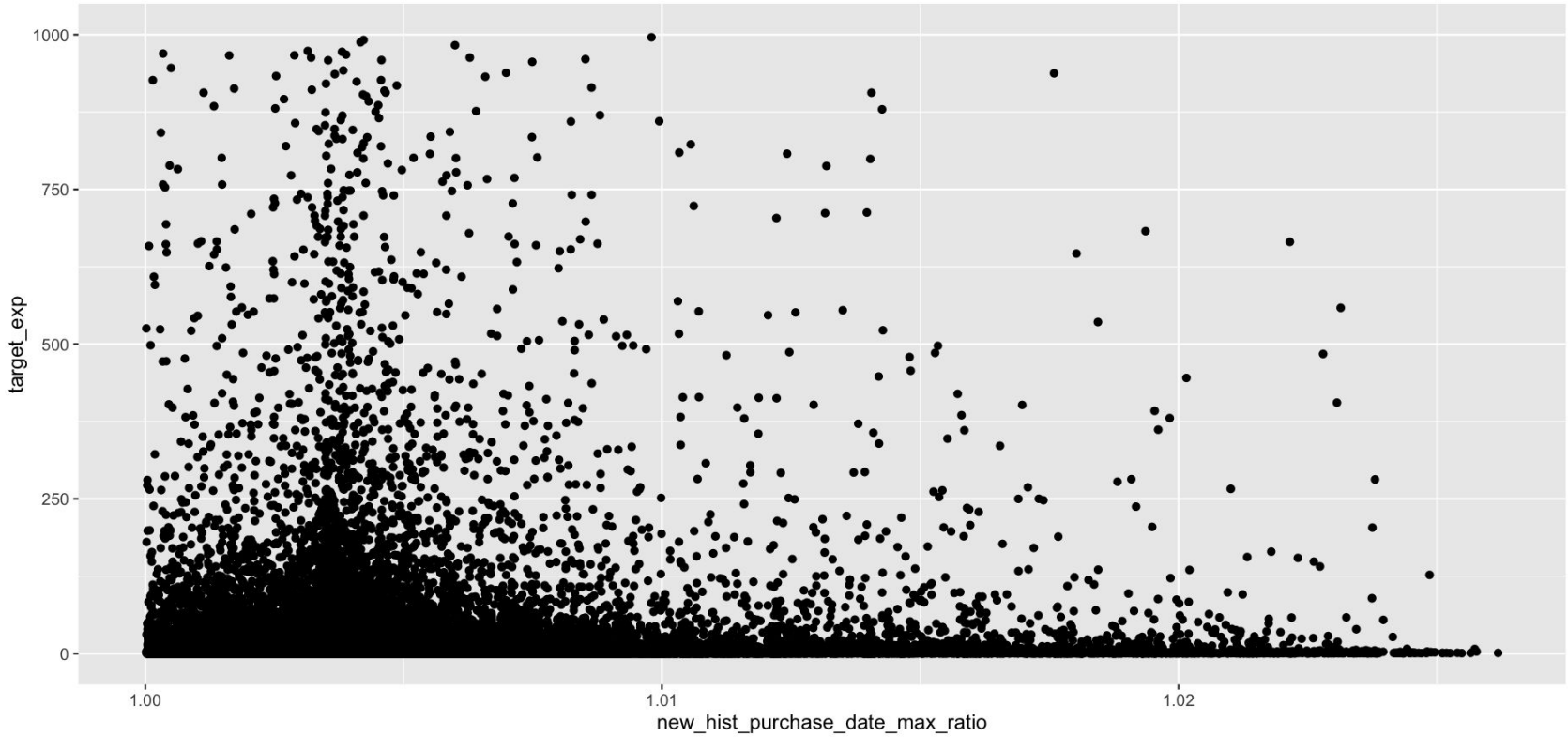
舊新相除算比例

基本款
hist_purchase_date_month_nunique

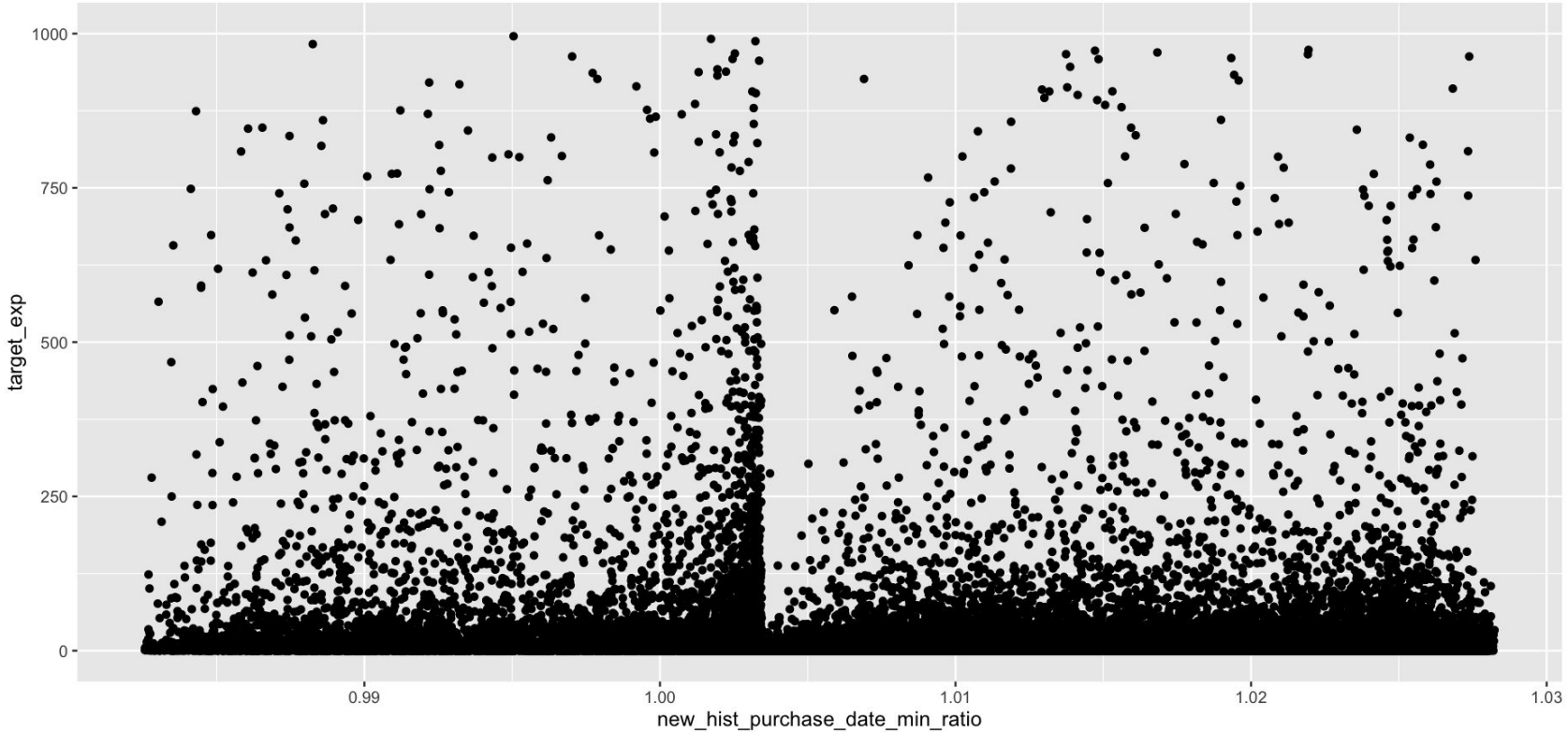
中階款
new_merchant_merchants_most_recent_sales_range_mean_max

進階款
new_hist_card_id_size_ratio

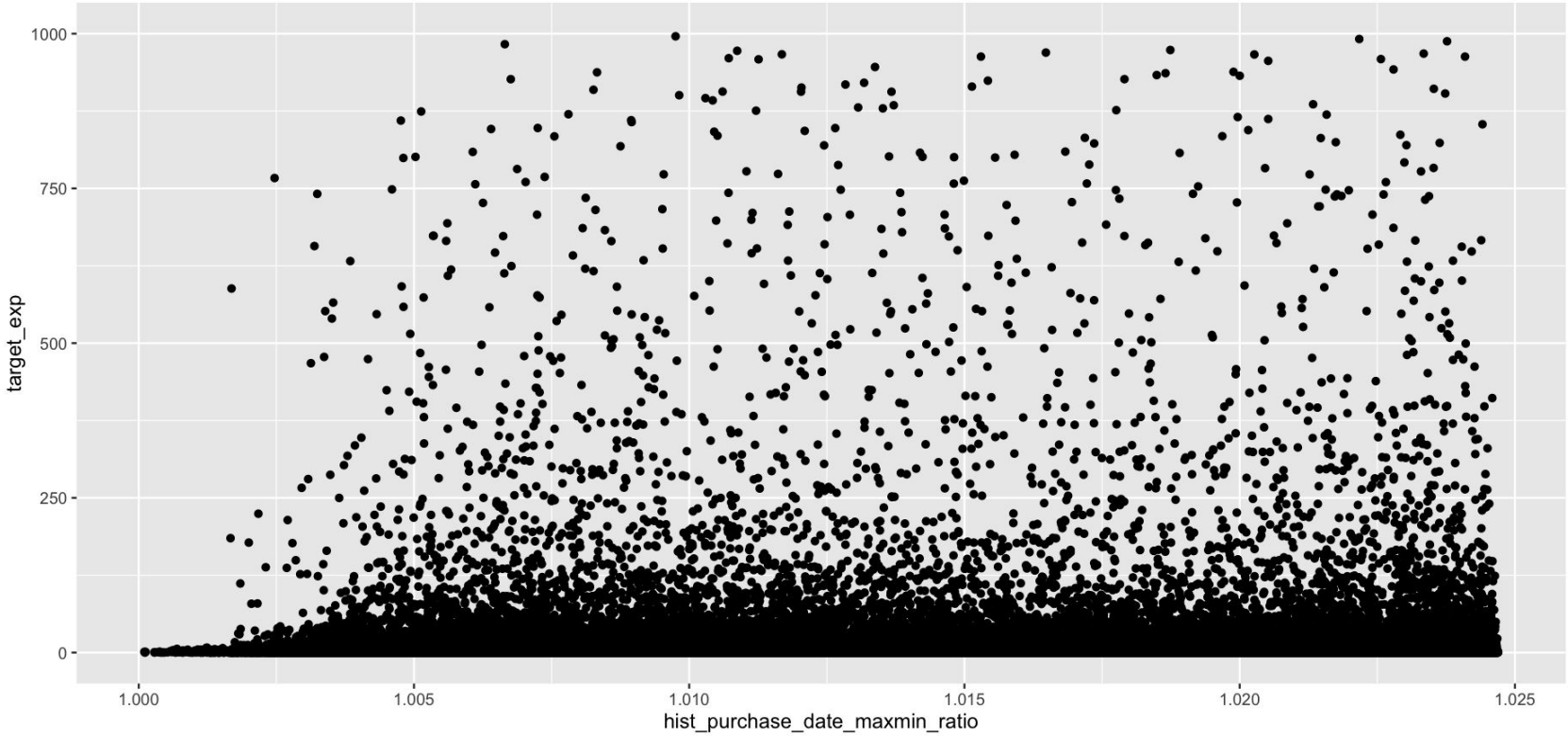
預測值(target)的分布



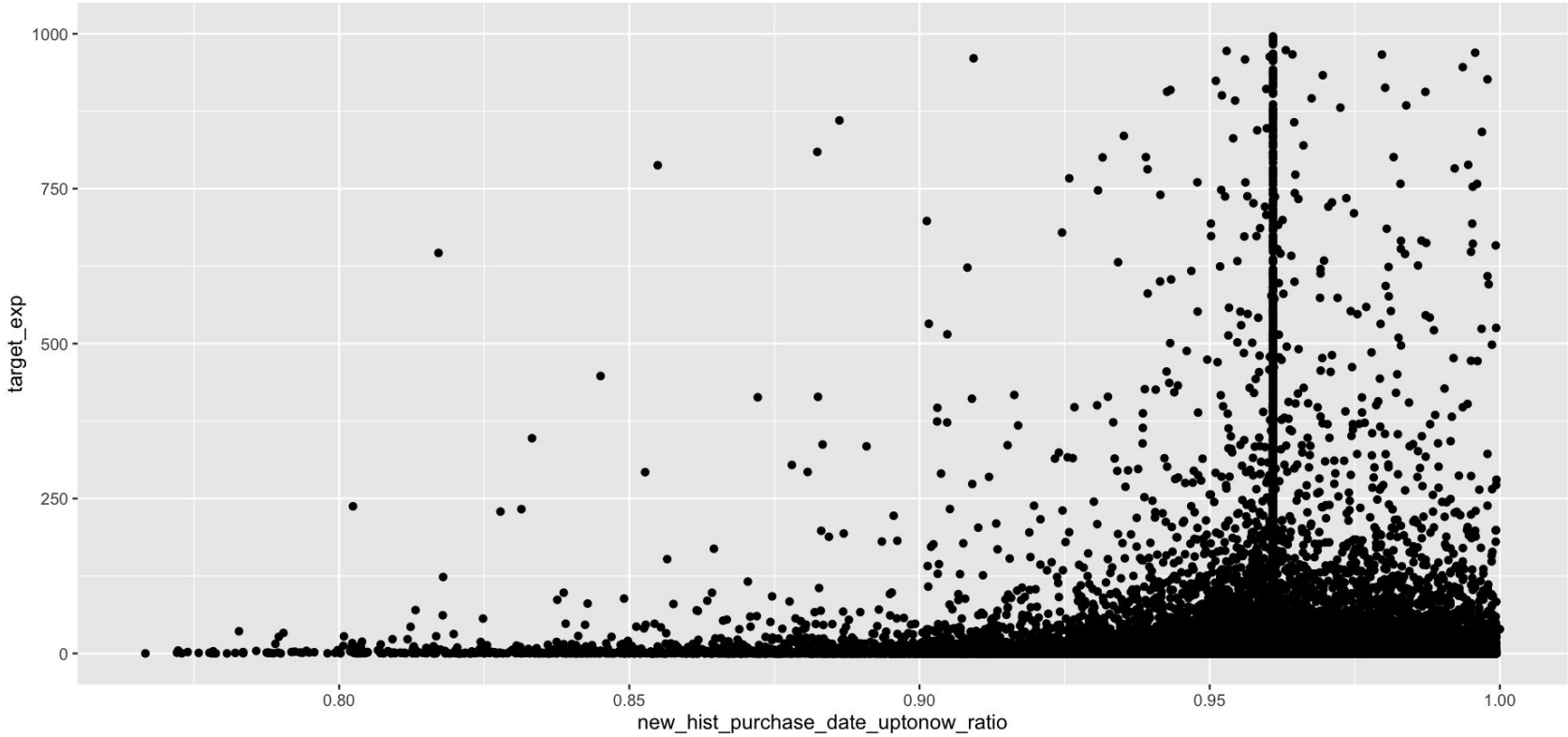
預測值(target)的分布



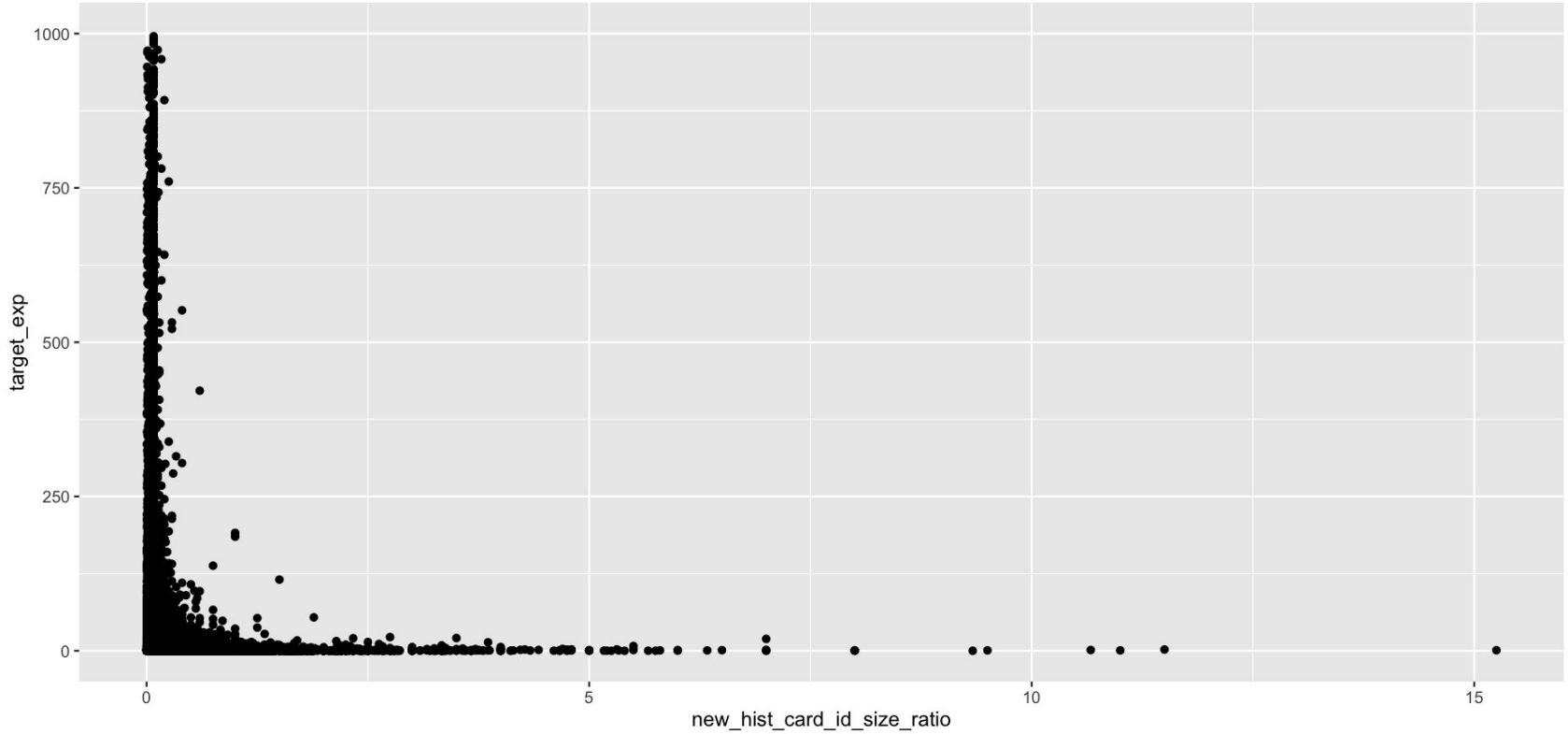
預測值(target)的分布



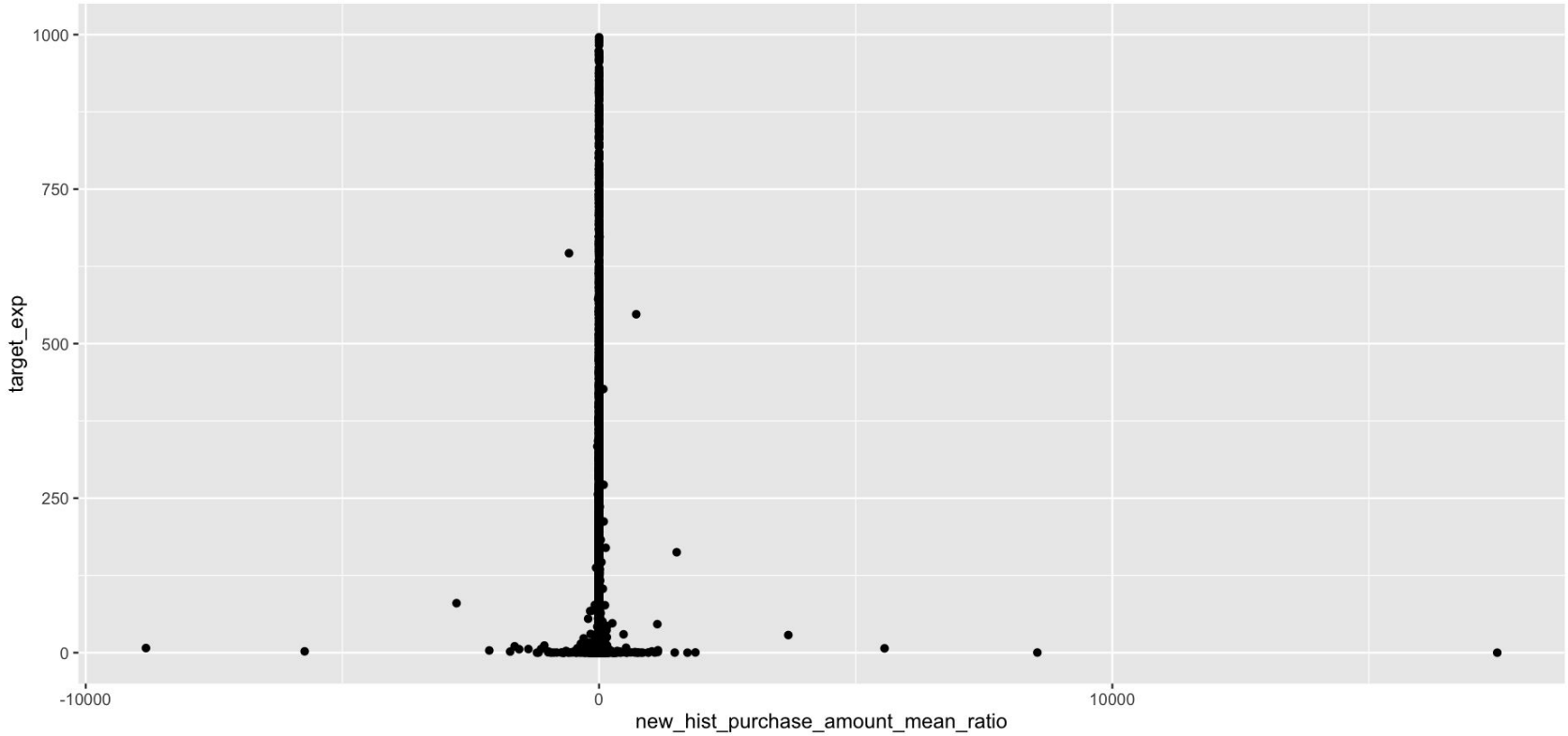
預測值(target)的分布



預測值(target)的分布



預測值(target)的分布



Shiny APP-Data Analysis

Modeling

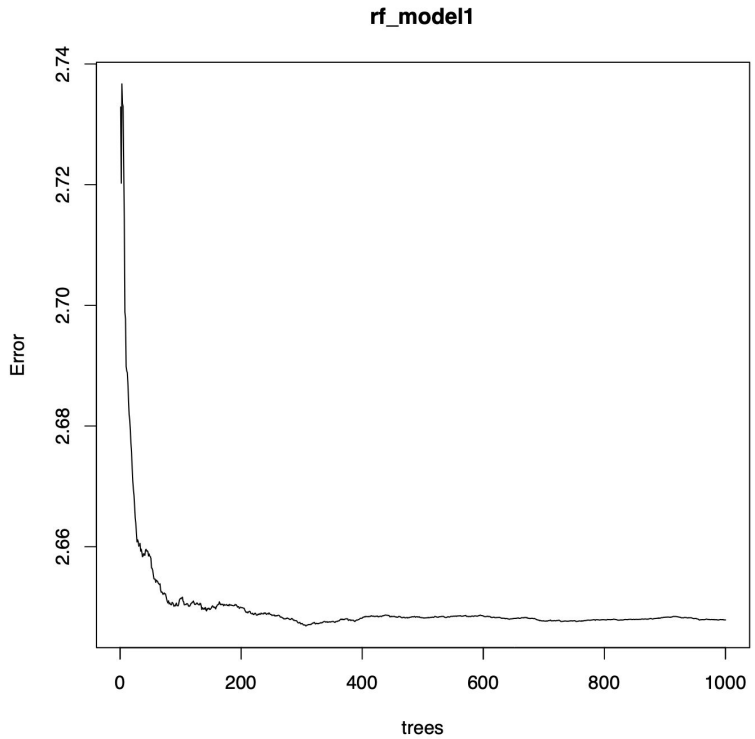
- 最終使用資料集大小
  - Training Set: 197710 * 85 (83 feautures + 1 target + 1 exp(target))
  - Testing Set: 123623 * 83

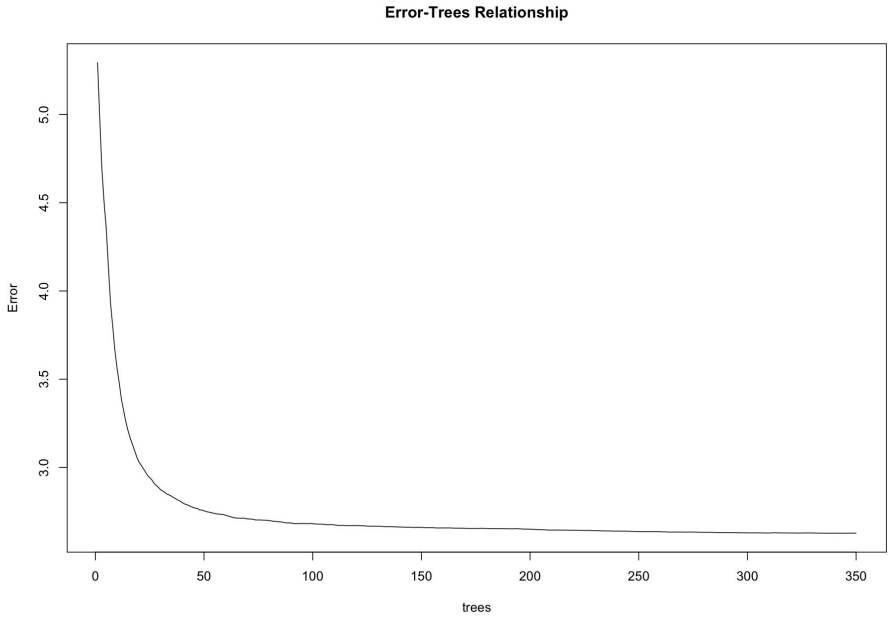
Modeling

- 使用Random forest
  - 隨機選取50000筆資料，隨機分8成為training data，剩下為testing data
  - 套用RandomForest()
  - 建構五個model
    - 差別：選擇不同seed
    - 去掉最大和最小的預測值，以剩下三個預測數的平均，作為最終的預測值

Modeling

- 調整：ntree
  - plot(rfmodel)

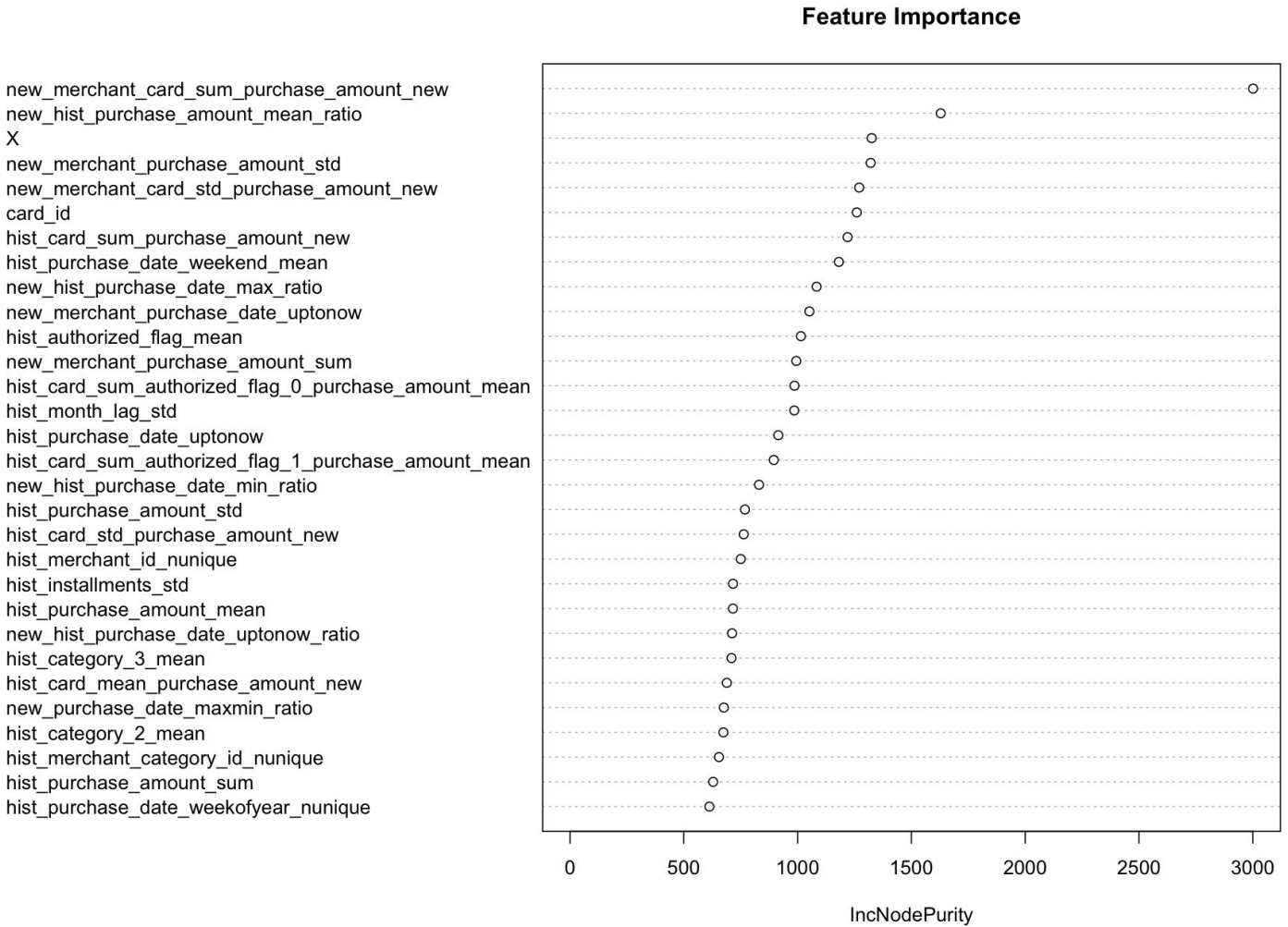




Modeling

- 調整
  - mtry:
    - 每次在決定切割變數時，所隨機抽樣的潛在變數清單數量
    - tuneRF()
  - maxnodes
    - 內部節點最大個數值
    - ？

```
mtry = 28  OOB error = 3.630661
Searching left ...
mtry = 14       OOB error = 3.716262
-0.02357708 0.05
Searching right ...
mtry = 56       OOB error = 3.634553
-0.00107188 0.05
   mtry OOBError
14   14 3.716262
28   28 3.630661
56   56 3.634553
[1] 0.06454707
```


Feature Importance
new_merchant_card_sum_purchase_amount_new
new_hist_purchase_amount_mean_ratio
X
new_merchant_purchase_amount_std
new_merchant_card_std_purchase_amount_new
card_id
hist_card_sum_purchase_amount_new
hist_purchase_date_weekend_mean
new_hist_purchase_date_max_ratio
new_merchant_purchase_date_uptonow
hist_authorized_flag_mean
new_merchant_purchase_amount_sum
hist_card_sum_authorized_flag_0_purchase_amount_mean
hist_month_lag_std
hist_purchase_date_uptonow
hist_card_sum_authorized_flag_1_purchase_amount_mean
new_hist_purchase_date_min_ratio
hist_purchase_amount_std
hist_card_std_purchase_amount_new
hist_merchant_id_nunique
hist_installments_std
hist_purchase_amount_mean
new_hist_purchase_date_uptonow_ratio
hist_category_3_mean
hist_card_mean_purchase_amount_new
new_purchase_date_maxmin_ratio
hist_category_2_mean
hist_merchant_category_id_nunique
hist_purchase_amount_sum
hist_purchase_date_weekofyear_nunique
0
500
1000
1500
2000
2500
3000
IncNodePurity

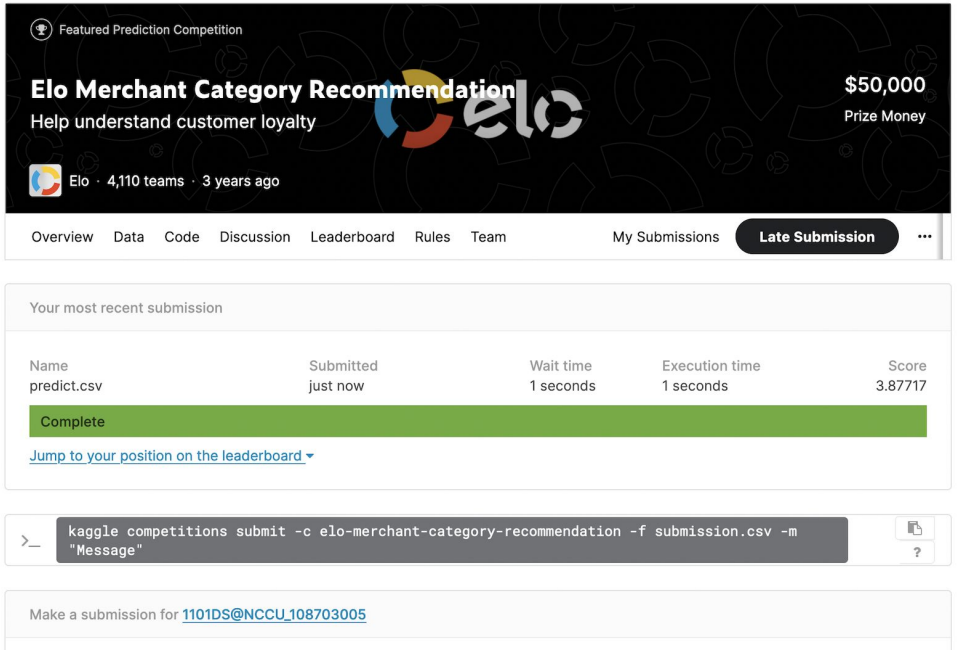
Output

- Performance
  - 計算MSE、RMSE、R-squared

model	MSE	RMSE	R_squared
model1	2.50392833755562	1.58238059187909	0.145607831457867
model2	2.53340248087143	1.59166657339766	0.141162036549647
model3	2.53713669648272	1.59283919354175	0.142224570445541
model4	2.52434626642342	1.58881914213778	0.149136665173404
model5	2.54223940374232	1.59444015370359	0.140129245478388

Output

- kaggle評測:
  - kaggle提供的null model:3.87852
  - 若用未preprocessing的資料:4.58
- 可改進處:
  - nodesize
  - 用所有資料



Shiny APP-Model Analysis

困難-Anonymized Feature

問題：許多資料的數值和類別是匿名的，因此無法用常理判斷處理方式及是否為合適特徵。

解決：將許多特徵的分佈畫出，以及將商品特徵與消費者特徵進行比較。例如：在ategory_3中，值為A、B、C。根據分析我們發現在target的均值上，A>B>C，因此將特徵轉換為A:2、B:1、C:0

困難-Feature Enigeering

問題:資料龐大但資訊分散，需要進行多層處理。主表上只有3個特徵，最後我們共造出84個特徵

解決:先將商品的主表中重複和缺值的商品進行處理，接著分別針對過去和近期資料groupby消費者進行整合，最後將過去和近期資料進行比較

困難-Shiny APP & Preprocessing

問題:input 的檔案很大，導致部署到ShinyIO時，會因為out of memory無法正常運作。

Preprocessing處理資料時也會因為檔案太大在跑不動function

modeling也要跑很久

解決:將 input 檔案處理過後再上傳，只留下處理後的資料，去除原始資料

把function拆開一行一行執行

降低資料量，隨機選取

QA時間

謝謝大家聆聽！寒假愉快～