



國立臺灣科技大學  
資訊工程系

---

## 碩士學位論文

巨量資料探勘框架：基於極限梯度提升之預測免費手機  
遊戲中潛在新進付費玩家

Big Data Mining Framework: Predicting Potential New  
Paying Player in Mobile Free-to-Play Games Based on  
Extreme Gradient Boosting

研 究 生：廖宣瑋

學 號：M10715084

指導教授：戴文凱博士

中華民國一〇九年七月二十一日



M10715084



## 碩士學位論文指導教授推薦書

本校 資訊工程系 廖宣瑋(LIAO, HSUAN-WEI) 君

所提之論文：

巨量資料探勘框架：基於極限梯度提升之預測免費手機遊戲中潛  
在新進付費玩家

係由本人指導撰述，同意提付審查。

指導教授：戴文凱

指導教授

109 年 07 月 21 日



# 碩士學位考試委員審定書



M10715084

指導教授：戴文凱

本校 資訊工程系 廖宣瑋 君

所提之論文：

巨量資料探勘框架：基於極限梯度提升之預測免費手機遊戲中潛  
在新進付費玩家

經本委員會審定通過，特此證明。

學校考試委員會

委

員：

張國清  
戴文凱

指導教授：

戴文凱

學程主任：

系(學程)主任、所長：

戴文凱

中華民國 109 年 7 月 21 日

# 中 文 摘 要

目前市面上之手機遊戲多以免費遊玩商業模式 (Free-to-Play, F2P) 為主，使得遊戲內購買 (In-App Purchase, IAP) 顯得越來越重要，已然成為遊戲開發商營運之重點，為了能夠推出成功吸引各式玩家的精準行銷，需要資料分析團隊針對付費玩家進行研究，並且希望能夠在新進玩家族群中，成功預測出潛在付費玩家，以利提升 IAP 的意願，因此，如何在付費玩家資料中，有效探勘出資料特徵並透過機器學習進行預測，則為此次研究的目標。

本論文對此議題提出一巨量資料探勘框架，將需先將資料進行前處理以及預測前之資料分析，隨後訓練機器學習與其最佳化處理，最後再依預測之結果導入資料特徵重要性分析之中，完成整體預測與分析之工作，此框架將由四大階段組成：(1) 資料前處理階段、(2) 資料分析階段、(3) 機器學習階段及 (4) 預測結果分析階段。

根據實驗結果，藉由我們提出的巨量資料探勘框架，利用無價值玩家觀察期清理了無價值的資料，並藉由付費玩家定義期準備了付費玩家與非付費玩家目標值，利用資料特徵探勘期探勘出了有價值的玩家遊戲行為軌跡。透過探索性資料分析 (Exploratory Data Analysis, EDA) 找出不合理資料特徵與高資訊量資料特徵，推測出有價值的資料特徵。能夠經由學習模型之預測，預測出潛在之新進付費玩家，並依其預測結果，分析資料特徵重要性，了解到玩家消費原因與遊戲之連動性。整體來說，該框架將能使得預測付費玩家之時間成本與人力成本有效降低，並得到對於行銷有利的資訊。

關鍵字：付費預測、免費遊玩遊戲、巨量資料、資料探勘、機器學習、極限梯度提升

# ABSTRACT

In the last few years, most mobile games on the market are dominated by Free-to-Play ( F2P ) business models, which makes In-App Purchase ( IAP ) more and more important, and has become the focus of game developer operations. In order to be able to launch accurate marketing that successfully attracts all types of players, it is necessary for the data analysis team to conduct research on paying players, and hope to be able to successfully predict potential paying players in the new player group, so as to improve the willingness of IAP. Among the payer data, effective mining of data features and prediction through machine learning are goals of this research.

This paper proposes a big data mining framework for this topic. It will need to process the data and data analysis before prediction, then train and optimize the model via machine learning algorithm. Finally, analyze the feature importance with the prediction results. This framework will be composed of four major stages: (1) data pre-processing stage, (2) data analysis stage, (3) machine learning stage, (4) feature importance analysis stage.

According to the experimental results, with the big data mining framework we proposed, the valueless data was cleaned up by the observation period of valueless players, and the target values of payer and non-payer were prepared through the definition period of payer. The valuable player game play record is prepared by the mining period of data mining. Through Exploratory Data Analysis ( EDA ) , unreasonable data features and high-information data features are found to infer valuable data features. Through the prediction of the learning model, it can predict potential new paying players, and analyze the importance of features according to the prediction results, understand the player's consumption reasons and the connection with game. Overall, the framework will effectively reduce the time cost and labor cost of predicting paying players, and obtain favorable information for marketing.

Keywords: Purchases Prediction, Free-to-Play, Big Data, Data Mining, Machine Learning, Extreme Gradient Boosting

## 誌

## 謝

在兩年的碩士生涯中，我的指導教授戴文凱博士給予了我許多的教導與鼓勵，積極帶領我參與各領域之計畫，適時補足我的不足，並給予指導；而在論文撰寫時，老師也是細心地叮囑我關於論文需注意之事項，並時刻追蹤進度，協助我完成本論文，再次誠心感謝戴老師的細心指導。

特別感謝張國清博士於本論文之研究階段時，給予許多的幫忙與建議，共同解析難題，並將其克服，進而提升實驗結果之成效，且不辭辛勞地前來學校與我討論，再次誠心感謝張博士的慷慨協助。

此外，十分感謝我的學長姐：益銓、竣生、秣安、國彥、奎谷、德潔、國軒、允斌與濬安；我的同學博安、岳儒、子樂、聖文、俊儒、承達與政一；我的學弟妹：增宇、維軒、孟傑、竹萱與珮如，以及其餘所有 GAMELab 中的成員，在我遇到困難時，給予相當大的幫忙，並與我共同討論，借助大家的支持繼續努力完成本論文，再次誠心感謝大家的陪伴與幫助；另外，祝福維軒能夠透過本論文，持續致力於資料科學之研究，並替 GAMELab 在此領域提供更多的知識與資源。

最後衷心感謝一路上支持與支助我求學的家人們以及我的女朋友：晏琦，因為大家提供於我非常多的鼓勵與溫暖的依靠，讓我在碩士生涯中能夠放心的進行研究與學習，順利完成碩士學位，期許能夠在將來繼續與大家共患難，誠心感謝。

# 目 錄

中文摘要 . . . . .	III
ABSTRACT . . . . .	IV
誌謝 . . . . .	V
目錄 . . . . .	VI
圖目錄 . . . . .	IX
表目錄 . . . . .	XI
符號說明 . . . . .	XII
1 緒論 . . . . .	1
1.1 研究背景與動機 . . . . .	1
1.2 研究目標 . . . . .	2
1.3 研究方法概述 . . . . .	2
1.4 研究貢獻 . . . . .	3
1.5 本論文之章節結構 . . . . .	3
2 文獻探討 . . . . .	5
2.1 Free-to-Play 類型遊戲興起 . . . . .	5
2.2 資料前處理 . . . . .	7
2.3 學習模型選擇 . . . . .	7
2.4 資料不平衡處理及其評估方式 . . . . .	9
3 研究方法 . . . . .	10
3.1 資料前處理階段 . . . . .	11
3.1.1 整合資料 . . . . .	11

3.1.2	清理資料 . . . . .	12
3.1.2.1	空缺值處理 . . . . .	13
3.1.2.2	無價值玩家資料處理 . . . . .	13
3.1.3	目標值準備 . . . . .	14
3.1.4	資料特徵探勘 . . . . .	16
3.2	資料分析階段 . . . . .	18
3.2.1	探索性資料分析 ( Exploratory Data Analysis, EDA ) . . . . .	18
3.2.1.1	不合理之資料特徵 . . . . .	19
3.2.1.2	高資訊量之資料特徵 . . . . .	20
3.3	機器學習階段 . . . . .	20
3.3.1	分割訓練與測試資料集 . . . . .	21
3.3.2	學習模型選擇 . . . . .	21
3.3.3	資料不平衡處理 . . . . .	22
3.3.4	搜尋最佳參數解 . . . . .	23
3.3.5	交叉驗證 ( Cross Validation ) . . . . .	24
3.3.6	評估驗證最佳模型 . . . . .	25
3.4	預測結果分析階段 . . . . .	26
3.4.1	資料特徵重要性分析 . . . . .	26
4	實驗結果與分析 . . . . .	28
4.1	實驗系統架構 . . . . .	28
4.2	資料前處理評估 . . . . .	29
4.2.1	清理資料評估 . . . . .	29



4.2.2	目標值與資料特徵評估 . . . . .	31
4.3	資料分析評估 . . . . .	35
4.3.1	探索性資料分析評估 . . . . .	35
4.4	機器學習評估 . . . . .	41
4.4.1	分割訓練與測試資料集評估 . . . . .	42
4.4.2	資料不平衡處理評估 . . . . .	43
4.4.3	最佳模型評估 . . . . .	48
4.5	預測結果分析評估 . . . . .	51
4.5.1	資料特徵重要性評估 . . . . .	51
5	結論與未來研究 . . . . .	54
5.1	結論 . . . . .	54
5.2	未來研究 . . . . .	54
	參考文獻 . . . . .	55

# 圖 目 錄

圖 1.1	近年來遊戲領域資料探勘研究數折線圖 . . . . .	1
圖 2.1	Bagging 方式建樹示意圖 . . . . .	8
圖 2.2	Boosting 方式建樹示意圖 . . . . .	8
圖 3.1	本論文之巨量資料探勘框架示意圖 . . . . .	10
圖 3.2	資料庫群內資料之示意圖 . . . . .	11
圖 3.3	依各項遊戲為整合目標之示意圖 . . . . .	11
圖 3.4	原始資料集示意圖 . . . . .	12
圖 3.5	空缺值處理示意圖 . . . . .	13
圖 3.6	判別有價值與無價值玩家之示意圖 . . . . .	14
圖 3.7	定義付費玩家與非付費玩家之示意圖 . . . . .	15
圖 3.8	付費玩家與非付費玩家范氏圖 . . . . .	15
圖 3.9	資料特徵探勘期示意圖 . . . . .	16
圖 3.10	探索性資料分析之使用圖表類型 . . . . .	18
圖 3.11	不合理之資料特徵示意圖 . . . . .	19
圖 3.12	高資訊量之資料特徵示意圖 . . . . .	20
圖 3.13	分割訓練與測試資料集示意圖 . . . . .	21
圖 3.14	RepeatedStratifiedKFold 示意圖 . . . . .	24
圖 4.1	玩家資料空缺值示意圖 . . . . .	29
圖 4.2	付費玩家之消費速度圖 . . . . .	31

圖 4.3	觀察設備所在地之付費玩家與非付費玩家數量長條圖 . . . . .	35
圖 4.4	觀察設備所在地之付費玩家比例長條圖 . . . . .	36
圖 4.5	觀察 GameTypeE 59 號遊戲之總贏遊戲次數長條圖 . . . . .	36
圖 4.6	觀察 GameTypeA 資料特徵散佈圖 . . . . .	37
圖 4.7	觀察 GameTypeA 資料特徵關聯散佈圖 . . . . .	38
圖 4.8	觀察 GameTypeD 65 號遊戲資料特徵關聯散佈圖 . . . . .	39
圖 4.9	觀察 GameTypeE 62 號遊戲之總贏遊戲次數長條圖 . . . . .	40
圖 4.10	觀察 GameTypeE 62 號遊戲之獲得遊戲貨幣 A 之總額散佈圖 . . .	41
圖 4.11	各週之新進玩家數圖 . . . . .	42
圖 4.12	ROC Curve 示意圖 . . . . .	44
圖 4.13	PR Curve 示意圖 . . . . .	44
圖 4.14	不平衡資料中 ROC Curve 失準示意圖 . . . . .	45
圖 4.15	不平衡資料處理前後比較之 ROC Curve 圖 . . . . .	46
圖 4.16	不平衡資料處理前後比較之 PR Curve 圖 . . . . .	47
圖 4.17	三種學習模型之 ROC Curve 比較圖 . . . . .	49
圖 4.18	三種學習模型之 PR Curve 比較圖 . . . . .	49
圖 4.19	Decision Tree 資料特徵重要性比較圖 . . . . .	51
圖 4.20	Random Forest 資料特徵重要性比較圖 . . . . .	52
圖 4.21	XGBoost 資料特徵重要性比較圖 . . . . .	52

# 表 目 錄

表 2.1	近年來機器學習應用於遊戲領域表 . . . . .	6
表 3.1	資料特徵種類表 . . . . .	17
表 3.2	學習模型參數調教表 . . . . .	23
表 3.3	學習模型參數說明表 . . . . .	23
表 3.4	$\beta$ 數值意義表 . . . . .	25
表 4.1	實驗系統架構之研究環境表 . . . . .	28
表 4.2	無價值玩家觀察表 . . . . .	30
表 4.3	有價值玩家觀察表 . . . . .	30
表 4.4	付費玩家及非付費玩家定義表 . . . . .	31
表 4.5	各遊戲行為軌跡探勘表 . . . . .	33
表 4.6	各遊戲行為軌跡資料特徵數表 . . . . .	34
表 4.7	資料特徵總數表 . . . . .	34
表 4.8	訓練與測試資料集玩家數表 . . . . .	42
表 4.9	Confusion Matrix . . . . .	43
表 4.10	最佳模型評估表 . . . . .	48
表 4.11	最佳模型參數解表 . . . . .	50
表 4.12	參數搜尋範圍表 . . . . .	50

## 符 號 說 明

$N$	無價值玩家觀察期
$M$	付費玩家定義期
$G$	資料特徵探勘期
$class\ 1$	付費玩家
$class\ 0$	非付費玩家
$N_1$	付費玩家樣本數
$N_0$	非付費玩家樣本數
$\lfloor \cdot \rfloor$	地板函數
$G(\cdot)$	<i>Gini Impurity</i>
$GI(\cdot)$	<i>Gini Importance</i>
$fi(\cdot)$	資料特徵重要性 ( <i>Feature Importance</i> )
$D_p$	父節點
$D_{left}$	左子節點
$D_{right}$	右子節點
$N_p$	父節點樣本數
$N_{left}$	左子節點樣本數
$N_{right}$	右子節點樣本數
$x$	欲求其重要性之資料特徵
$k$	節點分割時所用資料特徵為 $x$ 之所有節點
$l$	樹中所有節點
$t$	學習模型中的所有樹

# 第 1 章 緒論

## 1.1 研究背景與動機

近年來，遊戲玩家對於手機遊戲之投入顯得越來越重，尤其是於免付費類型 (Free-to-Play, F2P) 之手遊，而非傳統買斷制或月費制類型遊戲。此一商業模式之改變，將使得遊戲商更加注重於遊戲內購買 (In-App Purchases, IAP)，並且在 F2P 的模式之下，遊戲商之營收有非常顯著的成長，而玩家更加傾向於遊玩 F2P 類型遊戲，例如：「魔獸世界 (World of Warcraft)」由於其月費制的商業模式，在 2010 年至 2013 年間，因其玩家流失至 F2P 類型同種遊戲，玩家數大量下滑了 30 % 之多 [1]；「絕地要塞 2 (Team Fortress 2)」於 2007 年為發售買斷制型式遊戲，在 2012 年改為 F2P 商業模式，並推出 IAP 供玩家消費，此一轉變大幅提高遊戲營收達 12 倍之多 [2]。

因此，如何在 F2P 類型遊戲中，提高玩家之付費意願，使其透過 IAP 方式增加遊戲商之營收，顯得格外重要。透過上述情境發想，若能提出一巨量資料探勘框架於遊戲領域，將可以大幅提升預測付費玩家之效率，並且利用預測結果進行玩家消費引導以及了解付費玩家之消費原因與動機。另外，將有助於增加遊戲領域巨量資料探勘之資源，因近幾年來巨量資料研究於遊戲領域呈現下滑狀況 [3]，如圖 1.1，於 2013 年開始便無顯著成長。

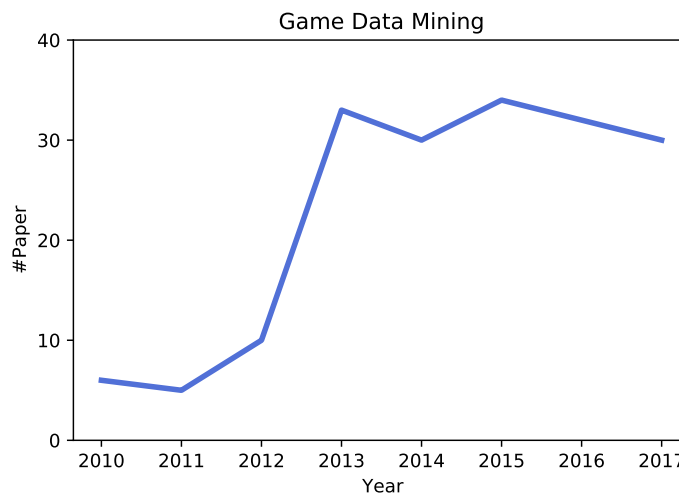


圖 1.1: 近年來遊戲領域資料探勘研究數折線圖 (此圖出自 [3])

## 1.2 研究目標

由於提前了解到玩家之付費意願，將有助於遊戲商進行精準行銷於正確的玩家上，使其提升 IAP 之意願，讓遊戲商可於 F2P 類型遊戲中獲得良好的營收，又因玩家於進入遊戲初期，將會擁有較高的付費意願。因此，若能藉由一巨量資料探勘框架，來協助遊戲商預測潛在之新進付費玩家，將可減少大量人力成本與時間成本，對於遊戲商有很大的幫助。

本論文的研究目標為預測潛在之新進付費玩家與了解玩家之付費原因。我們將提出一巨量資料探勘框架，包含對於資料集之前處理，並藉由資料分析方法來探索資料之特性，隨後採用機器學習之分類預測，來預測潛在之新進付費玩家，最後依其機器學習預測結果，分析各款遊戲中的突出性，來進行玩家付費原因的解釋說明。

## 1.3 研究方法概述

本論文對此議題提出一巨量資料探勘框架：此框架將由四大階段組成，(1) 資料前處理階段：首先將從資料庫群中整合所有所需資料，並對其採用直接刪除空缺值樣本方式進行清理，並透過無價值玩家觀察期，刪除無價值玩家資料，以獲取有價值之原始資料，再著手準備目標值與探勘資料特徵，以利後續分析及機器學習使用；(2) 資料分析階段：使用前階段產出之有價值原始資料並透過長條圖與散佈圖來觀察資料特性，進行探索性資料分析 (Exploratory Data Analysis, EDA)，藉由付費玩家與非付費玩家資料分佈來檢查是否有不合適之資料特徵以及觀察資料特徵是否可以提供給學習模型較多之資訊；(3) 機器學習階段：首先將有價值原始資料集進行分割為訓練及測試集，隨後針對訓練集進行少數群樣本權重值放大以處理不平衡資料與交叉驗證搭配參數表，以獲得最佳模型，其中學習模型選用 Decision Tree、Random Forest 及 Extreme Gradient Boosting，最後藉由測試集來驗證評估最佳模型，產出預測結果；(4) 預測結果分析階段：使用前階段產出之預測結果進行資料特徵重要性分析，透過計算各資料特徵於各學習模型中之 *Gini Importance*，以利更加了解及解釋資料特徵與遊戲所提供之體驗綜合評估。

在方法驗證上，本論文將藉由 *Confusion Matrix* 所延伸之 ROC Curve [4] 與 PR Curve [5] 來協助驗證學習模型之優劣，並同時利用 *Weighted  $F_{\beta\alpha}$  - Score* [6] 來選出最佳模型與最佳參數解，隨後計算 *Feature Importance* 於各資料特徵中，

以了解到何者於學習模型中貢獻了最多的資訊量，以利學習模型進行訓練與分類。

## 1.4 研究貢獻

本論文之研究貢獻為：

1. 提出一無價值玩家觀察期，可將無價值之玩家資料給予刪除，以利學習模型更加著重於有價值的資訊上。
2. 提出一付費玩家定義期，可將目標值之準備侷限在新進之玩家上。
3. 提出一資料特徵探勘期，配合無價值玩家觀察期，以利學習模型更加著重於新進玩家的資訊上。
4. 於資料集中進行資料特徵之探勘，藉由不同種類與面向之方式，挑選出適合用來呈現付費玩家的資料。
5. 整理出適合於不平衡資料集中的評估值方式，將對於學習模型之預測結果提供合理的評估，進而進行比較。
6. 提出一權重值，用於改善資料不平衡之處理，透過將少數群進行放大樣本權重值，使得學習模型更加著重於少數群之資訊。
7. 整理出資料特徵重要性之計算，以利分析資料特徵的突出性與其貢獻的資訊量。
8. 提出一巨量資料探勘框架，有效的減少研究時間，並針對各議題進行完善的規劃。

## 1.5 本論文之章節結構

本論文第 2 章為文獻探討，針對 Free-to-Play 類型遊戲興起介紹，並探討關於資料前處理、學習模型選擇、資料不平衡處理以及其評估方式的相關文獻。第 3 章為研究方法，詳述講述各階段之研究流程，並分為四節來介紹資料前處



理、資料分析、機器學習及預測結果分析。第 4 章為實驗結果與分析，分為五節：第 4.1 節實驗系統架構；第 4.2 節資料前處理評估；第 4.3 節資料分析評估；第 4.4 節機器學習評估；第 4.5 節預測結果分析評估。第 5 章為結論與未來研究，總結本論文提出的方法與實驗結果，並討論未來的研究方向。

## 第 2 章 文獻探討

本章節針對 Free-to-Play 類型遊戲興起介紹，並探討關於資料前處理、學習模型選擇、資料不平衡處理以及其評估方式的相關文獻。

### 2.1 Free-to-Play 類型遊戲興起

Free-to-Play (F2P) 類型遊戲，是一種玩家無需支付任何費用，即可遊玩該遊戲之大部分內容，與付費型遊戲（買斷制或月費制）形成對比。而在 F2P 類型遊戲中，針對遊戲商還是可以藉由遊戲內購買（In-App Purchase, IAP）或遊戲內置入廣告方式來賺取營收 [7]。近幾年內，遊戲商皆轉以開發 F2P 類型遊戲為主，因其類型所帶來之營收，已遠大於付費型遊戲 [3]。另外，還因為 F2P 類型遊戲能夠有效的讓玩家流失量降低，透過其無需支付任何費用就能遊玩遊戲的特性，使玩家進入遊戲的門檻大為降低 [8]。

因此，如何讓玩家於 F2P 類型遊戲內消費成為了值得探討的議題。近幾年來，許多研究團隊使用機器學習之方式，來進行遊戲領域資料科學方面的研究，大多是以預測玩家是否流失為主 [3] [8] [9]，如表 2.1。其中，僅有少數研究團隊進行預測玩家是否會消費之研究 [10]，原因出在於資料之取得困難，可能是礙於遊戲商之營運機密，且遊戲商鮮少與學術研究團隊合作探討此議題 [3]，故投注更多資源於此議題顯得更為重要。

Author(s)	Method		Category		
	Quantitative	Qualitative	Motivation to use Game Analytics	CLV prediction	Churn prediction
Hadiji et al. (2014)	x				x
Runge et al. (2014)	x				x
Hanner and Zarnekow (2015)	x			x	
Koskenvoima and Mäntimäki (2014)		x	x		
Sifa et al. (2015)	x			x	
Lee et al. (2016)	x				x
Perianez et al. (2016)	x				x
Voigt and Hinz (2016)	x			x	
Milosevic et al. (2017)	x				x
Demediuk et al. (2018)	x				x
Drachen et al. (2018)	x			x	

表 2.1: 近年來機器學習應用於遊戲領域表（此表出自 [8]）

## 2.2 資料前處理

在將巨量資料應用於機器學習前，資料的前處理也是極為重要，相較於在學習模型上進行深入研究與改進，透過資料特徵之轉化及選擇顯得更為重要且有效 [3]。首先將對資料進行清理，只收集有價值之資料，例如：Tamassia 等人只收集遊玩時間超過給定門檻之玩家 [11]、Periáñez 等人只收集消費金額遠高於一般玩家者 [12] 或 Runge 等人只取付費玩家中前 10 % 者 [13]，上述之清理方式皆只著重於具有高資訊量的資料，而不將無價值的資料放入機器學習中。

而針對資料特徵之探勘，Sifa 等人 [10] 將探勘玩家基本資料及玩家行為，再將其進行轉化，例如：取平均值與偏差值於玩家遊玩時間、將玩家國籍分類等等。Lee 等人 [9] 將探勘玩家行為、玩家購買商品數量、玩家遊戲內交易及玩家遊戲內社交，主要針對玩家於遊戲內的行為軌跡。Martínez 等人 [14] 將探勘顧客國籍、兩次購買時間差及最後一次購買時間點等等，再將其進行轉化，例如：取平均值、中位數及最大最小值等等，透過數值間的轉化，提升資料集的維度。Hadiji 等人 [15] 將探勘玩家消費商品數量、玩家遊玩天數等等。

## 2.3 學習模型選擇

在機器學習中，應用於分類預測上為樹狀結構之學習模型最為主流且有效，因其建樹之方式，可以清楚的解釋該筆樣本之預測路徑，進而針對各資料特徵進行重要性的計算與分析 [3]。另外，在樹狀結構之學習模型中，主流以 Bagging 及 Boosting 兩種建樹想法為準：

- Bagging 為從訓練資料集中，隨機取樣並訓練成多份分類器，而每次訓練資料取出後放回，再抽取，最後之預測結果將由多個分類器投票選出，採多數決，且各分類器間的權重關係皆為相等 [16]，如圖 2.1，例如：Random Forest [17] 即為 Bagging + Decision Tree [18]。
- Boosting 為從訓練資料集中，每次訓練使用相同資料，而第  $n$  個分類器於訓練時，將針對第  $n-1$  個分類器分類錯誤的資料增大其權重值，以修正分錯的資訊，希望將分錯的資料減少，預測結果將由多個分類器投票選出，各分類器間的權重關係不同，錯誤率越低的分類器，擁有越高的權重 [19]，如圖 2.2，例如：Extreme Gradient Boosting [20] 即為 Boosting + Decision Tree。

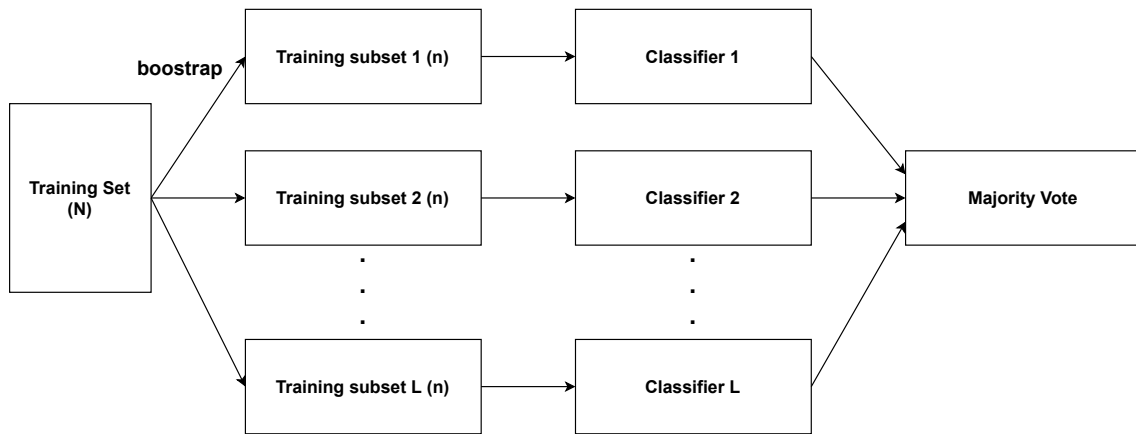


圖 2.1: Bagging 方式建樹示意圖（此圖出自 [21]）

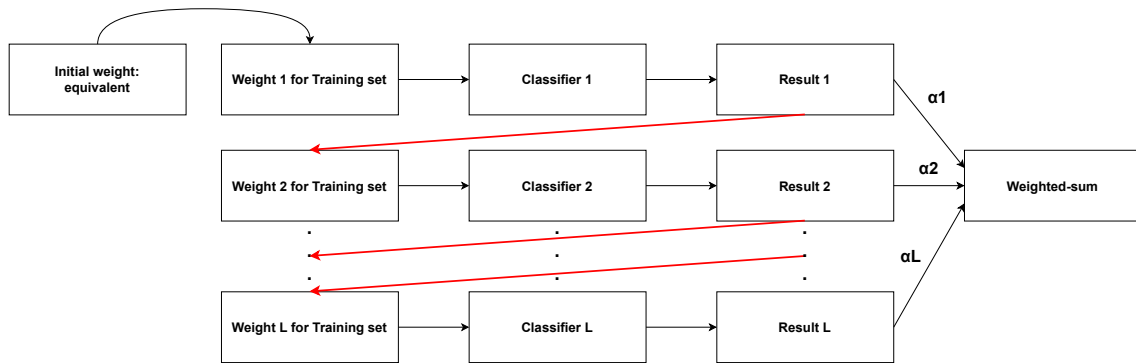


圖 2.2: Boosting 方式建樹示意圖（此圖出自 [21]）

Chen 與 Guestrin [20] 實作出高效率的 Gradient Boosting，稱其為 Extreme Gradient Boosting (XGBoost)，除了使用 Boosting 的方式建樹外，還針對錯誤修正的步驟，引入 Gradient Descent 的概念，加速了學習模型的收斂速度，使其修正錯誤的能力更加精準，大幅減少訓練的時間成本。近年來透過 XGBoost 來訓練的研究越來越多，且其預測能力皆有不錯的表現 [14] [22] [23]，明顯優於 Bagging 建樹方式的其他學習模型。

從上述得知，選用樹狀結構之學習模型將有助於預測分類問題，且其中使用 XGBoost 之成效最佳。因此，為求本論文之預測付費玩家能夠達到預期，將採用 Decision Tree、Random Forest 與 XGBoost 來驗證樹狀結構之優勢以及 XGBoost 之最佳表現。

## 2.4 資料不平衡處理及其評估方式

在遊戲領域進行機器學習訓練時，往往會遭受到資料不平衡的影響；例如：於預測消費上，非付費玩家將會遠大於付費玩家，導致付費玩家資料過少 [10]。於預測流失上，流失者將會遠大於非流失者，導致非流失者資料過少 [9]，前述研究都採以針對資料集進行處理的方式解決資料不平衡，例如：SMOTE，於少數群添加假資料，使得少數群之樣本數與多數群相等 [24]。本論文因不只預測玩家是否會付費，還需分析玩家之消費原因，故需要透過學習模型產出之預測結果進行分析，如在資料集中填入假資料，將會使得分析失準，無法得到有效的資訊，所以我們將採用在機器學習訓練時，放大少數群之樣本權重值，使得學習模型更加著重於少數群的資訊，如同 Boosting 方式建樹時，藉由權重值的不同，修正分類錯誤的資訊 [19]。

在評估資料不平衡資料集時，如果單純計算學習模型之 *precision*、*recall* 或 *F - Score* [25]，將導致多數群之評估結果壓過少數群之評估結果，使得最終評估失真，無法有效驗證學習模型之成效。因此，在評估不平衡資料時，Sifa 等人額外運用 *G - Mean* [26] 來評估學習模型之成效 [10]。藉由上述的概念，本論文將採用 *Weighted F<sub>beta</sub> - Score* 來評估不平衡資料，使得少數群之評估不被多數群所壓過，使用樣本間的數量權重差來計算多數群與少數群的 *F<sub>beta</sub> - Score*，希望能夠合理的評估學習模型間的表現。

### 第 3 章 研究方法

針對遊戲領域巨量資料進行付費玩家預測，我們先將資料進行前處理以及預測前之資料分析，隨後訓練機器學習與其最佳化處理，最後再依預測之結果導入資料特徵重要性分析之中，完成整體預測與分析之工作。

為求研究效率能夠快速且有效，本論文對此議題提出一巨量資料探勘框架：此框架將由四大階段組成，(1) 資料前處理階段：首先將從資料庫群中整合所有所需資料，並對其進行清理，以獲取有價值之原始資料，再著手準備目標值與探勘資料特徵，以利後續分析及機器學習使用；(2) 資料分析階段：使用前階段產出之有價值原始資料進行探索性資料分析 ( Exploratory Data Analysis, EDA ) [27]，檢查是否有不合適之資料特徵以及觀察資料特徵是否可以提供給學習模型較多之資訊；(3) 機器學習階段：首先將有價值原始資料集進行分割為訓練及測試集，隨後針對訓練集進行不平衡資料處理與交叉驗證搭配配參數表，以獲得最佳模型，最後藉由測試集來驗證評估最佳模型，產出預測結果；(4) 預測結果分析階段：使用前階段產出之預測結果進行資料特徵重要性分析，以利更加了解及解釋資料特徵與遊戲所提供之體驗綜合評估。圖 3.1 為巨量資料探勘框架示意圖。

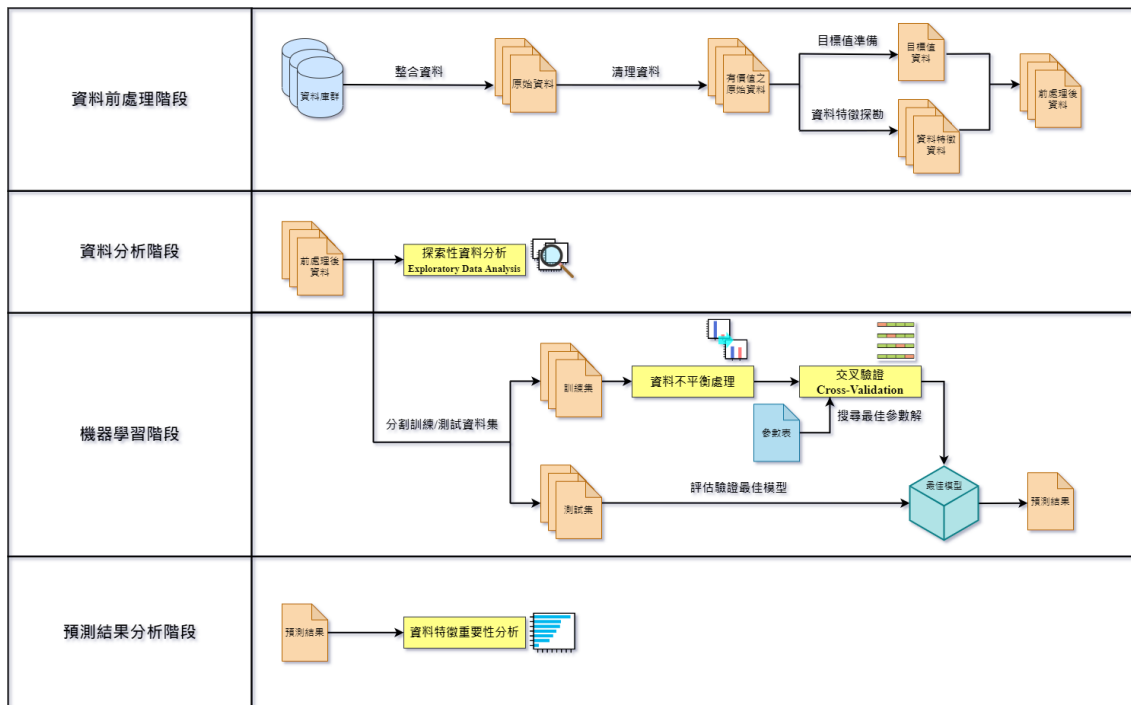


圖 3.1: 本論文之巨量資料探勘框架示意圖

## 3.1 資料前處理階段

此階段將著重於資料之整合與清理，為求能收集到有價值之原始資料，以提高後續分析研究之價值，同時進行目標值的準備與資料特徵的探勘，協助機器學習之訓練，目標產出有價值之玩家遊戲行為軌跡資料集。

### 3.1.1 整合資料

首先資料庫群中之資料皆以天為單位，記錄了各項遊戲之玩家行為軌跡，我們將收集 2019/08/01 至 2019/10/01 之資料，如圖 3.2，此步驟將依各項遊戲為整合目標，重整為多個原始資料集，每個原始資料集中，只會記錄該類遊戲之每位玩家行為軌跡，如圖 3.3，將可提升後續目標值準備及資料特徵探勘速度。

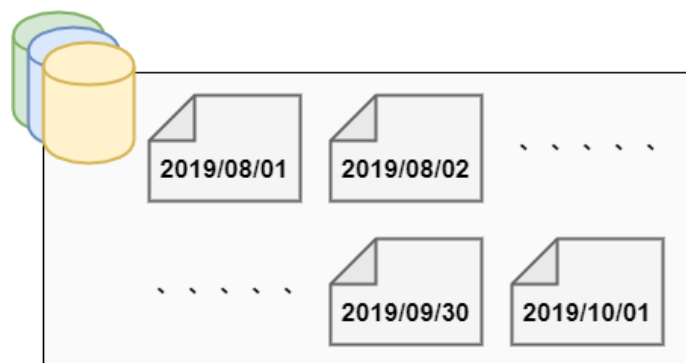


圖 3.2: 資料庫群內資料之示意圖

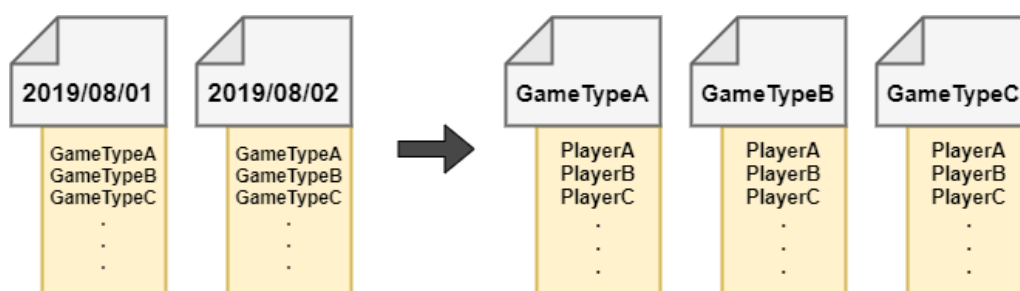


圖 3.3: 依各項遊戲為整合目標之示意圖



除了上述之玩家遊戲行為軌跡原始資料集外，我們還另外收集了玩家資料與遊戲平台玩家消費紀錄，最終此步驟將產出三大類原始資料集(圖 3.4)：

- 玩家資料 ( PlayerInfo. )
- 玩家消費紀錄 ( GameConsume )
- 玩家遊戲行為軌跡 ( GameTypeA, GameTypeB... )

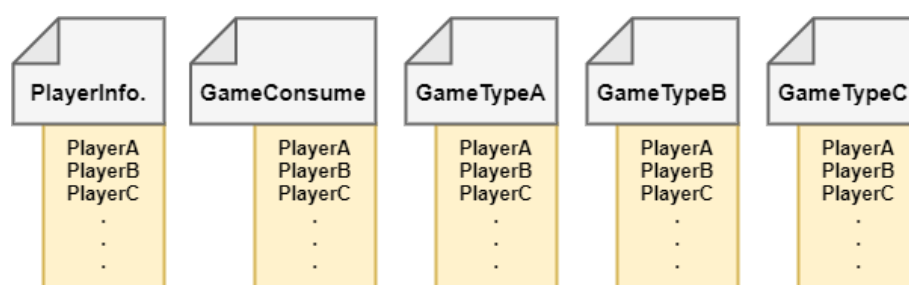


圖 3.4: 原始資料集示意圖

### 3.1.2 清理資料

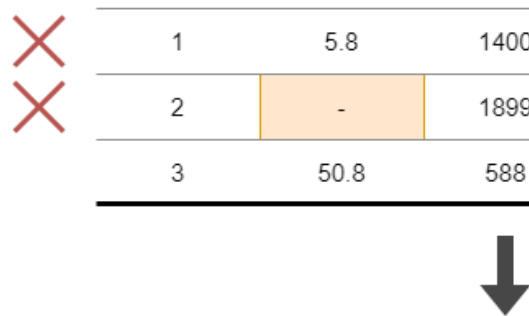
此步驟將針對兩大議題：空缺值處理與無價值玩家資料處理。

為求於進行資料分析及訓練機器學習時，能夠更加準確的了解及預測真實遊戲玩家之特性及付費意願，需要透過上述之處理，來過濾掉潛在的無用資料，使得整體研究能夠聚焦於更有價值的資料上。

### 3.1.2.1 空缺值處理

為了後續資料特徵重要性分析，希望能保持著資料間的真實性，我們將採用直接刪去具有空缺值之樣本方式，而不對資料集進行填入經過處理之假數值，每筆樣本只要在任何資料特徵中擁有一空缺值，即視為欲刪除之對象。

圖 3.5 為空缺值處理之示意圖，從圖中可以看出樣本 ID 1 及 2 分別在 Feature 3 及 1 擁有空缺值，將對兩者予以刪除，故最後只留下樣本 ID 0 及 3 之資料。



ID	Feature1	Feature2	Feature3	Feature4
0	11.5	1254	TW	90
1	5.8	1400	-	78
2	-	1899	JP	100
3	50.8	588	KR	68

ID	Feature1	Feature2	Feature3	Feature4
0	11.5	1254	TW	90
3	50.8	588	KR	68

圖 3.5: 空缺值處理示意圖

### 3.1.2.2 無價值玩家資料處理

對於遊戲領域巨量資料進行研究時，普遍會對所有玩家進行篩檢，以挑選出有價值之玩家族群 [3]，可使整體分析與預測更加貼近於真實遊戲情景。參考前述之概念，我們將對原始資料集中之玩家進行篩檢，定義一無價值玩家觀察期  $N$  天，如玩家在創立帳號日之  $N$  天內無任何遊戲行為軌跡，即視為無價值玩家，予以刪除該玩家及其所有行為軌跡。

圖 3.6 為判別有價值與無價值玩家之示意圖，假設  $N$  為 3 時，從圖中可以看出 Player A 及 B 於無價值玩家觀察期內，皆有遊戲行為軌跡，故定義為有價值玩家；而 Player C 則在觀察期後才有遊戲行為軌跡，將依舊視為無價值玩家；Player D 則在觀察期前後皆無遊戲行為軌跡，故定義為無價值玩家。

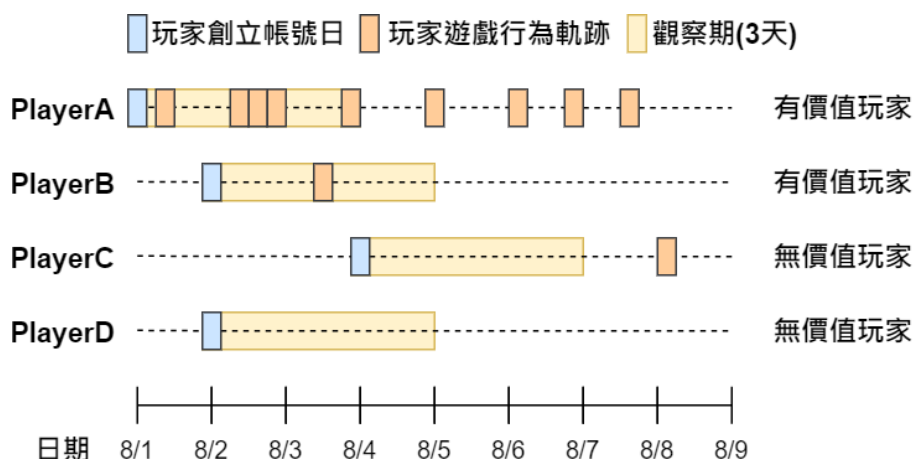


圖 3.6: 判別有價值與無價值玩家之示意圖 (假設  $N$  為 3 時)

經過前兩小節 3.1.2.1 及 3.1.2.2 之處理後，最終此步驟將產出有價值之原始資料集，提供給後續分析及訓練機器學習使用。

### 3.1.3 目標值準備

此步驟將準備供機器學習使用之目標值，即為後續預測所需之 *class*，首先定義一付費玩家定義期  $M$  天，再依  $M$  天來定義目標值「付費玩家」與「非付費玩家」，分別代表 *class 1* 及 *class 0*：

- 付費玩家 (*class 1*)：自玩家創立帳號之  $M$  天內，有消費行為者即為付費玩家；如逾  $M$  天後才消費者，則將不視為付費玩家。
- 非付費玩家 (*class 0*)：所有玩家扣除付費玩家，剩餘者皆為非付費玩家；包含  $M$  天後才消費者。

圖 3.7 為定義付費玩家與非付費玩家之示意圖，假設  $M$  為 7 時，從圖中可以看出 Player A 及 B 於付費玩家定義期內，皆有消費紀錄，故定義為付費玩家 (*class 1*)；而 Player C 則在定義期前後皆無消費紀錄，故定義為非付費玩家 (*class 0*)；Player D 則在定義期後才有消費紀錄，將依舊視為非付費玩家 (*class 0*)。

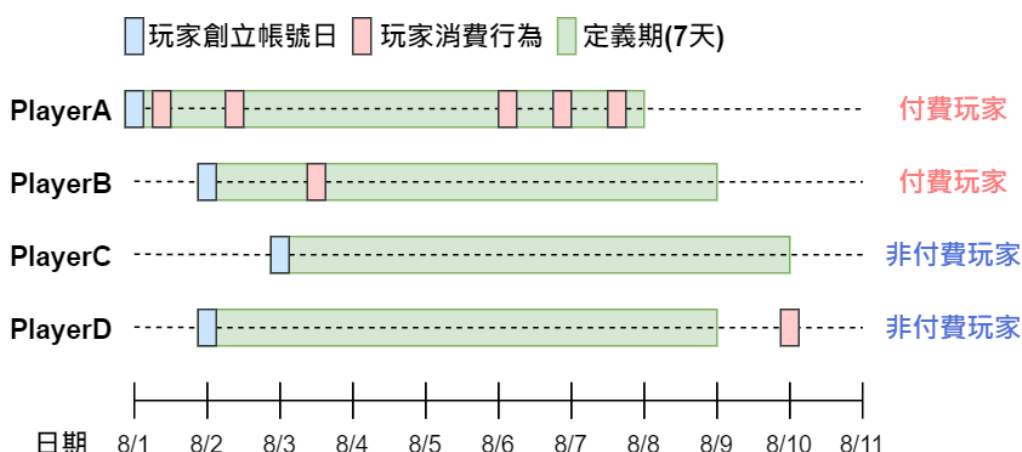


圖 3.7: 定義付費玩家與非付費玩家之示意圖 (假設  $M$  為 7 時)

本論文將預測目標聚焦於新進之潛在付費玩家，所以我們提出付費玩家定義期 ( $M$ ) 來侷限付費玩家之定義。圖 3.8 為付費玩家與非付費玩家范氏圖，可以從圖中看出，最外圍之黑圓框代表所有玩家 (於 3.1.2.2 小節中，篩檢後之有價值玩家)，而藍色底之圓形範圍代表所有非付費玩家 (*class 0*)，其中深藍色底之圓形範圍代表非付費玩家且無消費；淺藍色底之圓形範圍代表非付費玩家但有消費，內圈之綠圓框代表前述所提出之付費玩家定義期 ( $M$ )，綠圓框內之紅色底圓型範圍則代表所有付費玩家 (*class 1*)，可以由上述說明來了解到資料內付費玩家與非付費玩家之分佈狀況及其關係。

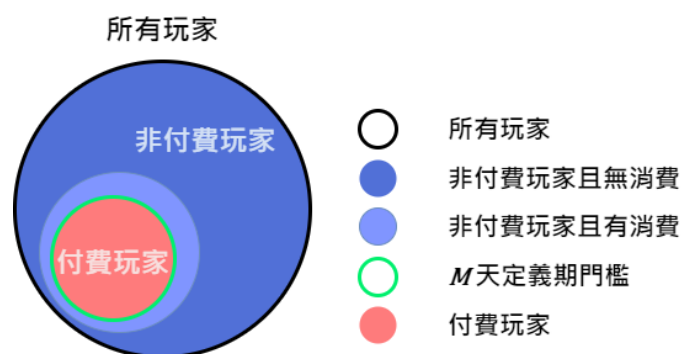


圖 3.8: 付費玩家與非付費玩家范氏圖

### 3.1.4 資料特徵探勘

此步驟將探勘供機器學習使用之資料特徵，即為後續訓練及測試所需之 Features，在遊戲領域巨量資料中，相較於在學習模型上進行深入研究與改進，透過資料特徵之轉化及選擇顯得更為重要且有效 [3]，所以我們將對資料集進行不同面向之探勘，以獲取更多的資訊，讓後續資料分析以及機器學習更加順利。

首先定義一資料特徵探勘期  $G$  天，依照資料特徵探勘期來對每位玩家進行資料特徵探勘，自玩家創立帳號日之  $G$  天內，探勘其所需之資料特徵，並且  $G$  值之設定必需介於 3.1.2.2 小節之無價值玩家觀察期 ( $N$ ) 以及 3.1.3 小節之付費玩家定義期 ( $M$ ) 之中，如  $N \leq G \leq M$ ，使得資料特徵之探勘有意義。

圖 3.9 為資料特徵探勘期示意圖，包含無價值玩家觀察期與付費玩家定義期，假設  $G$  為 3； $N$  為 3； $M$  為 7 時，從圖中可以看出，所有玩家皆是有價值之玩家，於無價值玩家觀察期內皆有遊戲行為軌跡，而 Player A 及 B 為付費玩家；Player C 及 D 為非付費玩家，取決於付費玩家定義期內是否有消費紀錄，再針對各玩家自創立帳號日之探勘期內進行探勘資料特徵之處理，探勘到之資料特徵皆有意義且在付費玩家定義期之內。

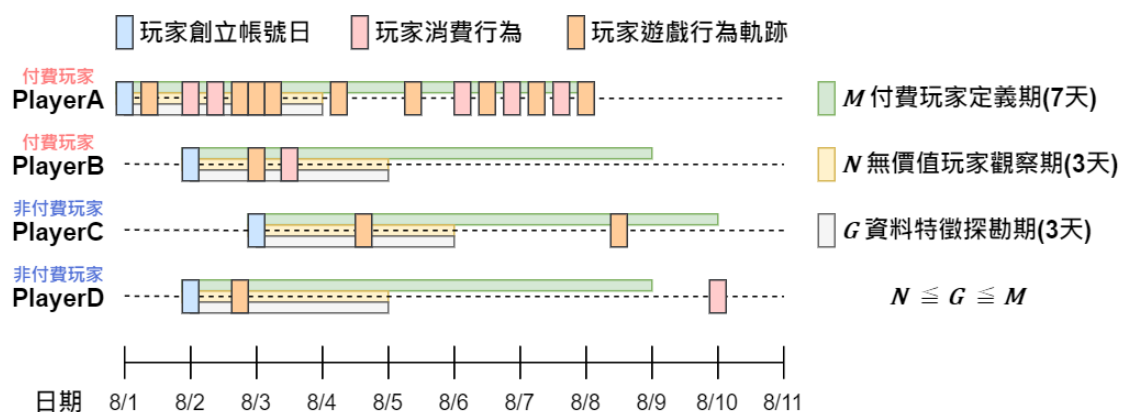


圖 3.9: 資料特徵探勘期示意圖 (包含無價值玩家觀察期 ( $N$ ) 與付費玩家定義期 ( $M$ ))(假設  $G$  為 3； $N$  為 3； $M$  為 7 時))

資料特徵之探勘面向將參考於 [9] [10] [14] 之探勘想法，主要聚焦於玩家之遊戲行為軌跡，並將其進行轉化與設計綜合指標特徵，最終我們將資料特徵種類分為三大類：

- 設備資訊：包含玩家之設備相關資訊
- 平台遊戲行為軌跡：包含玩家以平台為探勘範疇之遊戲行為軌跡
- 各遊戲行為軌跡：包含玩家以各遊戲為探勘範疇之遊戲行為軌跡

各面向之詳細探勘資料特徵如表 3.1，其中遊戲貨幣 A 及 B 為購買遊戲內禮包獲得，並為平台內共通之遊戲籌碼。

特徵種類	特徵
設備資訊	設備所在地
	設備平台
	設備廠牌
	設備型號
平台遊戲行為軌跡	遊戲貨幣 A 之餘額變化
各遊戲行為軌跡	遊玩天數
	遊戲貨幣 A 之餘額變化
	總贏遊戲次數
	總贏分
	獲得遊戲貨幣 A 之總額
	獲得遊戲貨幣 B 之總額

表 3.1: 資料特徵種類表

## 3.2 資料分析階段

此階段將著重於資料之分析，為求在訓練機器學習前，可以藉由資料分析之方法來了解到資料之特性，以提高後續解讀資料特徵之重要性與其相關之連結；另外，還將進行檢查是否有不適合作為資料特徵之資料以及觀察資料特徵是否可以提供給學習模型較多的資訊。

### 3.2.1 探索性資料分析 ( Exploratory Data Analysis, EDA )

我們將採用探索性資料分析 ( Exploratory Data Analysis, EDA ) [27] 來藉由圖表呈現協助了解資料之特性，並且檢查不適合之資料特徵與高資訊量之資料特徵，此步驟將利用下列兩種圖表來觀察資料特性：

- 直方圖：觀察資料特徵是否存在傾斜，如圖 3.10 (a)。
- 散佈圖：觀察資料特徵間之關聯性，如圖 3.10 (b)。

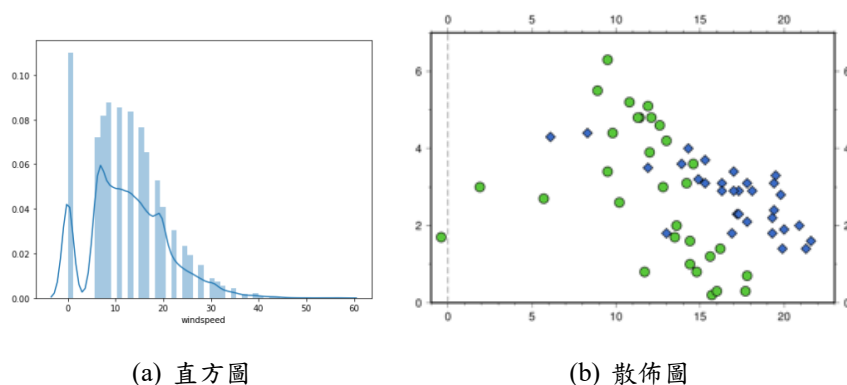


圖 3.10: 探索性資料分析之使用圖表類型

為求後續訓練機器學習之穩定度與合理性，需要透過排除不合理之資料特徵，來避免學習模型設置錯誤之門檻於付費玩家與非付費玩家；並且為了使後續資料特徵重要性之解讀能夠更加清晰，於產出預測結果前即針對資料集進行資料分析，提早推測高資訊量之資料特徵，我們將針對此兩大議題：不合理之資料特徵處理與高資訊量之資料特徵處理。

### 3.2.1.1 不合理之資料特徵

於前述 3.1.4 小節中探勘到的資料特徵，其中可能包含著不合理之資料特徵，此種資料特徵將有可能誤導於學習模型進行訓練，設置錯誤門檻於付費玩家與非付費玩家，故需將此類資料特徵排除，以利提升學習模型之可靠性以及避免後續資料特徵重要性分析有誤；例如：某項資料特徵僅有已付費玩家才可以遊玩，進而獲取遊戲行為軌跡；而非付費玩家則無法遊玩，所以無法獲取遊戲行為軌跡，此種情況將提供於學習模型不合理之資訊，誤認為無該筆遊戲行為軌跡，則有極高機率即是非付費玩家，但是還有可能僅為付費玩家尚未遊玩該遊戲，如圖 3.11。

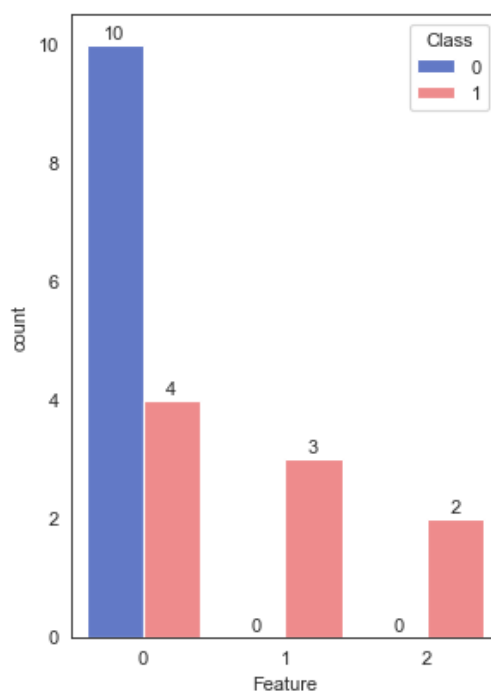


圖 3.11: 不合理之資料特徵示意圖



### 3.2.1.2 高資訊量之資料特徵

藉由觀察資料特徵之分佈是否有明顯差異性，而推測此資料特徵能夠提供給學習模型較多的資訊，將此類資料特徵認為是高資訊量之資料特徵，使得後續資料特徵重要性分析之解釋可以更加順利。

如圖 3.12，可以從圖中看出，非付費玩家資料分佈集中於 0 附近；而付費玩家資料分佈則分散於 y 軸上，可見此種資料特徵容易區分出付費玩家與非付費玩家，能夠提供給學習模型較多的資訊。

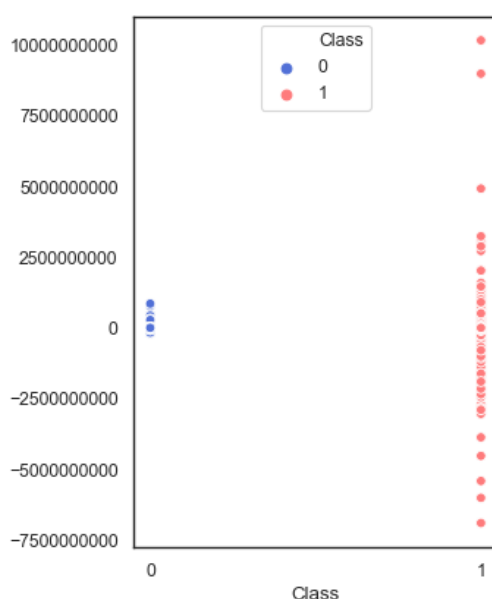


圖 3.12: 高資訊量之資料特徵示意圖

## 3.3 機器學習階段

此階段將著重於機器學習訓練以及資料不平衡處理，最後產出最佳模型之預測結果，提供給付費玩家之預測分析以及資料特徵重要性分析使用，我們將選擇樹狀結構之學習模型進行訓練，樹狀結構之學習模型對於巨量資料分類預測顯得更為合適，並且對於預測結果之解釋也相對清楚 [3] [10]，而本論文所挑選之學習模型包含：Decision Tree [18]、Random Forest [17] 與 Extreme Gradient Boosting [20]。



上述三種學習模型之資訊量計算以 *Gini Impurity* 為主，如式 3.1，其中  $c$  為 *Class*、 $p(i)$  為  $c$  之發生機率，透過 *Gini Impurity* 來衡量建樹時之分類準則，挑選出最適合用來進行分割之資料特徵及數值。

$$Gini\ Impurity(D) = G(D) = 1 - \sum_{i=1}^c p(i)^2 \quad (3.1)$$

最後將比較上述三種不同學習模型來挑選出最佳之模型，包含 Bagging、Boosting 不同方式之建樹差異。

### 3.3.3 資料不平衡處理

進行機器學習訓練於遊戲領域巨量資料時，往往將會遭受資料不平衡之問題，進而影響學習模型之成效與可靠度 [9] [10] [29]，普遍研究中將針對資料集進行預處理，設法解決資料不平衡之問題，例如：Under-Sampling：TomekLinks [30]、InstanceHardnessThreshold [31] 與 RandomUnderSampler；Over-Sampling：SMOTE [24] 與 RandomOverSampler；Combination of Under- and Over-sampling：SMOTETomek [32]；Ensemble：EasyEnsemble [33]。

為了確保資料間之真實性，我們於處理資料不平衡時，不希望針對資料集進行加工，如上述之 Under-Sampling、Over-Sampling 以及 Combination of Under- and Over-Sampling，此類處理方式皆將會對原始資料集進行破壞，無法呈現出真實資料集之特性。所以本論文將重點放於訓練學習模型時的樣本權重影響，而不對資料集進行直接處理，樣本權重設置如式 3.2，其中  $N_0$  為非付費玩家 (*class 0*) 之樣本數； $N_1$  為付費玩家 (*class 1*) 之樣本數，將計算  $N_0$  與  $N_1$  之比例差距，並取地板函數 (floor function)，此值則為付費玩家 (*class 1*) 樣本權重放大倍數，使得學習模型更加重視於少數群，即為更重視付費玩家之資訊。

$$class\ 0 : class\ 1 = 1 : \left\lfloor \frac{N_0}{N_1} \right\rfloor \quad (3.2)$$

### 3.3.4 搜尋最佳參數解

此步驟將對前述 3.3.2 小節挑選之學習模型進行搜尋最佳參數解，以調教出最適合該學習模型之參數。各學習模型之調教參數如表 3.2，針對各學習模型之結構不同，挑選不同的參數進行最佳化，各參數意義說明如表 3.3。

學習模型	Decision Tree	Random Forest	XGBoost
參數調教	max_depth	n_estimators	n_estimators
	min_samples_split	max_depth	max_depth
	min_samples_leaf	min_samples_split	
	min_samples_leaf		

表 3.2: 學習模型參數調教表

參數	參數說明
n_estimators	多樹結構之樹總數
max_depth	樹狀結構之最大深度限制
min_samples_split	節點分割之最小樣本數限制
min_samples_leaf	葉節點之最小樣本數限制

表 3.3: 學習模型參數說明表

### 3.3.5 交叉驗證 ( Cross Validation )

針對訓練資料集進行交叉驗證 ( Cross Validation )，並且搭配前頁之參數調教，最後輸出最佳模型，我們將參考 [34] 中所使用之 RepeatedStratifiedKFold 方法，其中使用 Stratified 方式分割，即為在各 Fold 中，付費玩家與非付費玩家之資料比例將會相等；使用 Repeated 方式反覆驗證，即為反覆執行上述之交叉驗證。透過上述之分割方式，可以在每次訓練學習模型時，使真實訓練集保持著原始訓練集的付費玩家與非付費玩家比例。

圖 3.14 為 RepeatedStratifiedKFold 示意圖，假設 Repeated 為 2；KFold 為 3 時，可以從圖中看出，首先依照前述 3.3.1 小節，從所有資料初步分割出原始訓練集與測試集，再針對原始訓練集進行 RepeatedStratifiedKFold，進行了兩次的交叉驗證，而每次分割原始訓練集時可以看到淺藍底之真實訓練集與淺橘底之驗證集中的付費玩家與非付費玩家比例與原始測試集相等，並且真實訓練集與驗證集中的付費玩家與非付費玩家比例也相等。



圖 3.14: RepeatedStratifiedKFold 示意圖 (假設 Repeated 為 2；KFold 為 3 時)

另外，因類別型資料特徵不適用於樹狀結構之學習模型，故在其應用於機器學習前，將此類資料特徵透過 One Hot Encoding 進行轉化，以利機器學習訓練。

### 3.3.6 評估驗證最佳模型

前述 3.3.5 小節中交叉驗證搭配 3.3.4 小節中的參數調教表所使用之評估值為  $Weighted F_{beta} - Score$ ，擇其最高值之學習模型，選定為最佳模型。 $Weighted F_{beta} - Score$  為在  $F_{beta} - Score$  評估值上導入樣本數權重概念，如式 3.3 與式 3.4，適合使用在評估資料不平衡之資料集中。

$$F_{beta} = (1 + \beta^2) \times \frac{precision \times recall}{(\beta^2 \times precision) + recall} \quad (3.3)$$

$$Weighted F_{beta} = \frac{N_1}{N_0 + N_1} \times F_{beta\ 1} + \frac{N_0}{N_0 + N_1} \times F_{beta\ 0} \quad (3.4)$$

其中  $beta (\beta)$  則為  $precision$  與  $recall$  之間的比重，如表 3.4。本論文預測玩家是否會付費，將著重於  $recall$ ，即為是否將所有潛在付費玩家預測出來，因為新進玩家有可能是經由廣告吸引而來，而該玩家身上即帶有廣告投放之成本，故希望能將有可能會付費的玩家全部預測出來，使得遊戲商能獲得相應的營收。所以  $beta (\beta)$  值之設置將為付費玩家 ( $class\ 1$ ) 樣本權重放大倍數，因本論文資料集為不平衡資料，故  $beta (\beta)$  值將必大於等於 1。

$\beta$ 數值範圍	說明
$0 < \beta < 1$	評估著重於 $precision$
$\beta = 1$	$precision$ 與 $recall$ 比重相當
$1 < \beta$	評估著重於 $recall$

表 3.4:  $\beta$  數值意義表

## 3.4 預測結果分析階段

此階段將著重於資料特徵重要性之分析，透過前述 3.3.6 小節所產出之預測結果，計算其資料特徵於各學習模型中各樹之重要性，並加總後正規化。產出之分析結果將與前述 3.2.1.2 小節中推測之資料特徵進行探討，並藉由最終結果對遊戲中的遊玩體驗進行評估與建議。

### 3.4.1 資料特徵重要性分析

我們將資料特徵重要性 (*Feature Importance, fi*) 定義為加總各樹中各資料特徵於節點分割時所提供之 *Gini Impurity* (見式 3.1)，稱為 *Gini Importance*，再將其正規化至區間 [0,1] 中。

式 3.5 為計算樹中各節點之 *Gini Importance*，其中  $D_p$  為父節點、 $N_p$  為父節點之樣本數、 $D_{left}$  為左子節點、 $N_{left}$  為左子節點之樣本數、 $D_{right}$  為右子節點、 $N_{right}$  為右子節點之樣本數，首先計算  $D_p$ 、 $D_{left}$  及  $D_{right}$  之 *Gini Impurity*，並計算  $D_{left}$  及  $D_{right}$  之樣本數權重比例，最後將  $D_p$  之 *Gini Impurity* 減去兩權重值。

$$Gini\ Importance(D_p) = GI(D_p) = G(D_p) - \frac{N_{left}}{N_p} \times G(D_{left}) - \frac{N_{right}}{N_p} \times G(D_{right}) \quad (3.5)$$

式 3.6 為計算資料特徵於單樹中之重要性，其中  $x$  為欲求其重要性之資料特徵、 $k$  為節點分割時所用資料特徵為  $x$  之所有節點、 $l$  為樹中所有節點，首先加總所有  $k$  之 *Gini Importance*，並加總  $l$  之 *Gini Importance*，最後將其進行正規化計算，落於區間 [0,1] 中，並總和為 1。

$$fi(t, x) = \frac{\sum_{k \in \text{node split based on } x} GI(D_k)}{\sum_{l \in \text{all nodes}} GI(D_l)} \quad (3.6)$$

式 3.7 為計算資料特徵於多樹中之重要性，其中  $x$  為欲求其重要性之資料特徵、 $t$  為學習模型中的所有樹、 $N_{trees}$  為樹總數，首先加總所有  $t$  中  $x$  的  $fi(t, x)$ ，並取其平均於  $N_{trees}$  中，最後即計算出  $x$  於學習模型內之資料特徵重要性 ( $fi$ )。

$$fi(x) = \frac{\sum_{t \in all\ trees} fi(t, x)}{N_{trees}} \quad (3.7)$$



## 第 4 章 實驗結果與分析

此章節中，我們將針對前述第 3 章之研究方法進行實驗結果評估與分析，其中包含：第 4.1 節 實驗系統架構、第 4.2 節 資料前處理評估、第 4.3 節 資料分析評估、第 4.4 節 機器學習評估以及第 4.5 節 預測結果分析評估。

### 4.1 實驗系統架構

本論文實驗系統架構將分為資料前處理端、資料分析端與機器學習端，如表 4.1，並均以 *Python* 做為開發語言。

- 資料前處理端：以 *Apache Spark* [35] 處理資料之前處理以及分割資料集所使用。
- 資料分析端：以 *missingno* [36]、*Seaborn* [37] 協助以圖表方式呈現資料特性。
- 機器學習端：以 *pandas* [38] [39]、*scikit-learn* [40] [41]、*XGBoost* [20] 處理機器學習訓練與評估。

系統端點	資料前處理端	資料分析端	機器學習端
	<i>Apache Spark</i>	<i>missingno</i>	<i>pandas</i>
研究環境		<i>Seaborn</i>	<i>scikit-learn</i>
			<i>XGBoost</i>

表 4.1: 實驗系統架構之研究環境表

我們所使用之資料集包含 1,824,962 位玩家及其遊戲行為軌跡，總容量約為 12.2 GB。

## 4.2 資料前處理評估

此階段將評估前章 3.1.2.1 小節之資料空缺值處理、3.1.2.2 小節之無價值玩家處理以及兩小節 3.1.3 及 3.1.4 之目標值與資料特徵處理，分別於 4.2.1 小節 清理資料評估及 4.2.2 小節 目標值與資料特徵評估說明。

### 4.2.1 清理資料評估

我們將從玩家資料中刪除有空缺值存在之玩家，並再刪除該玩家之遊戲行為軌跡，如圖 4.1，透過 *missingno* 產出圖表，Country 為設備所在地、DeviceName 為設備型號、DeviceBrand 為設備廠牌、Platform 為設備平台，可以從圖中看出，僅有設備廠牌存在空缺值，共有 18,037 位玩家，另外除了有空缺值存在外，還有值為「Unknown」之資料，將同樣視為空缺值予以刪除，共有 10,198 位玩家。共刪除 28,235 位玩家，約占總玩家 1.55 %。

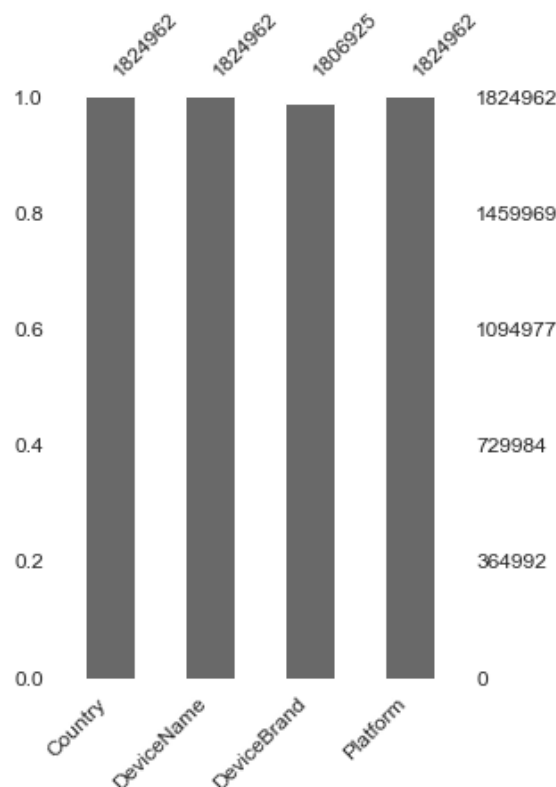


圖 4.1: 玩家資料空缺值示意圖 (X 軸為玩家資料之資料特徵名；Y 軸為該資料特徵非空缺值樣本數量)

我們將從各項遊戲原始資料集中撈取  $N$  天（無價值玩家觀察期）內的遊戲行為軌跡，並於  $N$  值代入 1, 2, 3, 4, 5，分別代表觀察 1, 2, 3, 4, 5 天內之玩家是否有遊玩過各項遊戲，實驗結果如表 4.2。

無價值玩家觀察期	無價值玩家數	無價值玩家佔比
$N = 1$	331,207	18.43 %
$N = 2$	305,058	16.98 %
$N = 3$	299,314	16.66 %
$N = 4$	295,826	16.46 %
$N = 5$	293,108	16.31 %

表 4.2: 無價值玩家觀察表

有價值玩家數將由玩家總數扣除空缺值玩家數與無價值玩家數，如表 4.3，從表中可以看出， $N$  值於 1 至 5 間之有價值玩家數差距不大，約在 1,500,000 位左右。本論文將挑選  $N = 3$  做後續實驗，即為 1,497,413 位有價值玩家。

玩家總數	空缺值玩家數	無價值玩家觀察期	無價值玩家數	有價值玩家數
1,824,962	28,235	$N = 1$	331,207	1,465,520
		$N = 2$	305,058	1,491,669
		$N = 3$	299,314	1,497,413
		$N = 4$	295,826	1,500,901
		$N = 5$	293,108	1,503,619
有價值玩家數 = 玩家總數 - 空缺值玩家數 - 無價值玩家數				

表 4.3: 有價值玩家觀察表

### 4.2.2 目標值與資料特徵評估

我們將從玩家資料與玩家消費紀錄中定義  $M$  天 (付費玩家定義期) 內有消費行為之玩家，並且  $M$  值最小值必須不小於  $N$  值，即為 3 天，故  $M$  值代入 3, 4, 5, 6, 7，分別代表觀察 3, 4, 5, 6, 7 天內之玩家是否有消費行為，如表 4.4，本論文將挑選  $M = 7$  做後續實驗，即為 41,584 位付費玩家及 1,455,829 位非付費玩家。

付費玩家定義期	付費玩家數	付費玩家佔比	非付費玩家	非付費玩家佔比
$M = 3$	35,088	2.43 %	1,462,325	97.57 %
$M = 4$	37,504	2.50 %	1,459,909	97.50 %
$M = 5$	39,137	2.61 %	1,458,276	97.39 %
$M = 6$	40,425	2.70 %	1,456,988	97.30 %
$M = 7$	41,584	2.78 %	1,455,829	97.22 %

表 4.4: 付費玩家及非付費玩家定義表

定義一消費速度，即為將付費日 - 創立帳號日，透過 *Seaborn* 產出圖表，圖 4.2 為觀察付費玩家之消費速度，從圖中可以看出，玩家於創立帳號日當天即消費者為多數，隨著天數增加，消費意願則降低；故本論文將聚焦於新進之潛在玩家，以保握住玩家高消費意願時期。

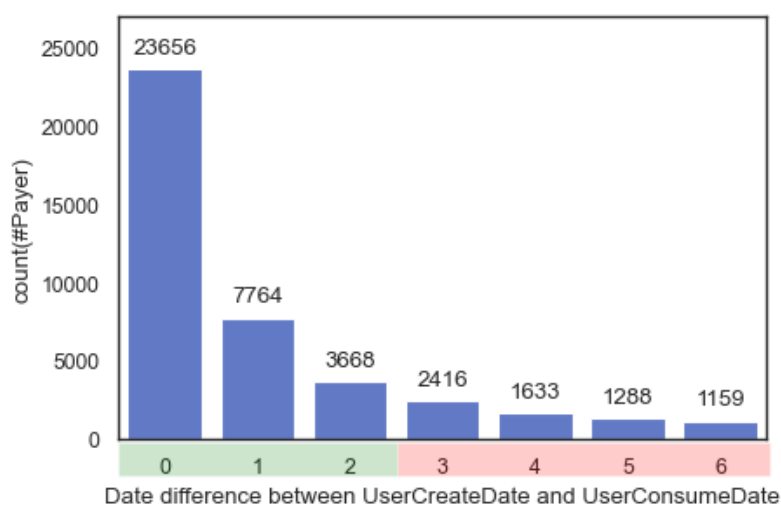


圖 4.2: 付費玩家之消費速度圖 (X 軸為付費玩家之消費速度；Y 軸為付費玩家數)

我們將從原始資料集中探勘出  $G$  天 ( 資料特徵探勘期 ) 內之資料特徵，參考於上述所用之  $N$  值 (  $N = 3$  ) 與  $M$  值 (  $M = 7$  )，選擇  $G = 3$  做後續實驗，即為探勘玩家 3 天內之資料特徵，並根據表 3.1，透過 *Apache Spark* 設置實驗環境以進行探勘。

在探勘各遊戲行為軌跡時，將依據各遊戲之提供資料不同，而探勘不同的資料特徵，如表 4.5，可以從表中看出，依各遊戲其資料特性不同而探勘不同資料特徵，而在 GameType C, D, E, F 在平台中不只一款遊戲，故除了分別探勘各款遊戲之資料特徵外，另外將各款遊戲加總整理為依加權資料特徵。

資料特徵 遊戲種類	遊玩天數	遊戲貨幣 A 之餘額變化	總贏遊戲次數	總贏分	獲得遊戲貨幣 A 之總額	獲得遊戲貨幣 B 之總額
GameType A	●	●	●	●		
GameType B	●					
GameType C*	●				●	
GameType D*	●		●	●		
GameType E*			●		●	
GameType F*	●		●		●	●
*：該遊戲種類於平台中不只一款遊戲，故除了分別探勘各款遊戲之特徵外，另外將各款遊戲加總整理為一加權特徵。						

表 4.5: 各遊戲行為軌跡探勘表

由表 4.5 中所探勘出的資料特徵總數如表 4.6，總資料特徵數將由資料特徵數乘以平台內遊戲款數並加上加權資料特徵數，可以從表中看出，共探勘出了 118 個資料特徵，其中以 GameType D 為最多，因其遊戲款數於平台中佔最多。

遊戲種類	資料特徵數	平台內 遊戲款數	加權資料特徵數	總資料特徵數
GameType A	4	1	0	4
GameType B	1	1	0	1
GameType C*	2	2	2	6
GameType D*	3	24	3	75
GameType E*	2	7	2	16
GameType F*	4	3	4	16
			總和	118
總資料特徵數 = 資料特徵數 × 平台內遊戲款數 + 加權資料特徵數				

表 4.6: 各遊戲行為軌跡資料特徵數表

表 4.7 為資料特徵總數表，從表中可以看出，在各遊戲行為軌跡中擁有最多的資料特徵，本論文希望可以透過玩家在各遊戲中的行為來預測其是否會付費。

特徵種類	特徵總數
設備資訊	4
平台遊戲行為軌跡	1
各遊戲行為軌跡	118
總和	123

表 4.7: 資料特徵總數表

## 4.3 資料分析評估

此階段將評估前章 3.2.1.1 小節之不合理資料特徵處理及 3.2.1.2 小節之高資訊量資料特徵推測，於 4.3.1 小節 探索性資料分析評估說明。

### 4.3.1 探索性資料分析評估

利用長條圖觀察設備所在地，如圖 4.3 及圖 4.4，從兩圖中可以看出，Country 3 的玩家數最多，但其付費玩家比例則偏低，可見雖有大量的玩家遊玩，卻無法提升其付費意願；而 Country 4 的玩家數雖不突出，但其付費玩家比例則最高，可見於該地之玩家相較於 Country 3 有更高的付費意願，可能是因為環境或消費行為不同所造成，所以相較於提升玩家數，更重要的是在於如何提高玩家付費意願。

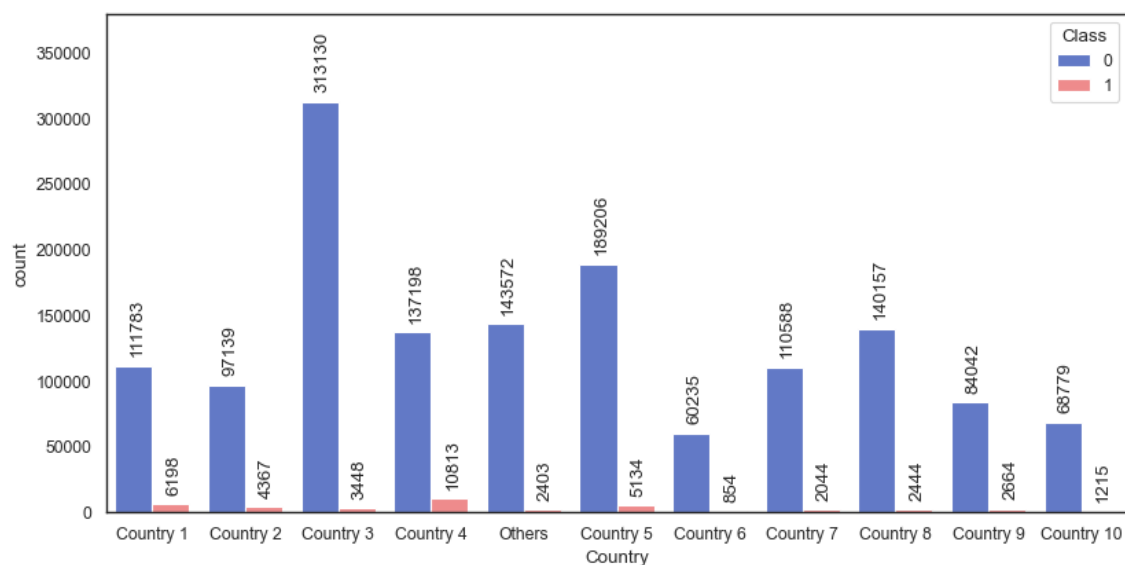


圖 4.3: 觀察設備所在地之付費玩家與非付費玩家數量長條圖 (X 軸為設備所在地；Y 軸為玩家數量)



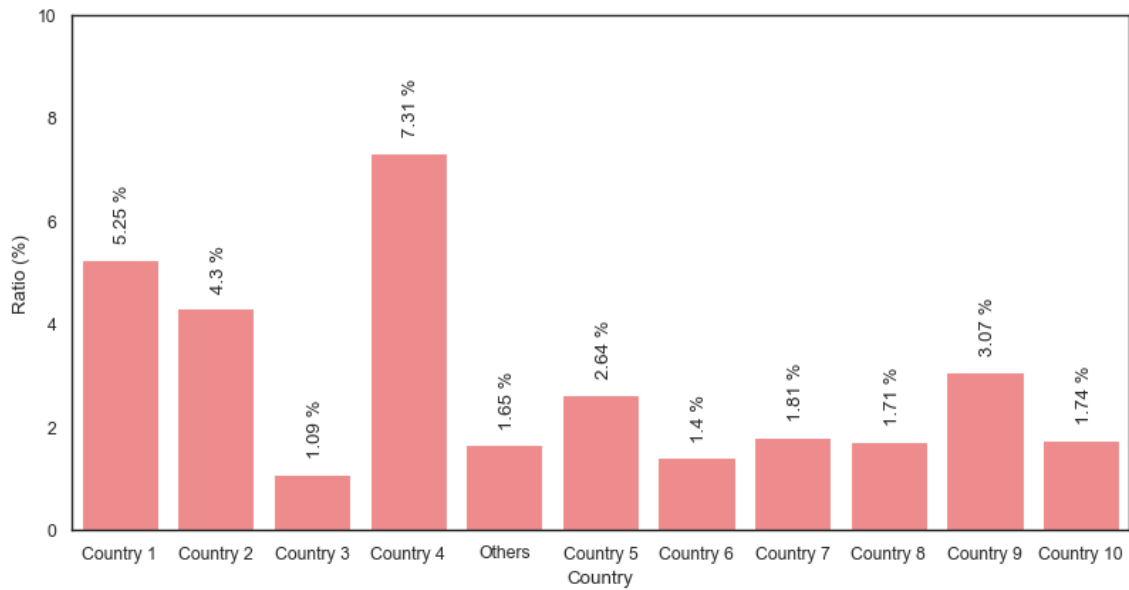


圖 4.4: 觀察設備所在地之付費玩家比例長條圖 (X 軸為設備所在地；Y 軸為付費玩家比例)

利用長條圖觀察在資料集中是否有不合適之資料特徵存在，如圖 4.5，該圖為 GameTypeE 59 遊戲之總贏遊戲次數，從圖中可以看出，僅有付費玩家有數值，而非付費玩家則全數皆為 0，造成此種現象之原因為因為該款遊戲僅有付費玩家可以遊玩，故將不適合當作資料特徵，予以刪除。

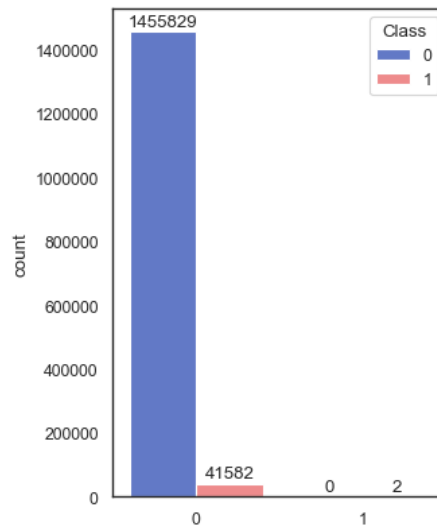
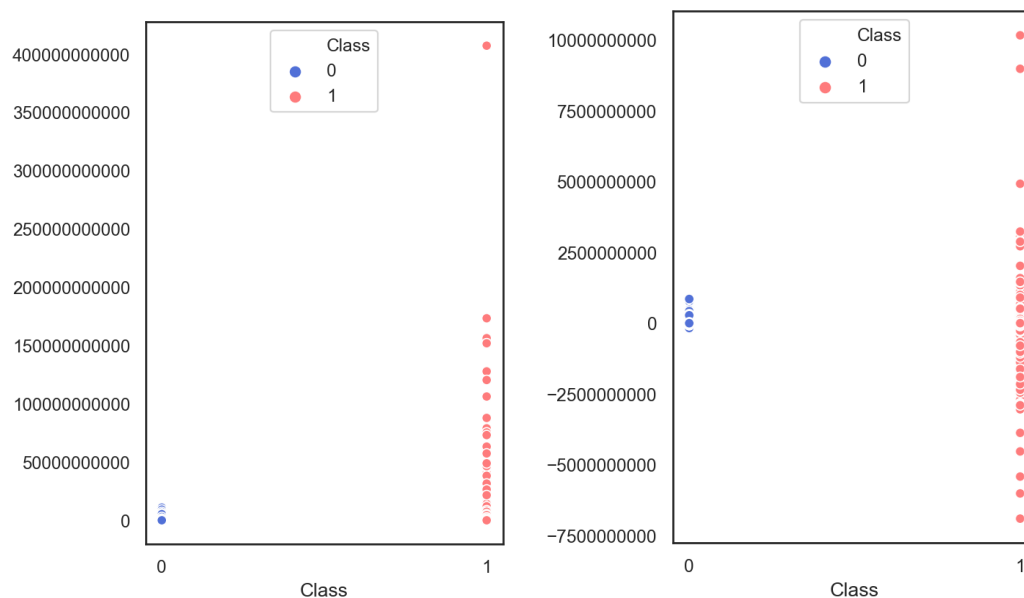


圖 4.5: 觀察 GameTypeE 59 號遊戲之總贏遊戲次數長條圖 (X 軸為總贏遊戲次數；Y 軸為玩家數量)

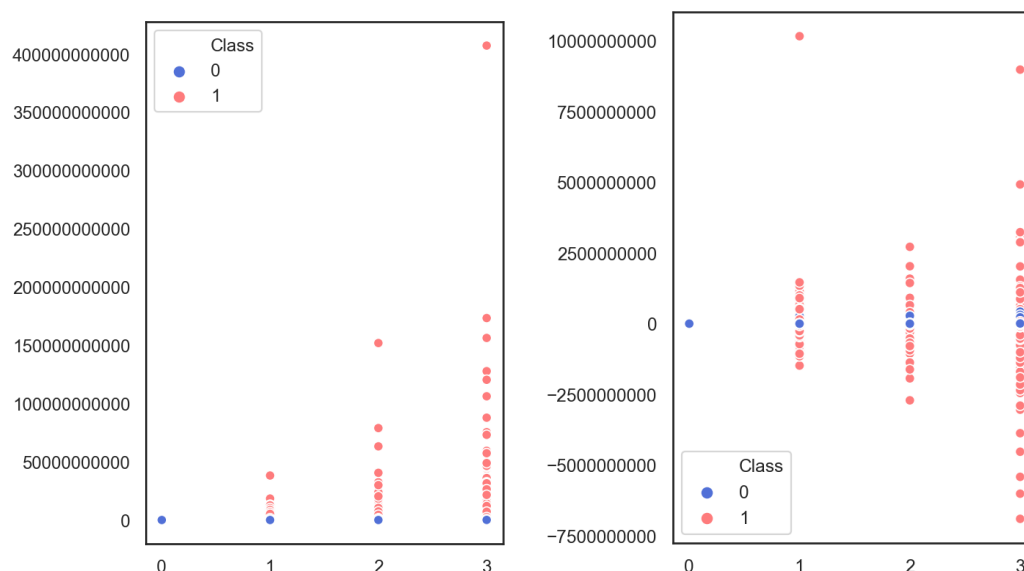
利用散佈圖觀察在資料集中的高資訊量資料特徵，如圖 4.6 (a) 及圖 4.6 (b)，該圖組為 GameTypeA 之總贏分與遊戲貨幣 A 之餘額變化，從兩圖中可以看出，付費玩家與非付費玩家之分佈有明顯差異，推測可以帶給學習模型很好的分類資訊。



(a) 總贏分散佈圖 (X 軸為非付費玩家與付費 (b) 遊戲貨幣 A 之餘額變化散佈圖 (X 軸為非  
 玩家；Y 軸為總贏分) 付費玩家與付費玩家；Y 軸為遊戲貨幣 A 之  
 餘額變化)

圖 4.6: 觀察 GameTypeA 資料特徵散佈圖

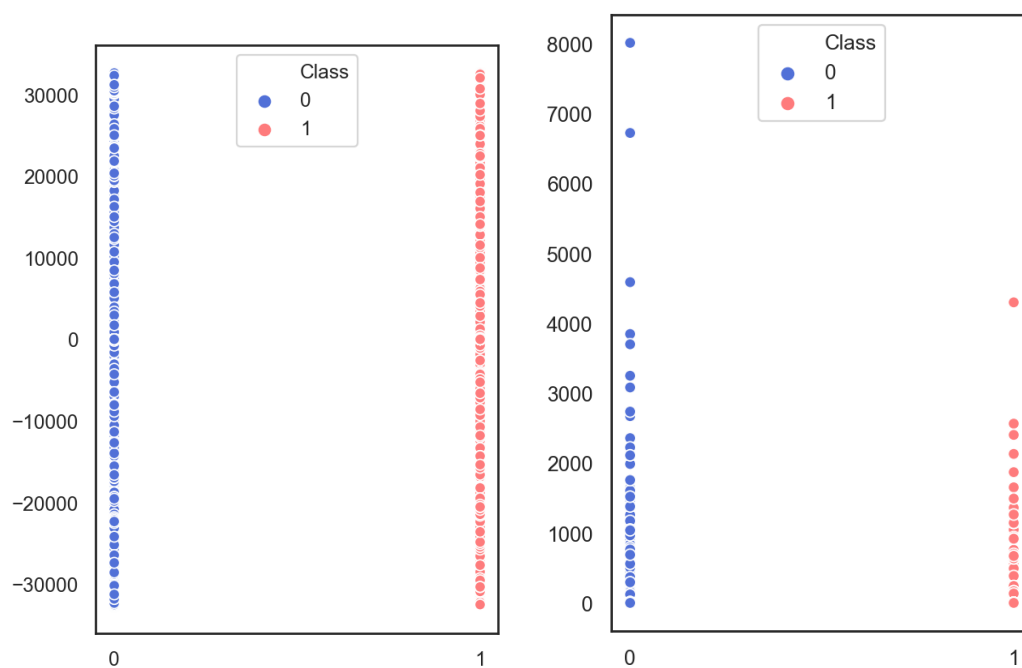
圖 4.7 (a) 及圖 4.7 (b)，該圖組為 GameTypeA 之遊玩天數與總贏分、遊戲貨幣 A 之餘額變化，從兩圖中可以看出，隨著遊玩天數的增加，數值的差異性則拉大，推測可以帶給學習模型很好的分類資訊。



(a) 遊玩天數與總贏分散佈圖 (X 軸為遊玩天數；Y 軸為總贏分)  
(b) 遊玩天數與遊戲貨幣 A 之餘額變化散佈圖 (X 軸為遊玩天數；Y 軸為遊戲貨幣 A 之餘額變化)

圖 4.7: 觀察 GameTypeA 資料特徵關聯散佈圖

圖 4.8 (a) 及圖 4.8 (b)，該圖組為 GameTypeD 65 號遊戲之總贏分與總贏遊戲次數，從兩圖中可以看出，付費玩家與非付費玩家之分佈明顯無差異，推測無法帶給學習模型很好的分類資訊。



(a) 總贏分散佈圖 (X 軸為非付費玩家與付費 (b) 總贏遊戲次數散佈圖 (X 軸為非付費玩家  
 玩家；Y 軸為總贏分) 與付費玩家；Y 軸為總贏遊戲次數)

圖 4.8: 觀察 GameTypeD 65 號遊戲資料特徵關聯散佈圖

圖 4.9，該圖為 GameTypeE 62 號遊戲之總贏遊戲次數，從圖中可以看出，在贏遊戲次數偏低時，非付費玩家佔了大多數，而付費玩家則相對偏少，推測可以帶給學習模型很好的分類資訊。

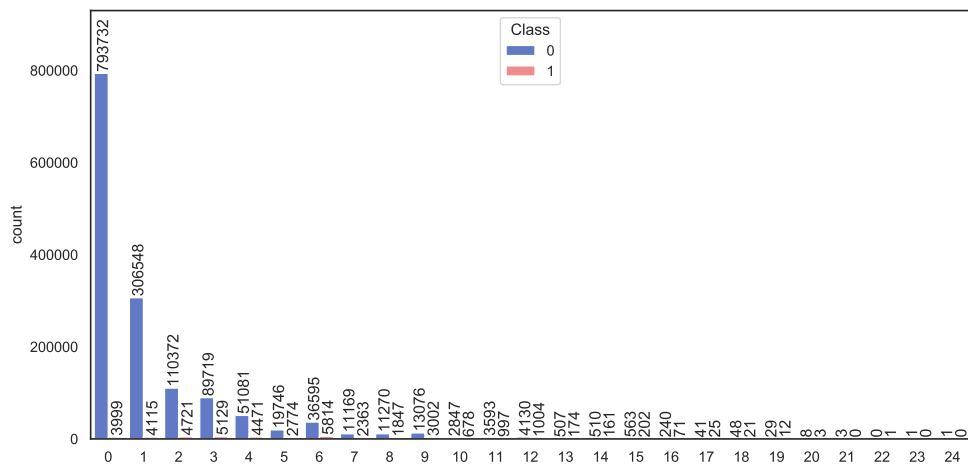


圖 4.9: 觀察 GameTypeE 62 號遊戲之總贏遊戲次數長條圖 ( X 軸為總贏遊戲次數； Y 軸為玩家數量 )

圖 4.10，該圖為 GameTypeE 62 號遊戲之獲得遊戲貨幣 A 之總額，從圖中可以看出，付費玩家資料分佈較廣，而非付費玩家則侷限在 10,000 左右，推測可以帶給學習模型很好的分類資訊。

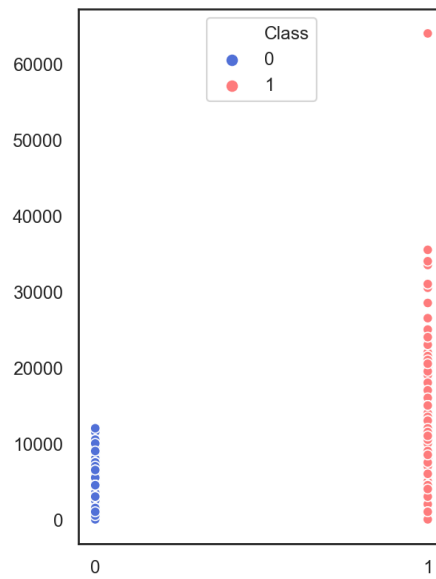


圖 4.10: 觀察 GameTypeE 62 號遊戲之獲得遊戲貨幣 A 之總額散佈圖 (X 軸為非付費玩家與付費玩家；Y 軸為獲得遊戲貨幣 A 之總額)

## 4.4 機器學習評估

此階段將評估前章 3.3.1 小節之分割訓練與測試資料集處理、3.3.2 小節之學習模型選擇與 3.3.3 小節之資料不平衡處理及 3.3.4 小節之搜尋最佳參數解處理與 3.3.5 小節之交叉驗證處理，分別於 4.4.1 小節 分割訓練與測試資料集評估、4.4.2 小節 資料不平衡處理評估及 4.4.3 小節 最佳模型評估說明。

以下實驗環境將透過 *pandas*、*scikit-learn* 以及 *XGBoost* 來進行機器學習之訓練與驗證，並透過 *Seaborn* 協助呈現驗證結果。

#### 4.4.1 分割訓練與測試資料集評估

將資料集依照 7：3 之比例分割，如圖 3.13，即為前 7 週設為訓練集；後 3 週設為測試集，圖 4.11 為各週之新進玩家數圖，深色底為付費玩家、淺色底為非付費玩家，從圖中可以看出，每週之付費玩家數及非付費玩家數皆大致相等，以週次分割資料集可以保存資料穩定性。

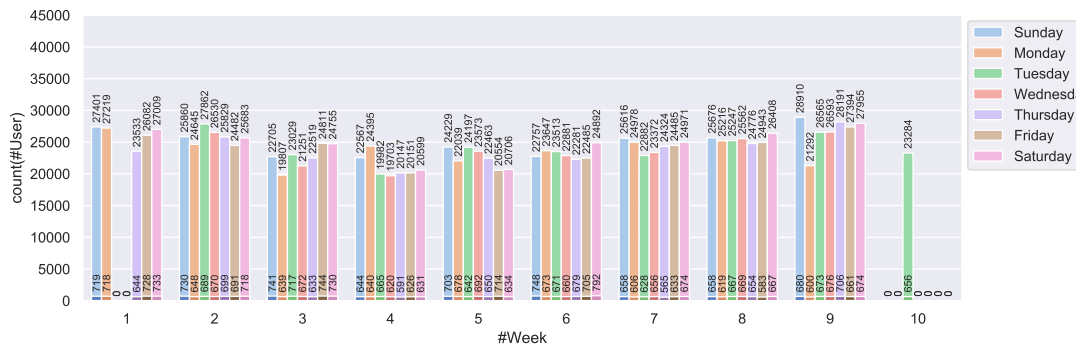


圖 4.11: 各週之新進玩家數圖 (X 軸為週次；Y 軸為玩家數)

表 4.8 為分割完資料集後之付費玩家數與非付費玩家數，訓練資料集與測試資料集之付費玩家與非付費玩家比例皆與原資料集大致相等，分別為 33 倍與 38 倍，原資料集為 35 倍。

資料集 \ 玩家數	玩家數	
	付費玩家	非付費玩家
訓練集	31,741	1,077,660
測試集	9,843	378,169

表 4.8: 訓練與測試資料集玩家數表

#### 4.4.2 資料不平衡處理評估

依照式 3.2 計算付費玩家樣本之放大權重，如式 4.1，最後將付費玩家之樣本權重放大 33 倍。以下將進行學習模型之評估，其中將藉由 *Confusion Matrix*、*Precision*、*Recall*、*True Positive Rate (TPR)* 及 *False Positive Rate (FPR)* 來說明評估，計算方式分別如表 4.9、式 4.2、式 4.3、式 4.4 及式 4.5。

$$class\ 0 : class\ 1 = 1 : \left\lfloor \frac{1,077,660}{31,741} \right\rfloor = 1 : 33 \quad (4.1)$$

	True class 1	True class 0
Predicted class 1	<i>True Positive (TP)</i>	<i>False Positive (FP)</i>
Predicted class 0	<i>False Negative (FN)</i>	<i>True Negative (TN)</i>

表 4.9: Confusion Matrix

$$Precision = \frac{TP}{TP + FP} \quad (4.2)$$

$$Recall = \frac{TP}{TP + FN} \quad (4.3)$$

$$True\ Positive\ Rate\ (TPR) = \frac{TP}{TP + FN} = Recall \quad (4.4)$$

$$False\ Positive\ Rate\ (FPR) = \frac{FP}{TN + FP} \quad (4.5)$$



我們將藉由 Receiver Operating Characteristic Curve ( ROC Curve ) 以及 Precision-Recall Curve ( PR Curve ) 來觀察學習模型間的 *Precision*、*Recall*、*True Positive Rate* ( *TPR* ) 及 *False Positive Rate* ( *FPR* )，如圖 4.12 及圖 4.13。前者於 X 軸及 Y 軸皆以值越大越理想，故曲線越趨近於右上角則越佳；後者於 X 軸為值越小越理想、Y 軸為值越大越理想，故曲線越趨近於左上角則越佳。將再分別利用 AUC 及 AP 來衡量兩曲線，皆為計算該曲線面積。

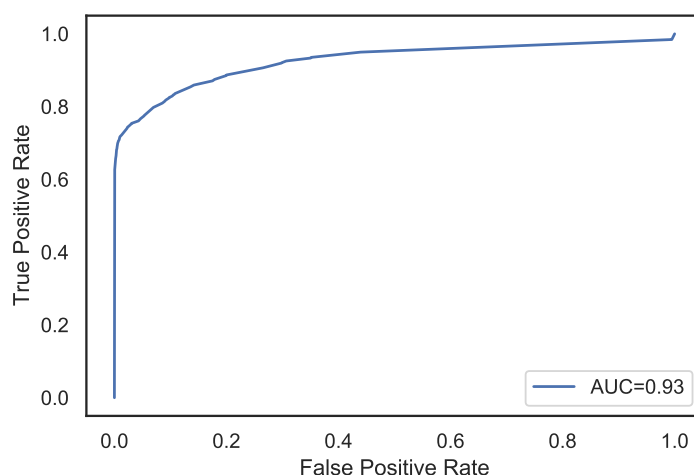


圖 4.12: ROC Curve 示意圖 ( X 軸為 *False Positive Rate* ( *FPR* ) ; Y 軸為 *True Positive Rate* ( *TPR* ) ; AUC 為其曲線面積 )

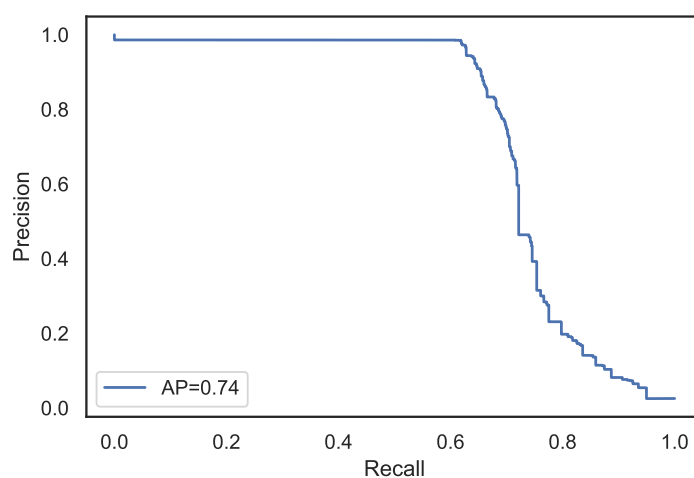
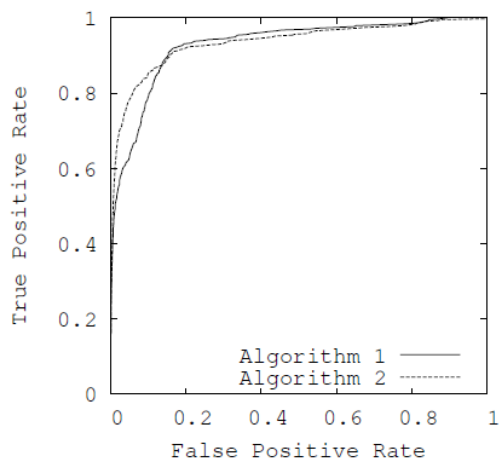
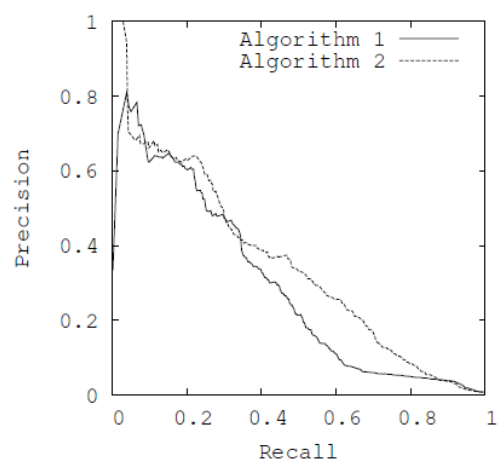


圖 4.13: PR Curve 示意圖 ( X 軸為 *Recall* ; Y 軸為 *Precision* ; AP 為其曲線面積 )

本論文將以評估 PR Curve 為重，因在不平衡資料集上進行評估時，ROC Curve 將無法準確的呈現出學習模型的好壞，常有在 ROC Curve 上表現良好，但其 PR Curve 卻不如預期，導致此情況原因為多數群之評估遠大於少數群之評估，故可在 ROC Curve 上擁有好的數值，卻在 PR Curve 中表現不佳 [42]，如圖 4.14。



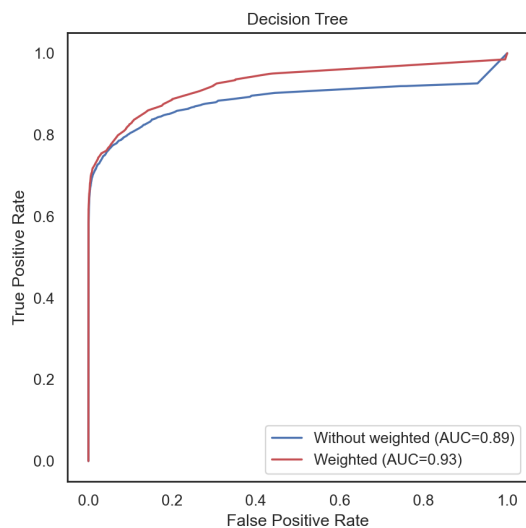
(a) Comparison in ROC space



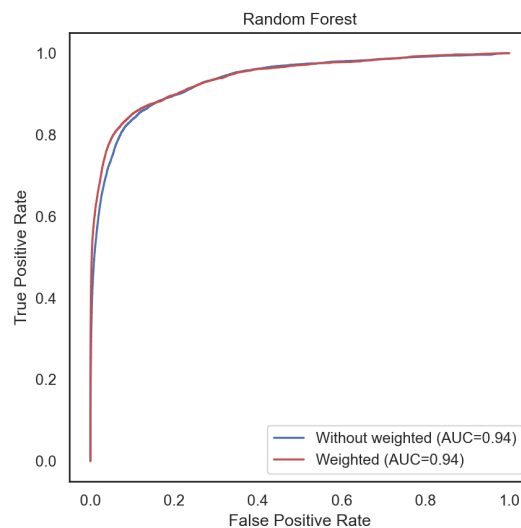
(b) Comparison in PR space

圖 4.14: 不平衡資料中 ROC Curve 失準示意圖 (此圖取自 [42])

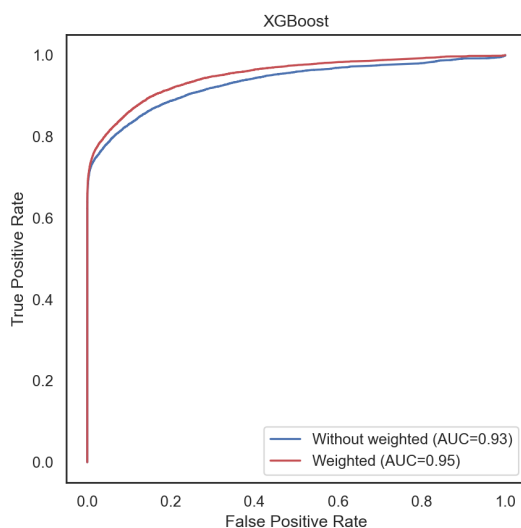
圖 4.15 為三種學習模型之 ROC Curve，並比較不平衡資料處理前後之差異，(a) 為 Decision Tree、(b) 為 Random Forest、(c) 為 XGBoost，藍色線為未加入權重值、紅色底為加入權重值，從圖組中可以看出，在付費玩家之樣本權重上進行放大，有助於學習模型之分類，使預設更加準確。AUC 最高值於 XGBoost 加入權重值，為 0.95。



(a) Decision Tree ROC Curve 圖



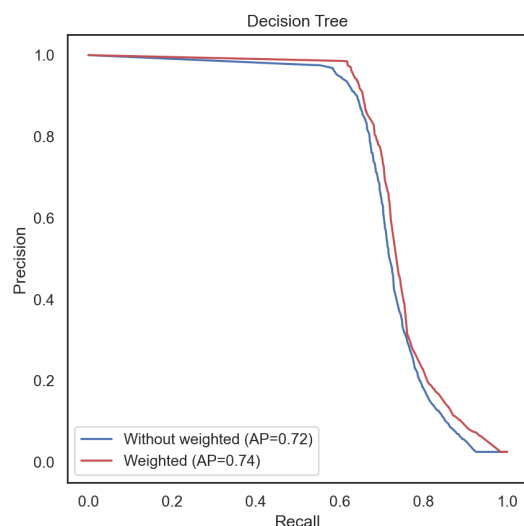
(b) Random Forest ROC Curve 圖



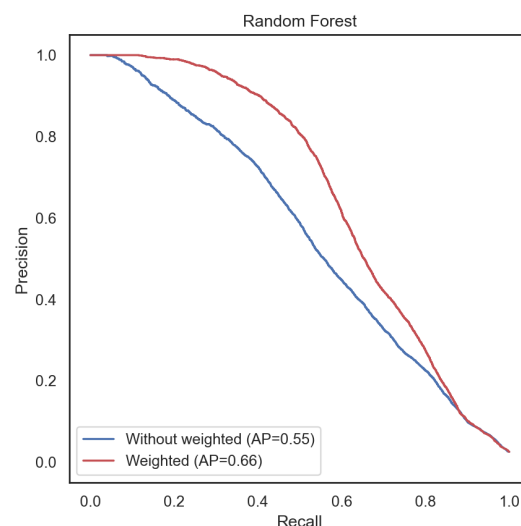
(c) XGBoost ROC Curve 圖

圖 4.15: 不平衡資料處理前後比較之 ROC Curve 圖 ( X 軸為 *False Positive Rate*( *FPR* ) ; Y 軸為 *True Positive Rate*( *TPR* ) )

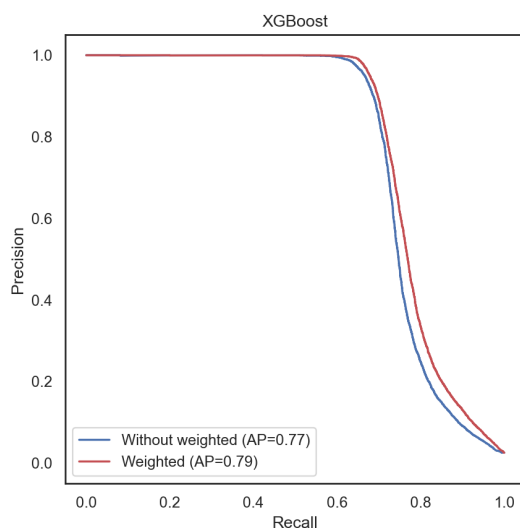
圖 4.16 為三種學習模型之 PR Curve，並比較不平衡資料處理前後之差異，(a) 為 Decision Tree、(b) 為 Random Forest、(c) 為 XGBoost，藍色線為未加入權重值、紅色底為加入權重值，從圖組中可以看出，在付費玩家之樣本權重上進行放大，有助於學習模型之分類，使預設更加準確。AP 最高值於 XGBoost 加入權重值，為 0.79。



(a) Decision Tree PR Curve 圖



(b) Random Forest PR Curve 圖



(c) XGBoost PR Curve 圖

圖 4.16: 不平衡資料處理前後比較之 PR Curve 圖 (X 軸為 *Recall* ; Y 軸為 *Precision* )

上述兩種評估方式皆為在加入權重值後，改進了學習模型的訓練，使其不受於資料不平衡之影響，且適用於三種學習模型。

### 4.4.3 最佳模型評估

利用 Repeated Stratified Cross Validation 來調教出最佳模型，我們將採用 Repeated 2 次及 5-Fold Cross Validation，最後使用測試資料集進行評估驗證，其中包含付費玩家 (*class 1*) 9,843 位及非付費玩家 (*class 0*) 378,169 位，驗證結果如表 4.10，從表中可以看出，XGBoost 的  $Weighted F_{beta} - Score$  為三者最高，預測能力最佳。

學習模型 \ 評估	$precision^+$	$recall^+$	$F_{\beta\alpha}^+$	$Weighted F_{\beta\alpha}$
	$precision^-$	$recall^-$	$F_{\beta\alpha}^-$	
Decision Tree	0.700	0.726	0.726	0.985
	0.993	0.992	0.992	
Random Forest	0.840	0.678	0.678	0.989
	0.992	0.997	0.997	
XGBoost	0.965	0.722	0.723	0.992
	0.993	0.999	0.999	
+：以正例 (付費玩家 $class\ 1$ ) 為評估對象進行計算				
-：以反例 (非付費玩家 $class\ 0$ ) 為評估對象進行計算				

表 4.10: 最佳模型評估表

圖 4.17 及圖 4.18 為三種學習模型之 ROC Curve 及 PR Curve 比較圖，並且都為加入權重值之結果，從圖組中可以看出，皆為 XGBoost 擁有最好的結果，綜合上述得到的實驗結果，我們認為 XGBoost 非常適用於遊戲領域巨量資料預測分類上，因其 Boosting 方式建樹，強化修正於分類錯誤樣本，有效的提升學習模型預測之準確度，並採用 Gradient Descent 來加速學習模型之收斂，減少建樹時間成本。

另外，此處實驗之 Random Forest 遭遇了前述所提到 ROC Curve 表現佳，卻在 PR Curve 表現差的問題，甚至評估結果差於基礎學習模型 Decision Tree，有可能是因學習模型有過擬合 (Overfitting) 的情形發生，所導致 PR Curve 表現不佳。

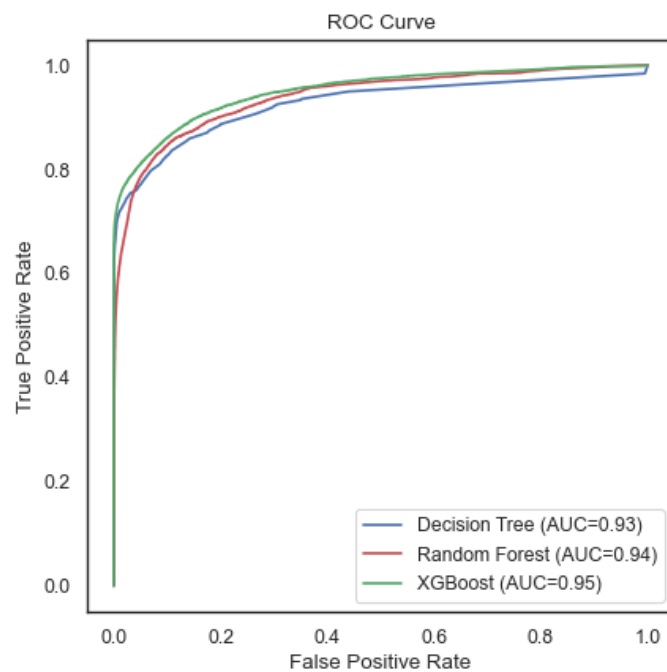


圖 4.17: 三種學習模型之 ROC Curve 比較圖

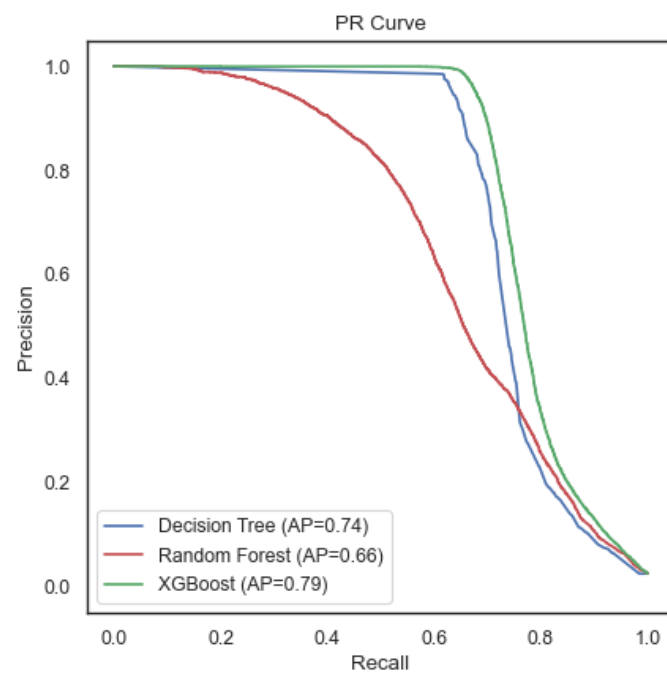


圖 4.18: 三種學習模型之 PR Curve 比較圖

表 4.11 為三種學習模型之最佳參數解，參數搜尋範圍如表 4.12，Decision Tree 因其為單樹結構，相較之下需要生成更深的樹；而 Random Forest 及 XGBoost 則因其為多樹結構，希望能以廣度發展，而非深度，相較之下需要生成更多的樹。

學習模型	Decision Tree	Random Forest	XGBoost
參數調教	max_depth=13	n_estimators=55	n_estimators=55
	min_samples_split=2	max_depth=13	max_depth=10
	min_samples_leaf=5	min_samples_split=2	
		min_samples_leaf=5	

表 4.11: 最佳模型參數解表

參數名	搜尋範圍
n_estimators	20, 25, 30, 35, 40, 45, 50, 55, 60
max_depth	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15
min_samples_split	2, 4, 6, 8, 10
min_samples_leaf	1, 5, 10, 15, 20

表 4.12: 參數搜尋範圍表

## 4.5 預測結果分析評估

此階段將評估前章 3.4.1 小節之資料特徵重要性分析處理，於 4.5.1 小節資料特徵重要性評估說明。

### 4.5.1 資料特徵重要性評估

將利用式 3.5、式 3.6 及式 3.7 計算之各資料特徵於各模型之資料特徵重要性 (*Feature Importance,  $f_i$* )，如圖 4.19、圖 4.20 及圖 4.21，分別為 Decision Tree、Random Forest 及 XGBoost 之資料特徵重要性比較圖。

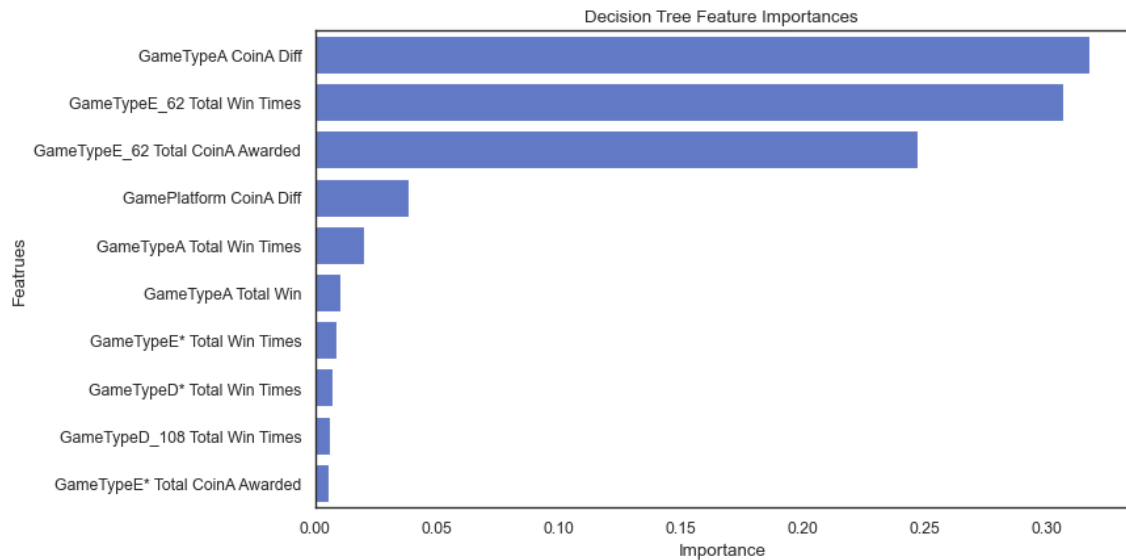


圖 4.19: Decision Tree 資料特徵重要性比較圖



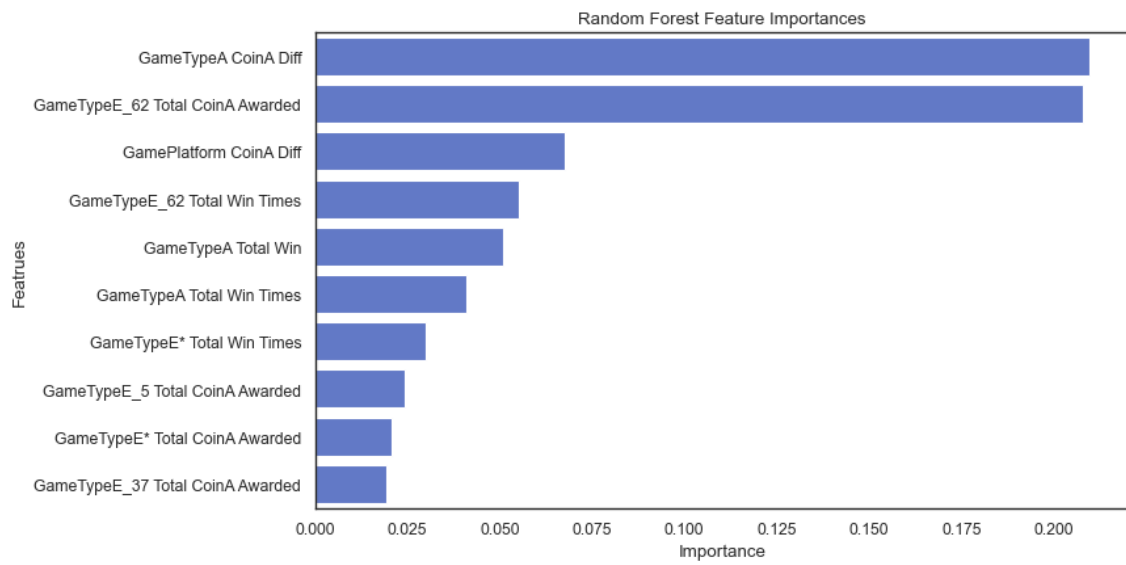


圖 4.20: Random Forest 資料特徵重要性比較圖

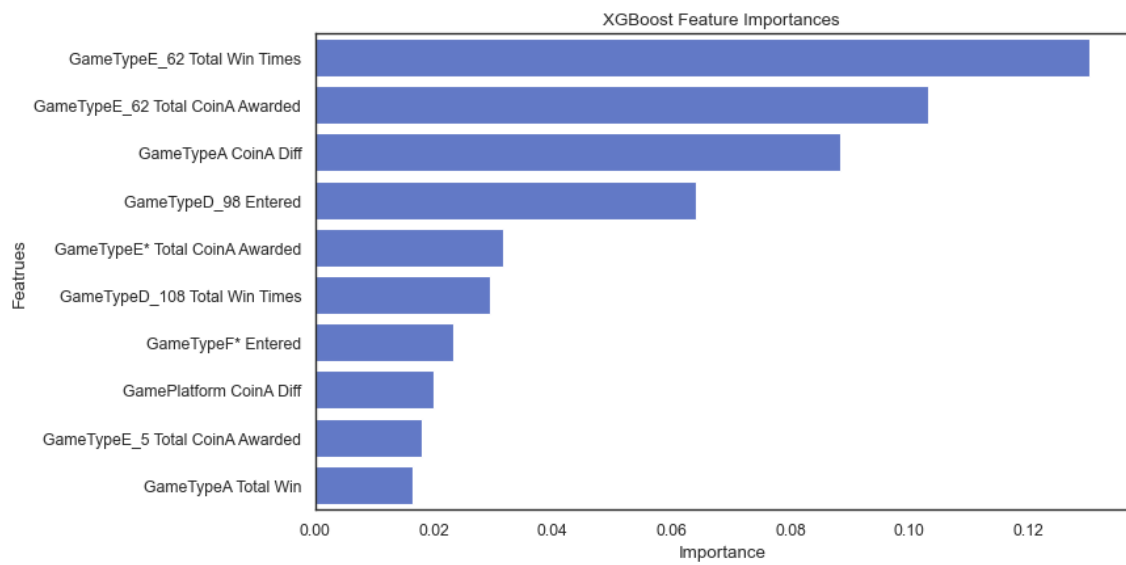


圖 4.21: XGBoost 資料特徵重要性比較圖

從三圖中可以看出，GameTypeE 62 號遊戲的獲得遊戲貨幣 A 之總額以及總贏遊戲次數皆在三種學習模型的前四名，可以說明此款遊戲在於玩家獲得獎勵及贏得遊戲時，對於其付費意願有明顯提升；GameTypeA 的遊戲貨幣 A 之餘額變化以及總贏分皆在前十名，可以說明此款遊戲在於玩家遊玩遊戲時的體驗起伏（無論輸或贏）及贏取分數時，對於其付費意願有明顯提升。

上述所提到的資料特徵皆在 3.2.1.2 小節中有進行推測該類資料特徵將有助於學習模型訓練，顯示先對資料進行分析，對於學習模型之解釋以及後續之利用是有相當程度上的幫助，可以透過探索性資料分析，在資料集應用於機器學習前，即先對資料進行探索，找出資料之問題或是高資訊量的資料特徵，進而增進學習模型的成效或是加強資料特徵的轉化。

另外，三種學習模型之前十名資料特徵皆為數值型的資料特徵，而無類別型資料特徵，因其資料特徵重要性之評估以計算 *Gini Importance* 為主，數值型將會比類別型來得更為顯著，未來將可在計算重要性分析中，對於不同類型的資料特徵加入權重值，使得類別型的資料特徵能夠突出，讓整體分析更加準確。

## 第 5 章 結論與未來研究

### 5.1 結論

本論文提出一巨量資料探勘框架，框架拆分為四大階段進行，其中包括資料前處理、資料分析以及機器學習訓練等步驟。透過此框架可了解到資料前處理中的整合資料需求，並在清理資料時，過濾掉無價值的玩家，以提高整體分析成效，並且在目標值準備與資料特徵探勘時，給予一時間門檻，定義出合理的資料集，使得後續機器學習更加順利；藉由探索性資料分析處理資料集中不合適的資料特徵，並提早推測有助於遊戲平台發展的資料特徵；經由加入權重值於少數群，解決遊戲領域遭遇到的資料不平衡問題，而不透過修改原資料集的內容，再利用學習模型所輸出的預測結果，預測出付費玩家；最後以資料特徵重要性分析與藉由前述之推測，整理出資料特徵突出之原因，例如：玩家於遊戲體驗之起伏、玩家獲得獎勵或贏得遊戲所提高其付費意願。我們提出的框架可應用於遊戲領域且著重於新進玩家並經實驗證實，藉由前處理後的資料集，並同時針對資料不平衡進行處理，在預測潛在之新進付費玩家上有不錯的表現！並且我們可以進一步分析各資料特徵的重要性，協助解釋預測結果與遊戲內玩家行為軌跡的連動性，將可在遊戲平台行銷策略上提供意見，以更符合玩家真實環境所需，提高往後玩家之消費意願。

### 5.2 未來研究

由於本論文的資料特徵探勘種類受限於遊戲平台所提供的原始資料集，如能獲得更進一步的詳細資料，將可更加準確的預測出付費玩家。此外我們使用到的學習模型只有三種樹狀結構之模型，未來將可進行更多實驗於不同類型的學習模型，甚至導入時間序的概念，了解到玩家遊玩遊戲的順序差異是否會影響消費意願。最後除了可預測玩家是否會付費之外，未來還可進行付費時間點、購買商品種類等之預測，尚有許多議題可於遊戲領域巨量資料中研究。

## 参 考 文 献

- [1] A. Drachen, C. Thureau, J. Togelius, G. N. Yannakakis, and C. Bauckhage, “Game data mining,” in *Game analytics*, pp. 205–253, Springer, 2013.
- [2] P. Miller, “Gdc 2012: How valve made team fortress 2 free-to-play,” *Gamasutra*. *Haettu*, vol. 7, p. 2012, 2012.
- [3] E. Lee, Y. Jang, D.-M. Yoon, J. Jeon, S.-i. Yang, S.-K. Lee, D.-W. Kim, P. P. Chen, A. Guitart, P. Bertens, f. Periañez, F. Hadiji, M. Müller, Y. Joo, j. Lee, I. Hwang, and K.-J. Kim, “Game data mining competition on churn prediction and survival analysis using commercial game log data,” *IEEE Transactions on Games*, vol. 11, no. 3, pp. 215–226, 2018.
- [4] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [5] D. Powers, “Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation,” *Mach. Learn. Technol.*, vol. 2, 01 2008.
- [6] C. Goutte and É. Gaussier, “A probabilistic interpretation of precision, recall and f-score, with implication for evaluation,” in *ECIR*, 2005.
- [7] Wikipedia contributors, “Free-to-play — Wikipedia, the free encyclopedia.” <https://en.wikipedia.org/w/index.php?title=Free-to-play&oldid=965292994>, 2020. [Online; accessed 10-July-2020].
- [8] R. Flunger, A. Mladenow, and C. Strauss, “Game analytics on free to play,” in *Big Data Innovations and Applications* (M. Younas, I. Awan, and S. Benbernou, eds.), (Cham), pp. 133–141, Springer International Publishing, 2019.
- [9] S.-K. Lee, S.-J. Hong, S.-I. Yang, and H. Lee, “Predicting churn in mobile free-to-play games,” in *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 1046–1048, IEEE, 2016.
- [10] R. Sifa, F. Hadiji, J. Runge, A. Drachen, K. Kersting, and C. Bauckhage, “Predicting purchase decisions in mobile free-to-play games,” in *Eleventh Artificial Intelligence and Interactive Digital Entertainment Conference*, 2015.

- [11] M. Tamassia, W. Raffè, R. Sifa, A. Drachen, F. Zambetta, and M. Hitchens, “Predicting player churn in destiny: A hidden markov models approach to predicting player departure in a major online game,” in *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8, IEEE, 2016.
- [12] Á. Periañez, A. Saas, A. Guitart, and C. Magne, “Churn prediction in mobile social games: Towards a complete assessment using survival ensembles,” in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 564–573, IEEE, 2016.
- [13] J. Runge, P. Gao, F. Garcin, and B. Faltings, “Churn prediction for high-value players in casual social games,” in *2014 IEEE conference on Computational Intelligence and Games*, pp. 1–8, IEEE, 2014.
- [14] A. Martínez, C. Schmuck, S. Pereverzyev Jr, C. Pirker, and M. Haltmeier, “A machine learning framework for customer purchase prediction in the non-contractual setting,” *European Journal of Operational Research*, vol. 281, no. 3, pp. 588–596, 2020.
- [15] F. Hadiji, R. Sifa, A. Drachen, C. Thureau, K. Kersting, and C. Bauckhage, “Predicting player churn in the wild,” in *2014 IEEE Conference on Computational Intelligence and Games*, pp. 1–8, 2014.
- [16] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [17] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [18] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.
- [19] Y. Freund, R. Schapire, and N. Abe, “A short introduction to boosting,” *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.
- [20] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

- [21] T. Huang, “機器學習: Ensemble learning 之 bagging、boosting 和 adaboost.” <https://medium.com/@chih.sheng.huang821/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92-ensemble-learning%E4%B9%8Bbagging-boosting%E5%92%8CadaBoost-af031229ebc3>. [Online; accessed 10-July-2020].
- [22] A. Semenov, P. Romov, S. Korolev, D. Yashkov, and K. Neklyudov, “Performance of machine learning algorithms in predicting game outcome from drafts in dota 2,” in *International Conference on Analysis of Images, Social Networks and Texts*, pp. 26–37, Springer, 2016.
- [23] A. Janusz, T. Tajmayer, and M. Świechowski, “Helping ai to play hearthstone: Aaia’17 data mining challenge,” in *2017 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pp. 121–125, IEEE, 2017.
- [24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “Smote: synthetic minority over-sampling technique,” *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.
- [25] N. Chinchor and B. M. Sundheim, “Muc-5 evaluation metrics,” in *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*, 1993.
- [26] M. Kubat, R. Holte, and S. Matwin, “Learning when negative examples abound,” in *European Conference on Machine Learning*, pp. 146–153, Springer, 1997.
- [27] J. W. Tukey, *Exploratory data analysis*, vol. 2. Reading, MA, 1977.
- [28] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg, “Top 10 algorithms in data mining,” *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.
- [29] N. V. Chawla, “Data mining for imbalanced datasets: An overview,” in *Data mining and knowledge discovery handbook*, pp. 875–886, Springer, 2009.
- [30] I. Tomek, “Two modifications of cnn,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, 1976.
- [31] M. R. Smith, T. Martinez, and C. Giraud-Carrier, “An instance level analysis of data complexity,” *Machine learning*, vol. 95, no. 2, pp. 225–256, 2014.

- [32] G. E. Batista, A. L. Bazzan, and M. C. Monard, “Balancing training data for automated annotation of keywords: a case study,” in *WOB*, pp. 10–18, 2003.
- [33] X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory undersampling for class-imbalance learning,” *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, 2008.
- [34] J. Brownlee, *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.
- [35] M. Armbrust, R. S. Xin, C. Lian, Y. Huai, D. Liu, J. K. Bradley, X. Meng, T. Kaftan, M. J. Franklin, A. Ghodsi, and M. Zaharia, “Spark sql: Relational data processing in spark,” in *Proceedings of the 2015 ACM SIGMOD international conference on management of data*, pp. 1383–1394, 2015.
- [36] A. Bilogur, “Missingno: a missing data visualization suite,” *Journal of Open Source Software*, vol. 3, no. 22, p. 547, 2018.
- [37] M. Waskom, O. Botvinnik, J. Ostblom, M. Gelbart, S. Lukauskas, P. Hobson, D. C. Gemperline, T. Augspurger, Y. Halchenko, J. B. Cole, J. Warmenhoven, J. de Ruiter, C. Pye, S. Hoyer, J. Vanderplas, S. Villalba, G. Kunter, E. Quintero, P. Bachant, M. Martin, K. Meyer, C. Swain, A. Miles, T. Brunner, D. O’Kane, T. Yarkoni, M. L. Williams, C. Evans, C. Fitzgerald, and Brian, “mwaskom/seaborn: v0.10.1 (april 2020),” Apr. 2020.
- [38] J. Reback, W. McKinney, jbrockmendel, J. V. den Bossche, T. Augspurger, P. Cloud, gfyong, Sinhrks, A. Klein, M. Roeschke, S. Hawkins, J. Tratner, C. She, W. Ayd, T. Petersen, M. Garcia, J. Schendel, A. Hayden, MomIsBestFriend, V. Jancauskas, P. Battiston, S. Seabold, chris b1, h vetinari, S. Hoyer, W. Overmeire, alimcmaster1, K. Dong, C. Whelan, and M. Mehyar, “pandas-dev/pandas: Pandas 1.0.3,” Mar. 2020.
- [39] Wes McKinney, “Data Structures for Statistical Computing in Python,” in *Proceedings of the 9th Python in Science Conference* (Stéfan van der Walt and Jarrod Millman, eds.), pp. 56 – 61, 2010.
- [40] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Courn-

- peau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [41] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, “API design for machine learning software: experiences from the scikit-learn project,” in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.
- [42] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, 2006.