



國立臺灣科技大學
資訊工程系

碩士學位論文

基於極限梯度提升的新進玩家流失預測模型應用於手機
免費遊戲數據集

New Player Churn Prediction Model Based On Extreme
Gradient Boosting Is Applied to Mobile Free-to-Play Game
Dataset

研 究 生：柯名鴻

學 號：M10915Q05

指導教授：戴文凱博士

中華民國一一年六月三十日

第 1 章 緒論

1.1 研究背景與動機

對於許多遊戲商而言，準確預測玩家流失對於長期成功至關重要。近年來，手機遊戲商大多以免費遊戲為主，不再是以往的買斷制或月費制，在免費遊戲的模式之下，遊戲商之營收有非常顯著的成長，例如：「絕地要塞 2」(Team Fortress 2) 原本為買斷制型式遊戲，新收入的來源僅限於沒購買過遊戲的人，這樣的商業模式與他們的發展策略並不完全吻合。於是，在 2011 年改為免費遊戲，除了擴大玩家受眾，還透過遊戲不斷地更新保持玩家的興趣，最終，遊戲營收提高達 12 倍之多 [1]。但是，免費遊戲在定義是否為流失玩家上極為困難，一旦玩家想離開了隨時都能停止遊玩，沒有必要告訴遊戲商其決定，遊戲商也容易因此失去挽回玩家的機會。

根據哈佛商業評論 [2] 中指出，企業爭取一個新顧客的成本是保留老顧客成本的 5 倍，一個公司如果能將其顧客流失率降低 5%，其利潤就能增加 25% - 85%。也就是說，獲得一個新玩家的成本比留住一個玩家要昂貴許多。因此，留住玩家成為重要議題，精準地預測玩家流失，即使是微小地提升，也可能導致營收有顯著的提升。

另外，根據網站 Swrve [3] 於 2014 年提供的報告指出，數十款的遊戲中，有 19.3% 的新玩家只玩一次特定遊戲，新玩家的次日留存率為 33.9%，而第 30 天的留存率只剩下 5.5%。而 Mustač 等人 [4] 則是針對歐洲一款休閒遊戲進行研究，研究中可以看到，此遊戲有大約 60% 的玩家在玩了一天後就離開了，3 天後，玩家們只剩下大約 20%。上述的 2 份報告都清楚表明了新進玩家的流失率是很嚴重的問題，因此，本文將針對新進玩家做流失預測分析，希望能因此提高新進玩家的留存率，進而有效地增加營收。

1.2 研究目標

本論文的研究目標是預測免費遊戲的新進玩家是否會流失。透過新進玩家在遊戲初期的遊玩歷程和儲值紀錄等特徵訓練出準確率最高的模型，並藉由模型的

結果分析個個特徵的突出性，來進行玩家流失原因的解釋說明，甚至能從中了解玩家們的喜好、趨勢，也可以讓市場操作人員有挽留玩家的操作依據，有了明確的挽留策略就能夠對症下藥，藉此來強化留存，並進一步增加營收。

1.3 研究方法概述

本論文將新進玩家創帳號後的天數切分為三個時期：(1) 觀察期：玩家創帳號前後幾天，會將此時期的玩家行為軌跡作為資料特徵來訓練模型；(2) 挽留期：於觀察期之後，作為給市場操作人員實施挽留策略的時間；(3) 表現期：於挽留期之後，決定玩家是否流失，如果玩家在此時期有任一登入紀錄則視為非流失玩家，反之將視為流失玩家。

此外，本文還運用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：首先從資料庫群中整合所有需要的資料，並過濾掉無價值玩家，再著手目標值準備、資料特徵探勘與特徵工程，以利後續分析及機器學習使用；(2) 資料分析階段：使用前階段產出之資料，透過統計圖表來觀察資料特性，進行探索性資料分析 (Exploratory Data Analysis) [5]，藉由流失玩家與非流失玩家的資料分佈來檢查是否有不合適之資料特徵，並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：首先將處理後的資料集分割為訓練集及測試集，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，並透過交叉驗證 (Cross Validation) 找出機器學習模型的最佳超參數以獲得最佳模型，其中學習模型選用決策樹 (Decision Tree)、隨機森林 (Random Forest) 及極限梯度提升 (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基尼重要性 (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，本論文將藉由混淆矩陣 (Confusion Matrix) 所延伸之接收者操作特徵曲線 (Receiver Operating Characteristic Curve) [6] 與精確召回曲線 (Precision-Recall Curve) [7] 來協助驗證學習模型之優劣，並同時利用 Weighted F_β - Score [8] 來選出最佳模型與最佳參數解，隨後計算特徵重要性 (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. 訓練一新進玩家流失預測模型並運用於遊戲領域，利用其預測結果從中了解玩家流失的原因與動機，並作為市場操作人員後續挽留玩家的操作依據，以強化留存並進一步增加營收。
2. 提出一資料特徵工程的方法：對資料集以統計手法建立資料特徵，並用多個時間框架做拆分，以獲得第一層特徵變數，再對第一層特徵變數做計算，進一步來得到第二層特徵變數，如變化量特徵等。
3. 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流失玩家的資料。
4. 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。
5. 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。

1.5 本論文之章節結構

第 2 章 文獻探討

本章節針對免費遊戲興起介紹，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 免費遊戲興起

免費遊戲，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。在免費遊戲中，遊戲商可以藉由遊戲內購買或遊戲內置入廣告等方式來賺取營收 [9]。近幾年內，遊戲商皆轉以開發免費遊戲為主，因其類型所帶來之營收，已遠大於付費型遊戲 [10]。另外，免費遊戲還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 [11]。

雖然免費遊戲能讓玩家流失量降低，但相對的，對於玩家是否流失變得極難定義，因為玩家可以在沒有任何通知的情況下停止遊戲，沒有明確的流失事件，例如玩家取消訂閱。儘管玩家流失模型在商業領域已經存在了幾十年，但隨著機器學習方法近年來的進步，它們的複雜性和準確性也都在提升，因此，即使在高維數據上，也可以應用極限梯度提升等機器學習來創建非常準確的模型 [12]。

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 [4] [10] [12]。

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 [13]、Periáñez 等人只收集消費金額遠高於一般玩家者 [14] 或 Runge 等人只取付費玩家中前 10 % 者 [15]，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 [16] 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。Xie 等人 [17] 使用遊戲內的事件發生頻率來預測玩家對遊戲的參與度。Lee 等人 [18] 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，主要針對玩家每日於遊戲內的行為軌跡。Gregory [12] 透過相對時間及絕對時間建立新的特徵，以提升資料集的特徵數量。Hadiji 等人 [19] 將探勘玩家消費商品數量、玩家遊玩天數等。

2.3 學習模型選擇

在機器學習中，於分類預測的應用上，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 [10]。另外，在樹狀結構之學習模型中，主流以裝袋算法 (Bagging) 及提升方法 (Boosting) 兩種建樹想法為準：

- 裝袋算法：從訓練資料集中，隨機取樣並訓練成多份分類器，而每次訓練時會將資料取出後放回，並再次抽取，最後的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 [20]，如圖 2.1。例如：隨機森林 [21] 即為裝袋算法 + 決策樹 [22]。
- 提升方法：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 [23]，如圖 2.2。例如：極限梯度提升 [24] 即為提升方法 + 決策樹。

Chen 與 Guestrin [24] 實作出高效率的梯度提升 (Gradient Boosting)，稱其為極限梯度提升，除了使用提升方法建樹外，還針對錯誤修正的步驟，引入梯度下降法 (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。近年來透過極限梯度提升來訓練的研究越來越多，且其預測能力皆有不錯的表現 [12] [25] [26] [27]，明顯優於裝袋算法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，且其中使用極限梯度提升之成效最佳。因此，為求本論文之預測新進玩家流失能夠達到預期，

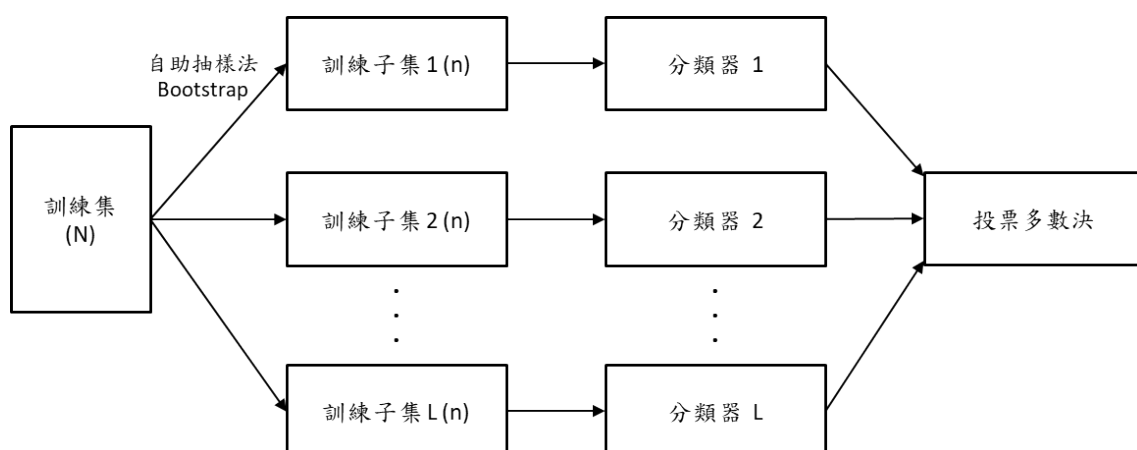


圖 2.1: 裝袋算法方式建樹示意圖

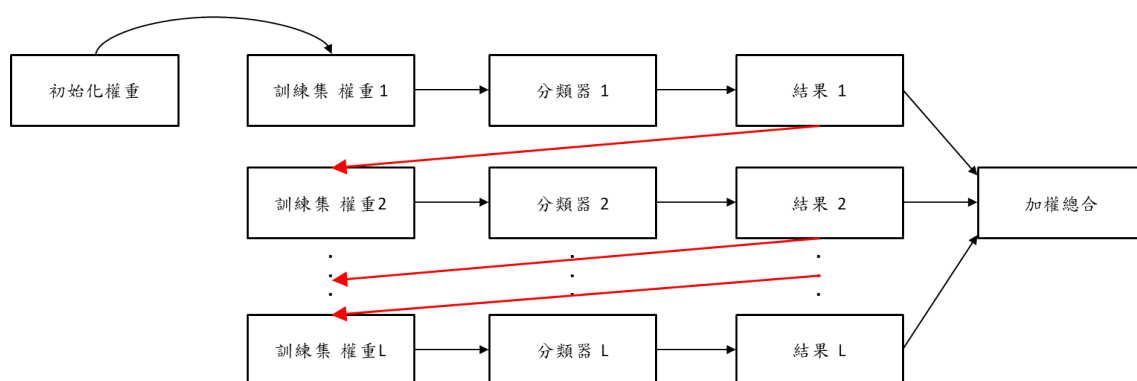


圖 2.2: 提升方法方式建樹示意圖

將採用決策樹、隨機森林與極限梯度提升來驗證樹狀結構之優勢以及極限梯度提升之最佳表現。

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：於預測是否付費上，非付費玩家會遠多於付費玩家，導致付費玩家資料過少 [16]。於預測是否流失上，流失玩家會遠多於非流失玩家，導致非流失玩家資料過少 [18]，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 (Synthetic Minority Over-sampling Technique)，於少數群添加模擬資料，使得少數群之樣本數與多數群相等 [28]。而本論文不只預測玩家是否會流失，還需分析其原因，如在資料集中填入模擬資料，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，如同提升方法建樹時，藉由權重值的不同，修正分類錯誤的資訊 [23]。

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 [29]，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 等人額外運用幾何平均數 (Geometric Mean) [30] 來評估學習模型之成效 [16]。藉由上述的概念，本論文將採用 Weighted F_{β} - 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_{β} - 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新進玩家流失預測，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，本論文對此議題運用一巨量資料探勘框架，圖 3.1 為巨量資料探勘框架示意圖，此框架將由五大階段組成：

- 資料前處理階段：首先將從資料庫群中整合所有所需資料，並過濾出有價值之原始資料，再著手目標值準備、資料特徵探勘與特徵工程，以利後續分析及機器學習使用。
- 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，觀察資料特徵是否可以提供給學習模型較多之資訊。
- 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交叉驗證搭配參數表，以獲得模型的最佳參數解，最後藉由測試集來驗證評估最佳模型，產出預測結果。
- 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。
- 產業應用分析階段：運用代理人模型 (Surrogate Model)。

3.1 資料前處理階段

此階段將著重於資料之整合與過濾，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、資料特徵的探勘與資料特徵工程，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。

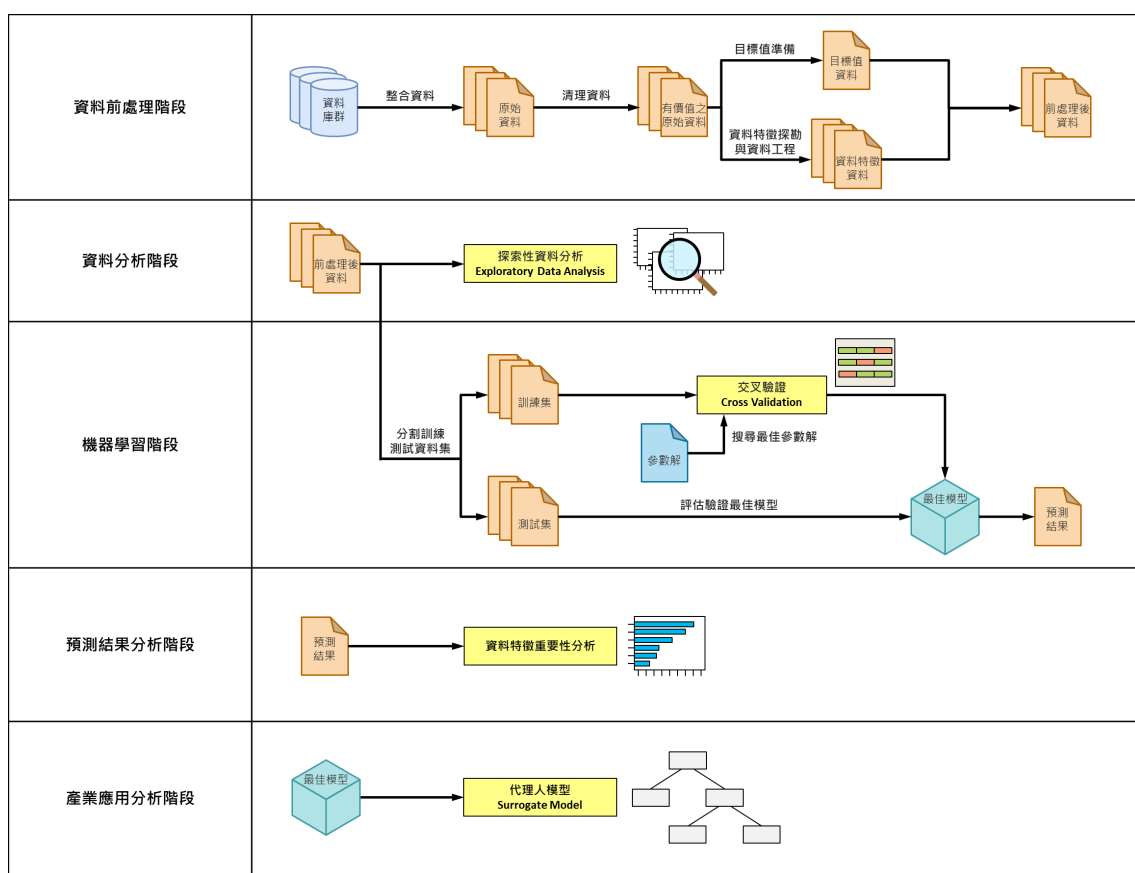


圖 3.1: 本論文之巨量資料探勘框架示意圖

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，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。

除了上述之玩家遊戲行為軌跡原始資料集外，還另外收集了玩家輪廓資料(含國家、玩家等級等)與玩家平台操作紀錄(含消費紀錄、客訴紀錄等)，最終此步驟將產出三大類原始資料集(圖 3.4)：

- 玩家輪廓資料(含國家、玩家等級等)
- 玩家平台操作紀錄(含消費紀錄、客訴紀錄等)
- 玩家遊戲行為軌跡

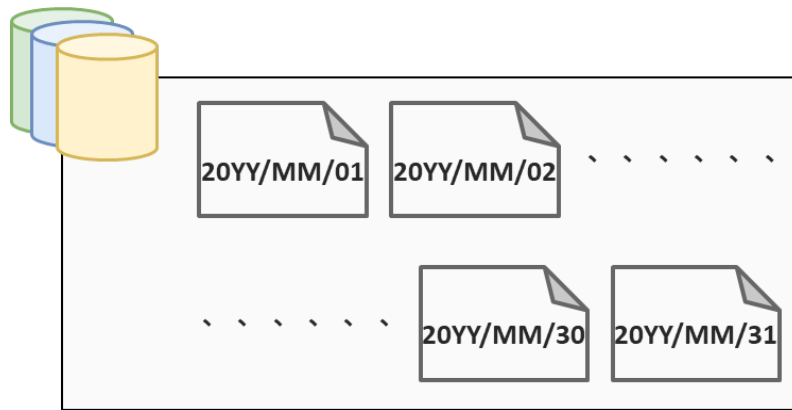


圖 3.2: 資料庫群內資料之示意圖

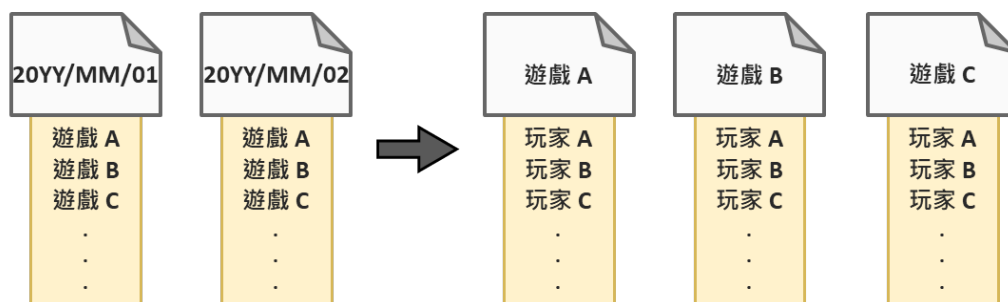


圖 3.3: 依各項遊戲為整合目標之示意圖

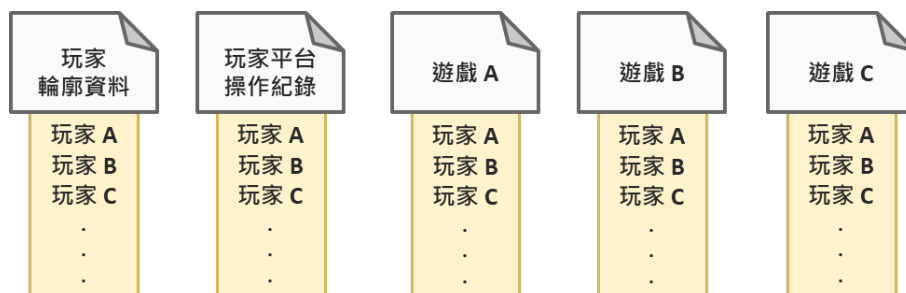


圖 3.4: 原始資料集示意圖

3.1.2 資料過濾

此步驟將針對兩大議題：刪除空缺值與無價值玩家資料處理。


為了要在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與是否流失，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 刪除空缺值

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樣本的方式，而不對資料集填入經過處理之假數值，每筆樣本只要在任意資料特徵中擁有一空缺值，即視為欲刪除之對象。

圖 3.5 為刪除空缺值之示意圖，從圖中可以看出樣本 ID 1 及 2 分別在特徵 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	特徵 1	特徵 2	特徵 3	特徵 4
0	12.5	198	TW	97
1	6.8	1300	-	70
2	-	1788	US	100
3	45.2	699	JP	65



ID	特徵 1	特徵 2	特徵 3	特徵 4
0	12.5	198	TW	97
3	45.2	699	JP	65

圖 3.5: 刪除空缺值示意圖

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 [10]，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。定義一時間框架於新進玩家創帳號後，又將其切分為三個時期：

- 觀察期：玩家創立帳號後前 O 天。觀察玩家在此時期的行為軌跡，並將對其進行特徵工程，因此，也將觀察期視為資料特徵探勘期。

- 挽留期：觀察期之後前 R 天。作為市場操作人員實施挽留策略的時間。
- 表現期：挽留期之後前 P 天。決定玩家是否流失，於玩家在觀察期有登入紀錄的前提下，如果玩家在此時期有任一登入紀錄則視為非流失玩家，反之將視為流失玩家。

如果該玩家創建帳號的日子較晚，尚未完整擁有上述三個時期的資料，將會造成後續特徵提取與目標值準備的不正確性，進而影響機器學習預測的準確度，因此將其視為無價值玩家，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。從圖中可以看出，玩家 1 及玩家 2 已在創帳號後完整經歷觀察期、挽留期與表現期，故視為有價值玩家；而玩家 3，因缺少完整的表現期資料，容易被誤判為流失玩家，故視為無價值玩家；玩家 4，除了表現期的資料，觀察期資料也並不完整，若將其視為有價值玩家，其錯誤的資料特徵將會影響機器學習之訓練，故也視為無價值玩家。

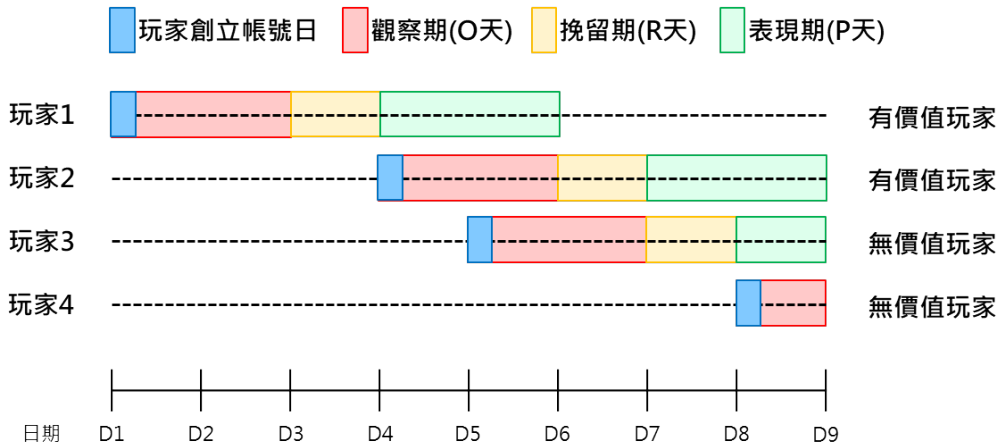


圖 3.6: 判別有價值與無價值玩家之示意圖 (以 O 、 R 、 P 分別為 2、1、2 為例)

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*。定義目標值「非流失玩家」與「流失玩家」，分別代表 *class 0* 及 *class 1*：

- 非流失玩家 (*class 0*)：觀察期有登入紀錄的玩家中，表現期間有登入紀錄者，視為非流失玩家。
- 流失玩家 (*class 1*)：觀察期有登入紀錄的玩家扣除非流失玩家，剩餘者皆為流失玩家；表現期後才有登入紀錄者同樣視為流失玩家。

圖 3.7 為定義非流失玩家與流失玩家之示意圖。從圖中可以看出玩家 1 及玩家 2 於觀察期及表現期中皆有登入紀錄，故定義為非流失玩家 (*class 0*)；而玩家 3 及玩家 4 則在表現期中無登入紀錄，故定義為流失玩家 (*class 1*)，即使玩家 4 在表現期後有登入紀錄，依舊將其視為流失玩家 (*class 1*)。

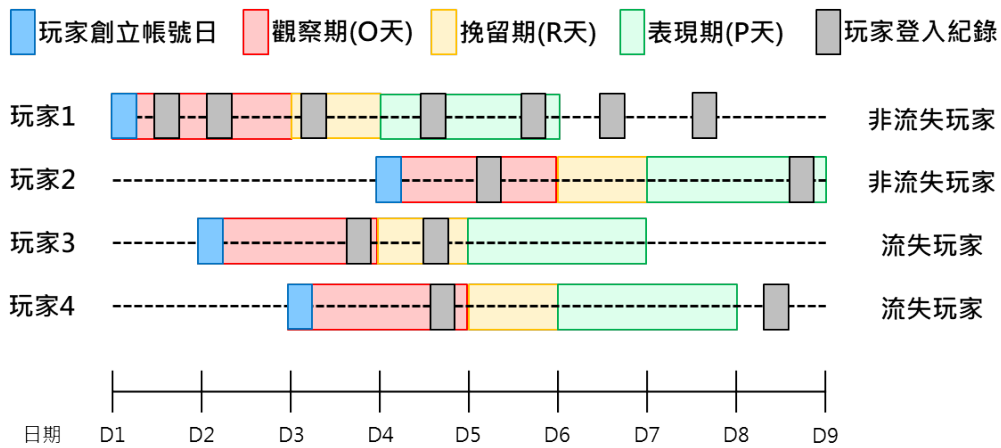


圖 3.7: 非流失玩家與流失玩家之示意圖 (以 O 、 R 、 P 分別為 2、1、2 為例)

本論文將預測目標聚焦於流失的新進玩家，所以我們透過觀察期與表現期來侷限流失玩家之定義。圖 3.8 為流失玩家與非流失玩家范氏圖，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流失玩家 (*class 0*)，內圈之綠圓框代表前述流失定義門檻，綠圓框內之紅色底圓型範圍則代表所有流失玩家 (*class 1*)，其中深紅色底之圓形範圍代表表現期後無登入紀錄的玩家；淺紅色底之圓形範圍代表表現期後有登入紀錄的玩家。可以由上述說明來了解到資料內流失玩家與非流失玩家之分佈狀況及其關係。

3.1.4 資料特徵探勘與特徵工程

此步驟將探勘供機器學習使用之資料特徵。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

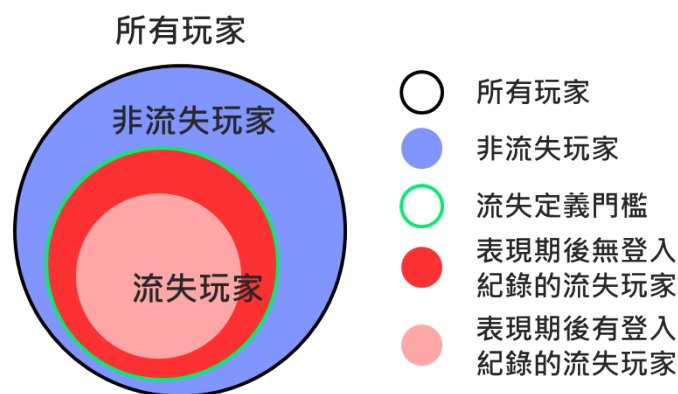


圖 3.8: 流失玩家與非流失玩家范氏圖

有效 [4] [10] [12]，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。

本文將觀察期作為資料特徵探勘期，對每位玩家進行資料特徵探勘，自玩家創立帳號日之 O 天內，探勘其所需之資料特徵。資料特徵之探勘面向將參考於 [16] [18] [25] 之探勘想法，主要聚焦於玩家之行為軌跡，並將其進行特徵工程與設計綜合指標特徵。

圖 3.9 為特徵工程示意圖。對資料集以多種統計方式建立資料特徵，並用多個時間框架做拆分，以獲得第一層特徵變數，再對第一層特徵變數做計算，進一步來得到第二層特徵變數，如變化量特徵等。

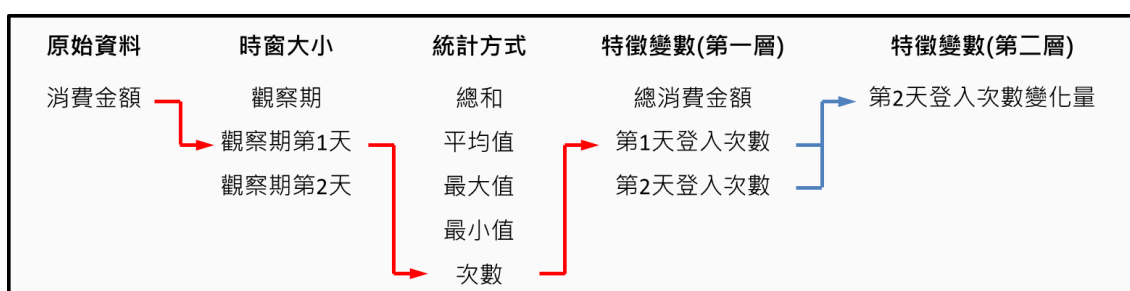


圖 3.9: 特徵工程示意圖

最終我們將資料特徵種類分為三大類：

- 玩家資料：包含玩家自身相關資訊。如創帳號國家、玩家等級等。
- 玩家平台操作紀錄：包含玩家以平台為探勘範疇之行為軌跡。如消費紀錄、客訴紀錄等。

- 玩家遊戲行為軌跡：包含玩家以遊戲為探勘範疇之行為軌跡。如押注次數、贏分等。

3.2 資料分析階段

此階段將著重於資料的分析。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

我們將採用探索性資料分析 [5] 來藉由圖表呈現協助了解資料之特性，並且檢查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：

- 長條圖：觀察資料特徵分布情況，如圖 3.10 (a)。
- 散佈圖：觀察資料特徵間之關聯性，如圖 3.10 (b)。

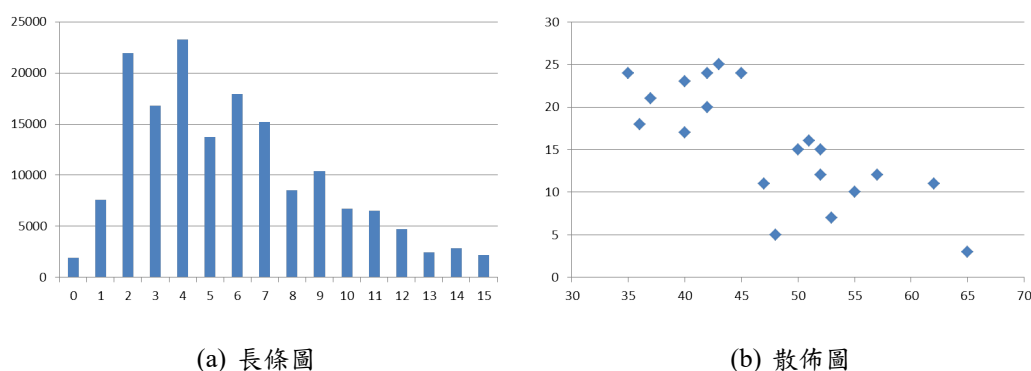


圖 3.10: 探索性資料分析之使用圖表類型

3.2.1.1 高資訊量之資料特徵

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。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。

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，提供給流失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 [10] [16]，而本論文所挑選之學習模型包含：決策樹 [22]、隨機森林 [21] 與極限梯度提升 [24]。

3.3.1 分割訓練與測試資料集

透過前述 ?? 小節排除不合理資料特徵之資料集，進行訓練集與測試集分割，並按照 7：3 之比例隨機分配。為了避免隨機切割時，目標類別分布不平衡，我們採分類隨機抽樣，即流失玩家與非流失玩家各別以 7：3 之比例隨機抽樣，如圖 3.11。

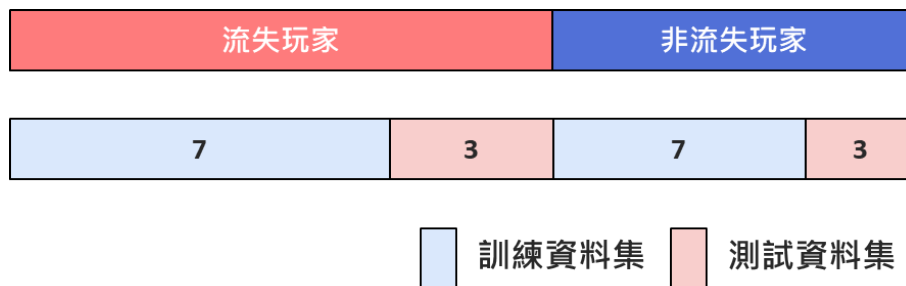


圖 3.11: 分割訓練與測試資料集示意圖

3.3.2 學習模型選擇

我們將選擇樹狀結構的學習模型進行訓練。樹狀結構的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 [10] [16]，是研究中常使用的學習模型 [31]，因其模型結構能夠清楚表現每筆樣本之預測路徑，可以協助我們了解到學習模型如何做出決策，如白箱模型 (white box model)；而

神經網路結構之學習模型，其模型內的結構難以解析，如黑箱模型 (black box model)，無法協助我們進行更一步的研究。而本論文所挑選之學習模型包含：

- 決策樹 [22]：樹狀結構學習模型之最基礎結構，單樹結構，採用 CART (Classification and Regression Tree) 演算法進行建樹。
- 隨機森林 [21]：多樹結構，採用裝袋算法建樹。
- 極限梯度提升 [24]：多樹結構，採用提升方法建樹。

上述三種學習模型之資訊量計算以基尼不純度 (Gini Impurity, G) 為主，如式 3.1，其中 c 為 *Class*、 $p(i)$ 為 c 之發生機率。透過基尼不純度 (G) 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，包含裝袋算法與提升方法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 [16] [18] [32]。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 [33]、InstanceHardnessThreshold [34] 與 RandomUnderSampler；Over-Sampling：SMOTE [28] 與 RandomOverSampler；Combination of Under- and Over-sampling：SMOTETomek [35]；Ensemble：EasyEnsemble [36]。

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 為非流失玩家 (*class 0*) 之樣本數； N_1 為流失玩家 (*class 1*) 之樣本數。將計算 N_0 與 N_1 之比例差距，此值則為非流失玩家 (*class 0*) 樣本權重放大倍數。

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

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 3.1，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 3.2。

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 3.1: 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 3.2: 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 [37] 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 3.12 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中的付費玩家與非付費玩家比例也相等。

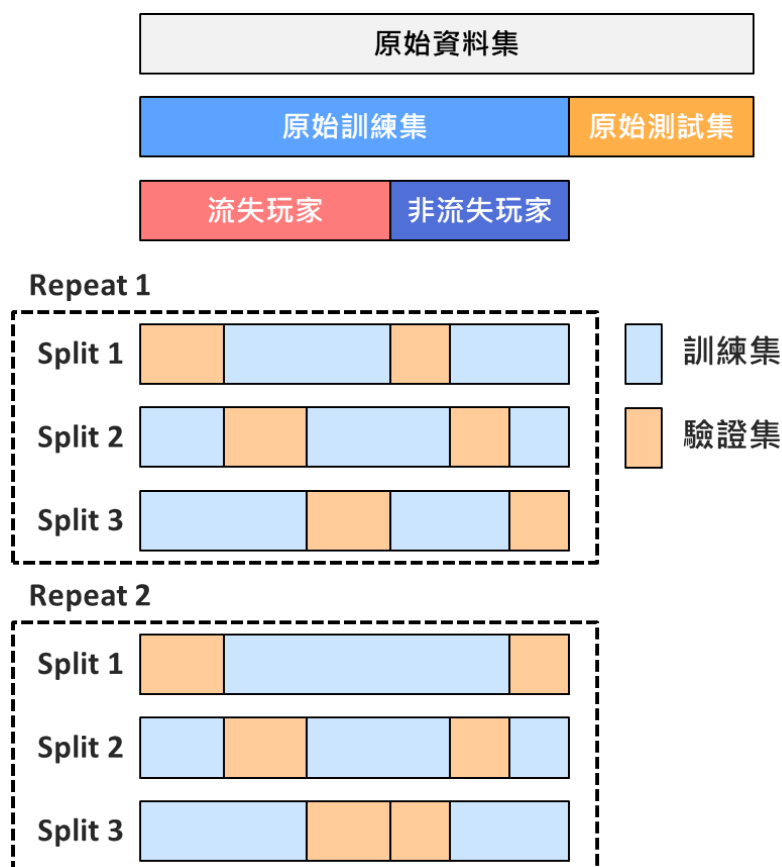


圖 3.12: RepeatedStratifiedKFold 示意圖 (假設 Repeated 為 2；KFold 為 3 時)

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_β - Score，擇其最高值之學習模型，選定為最佳模型。Weighted F_β - Score 為在 F_β - 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 β 則為 Precision 與 Recall 之間的比重，如表 3.3。本論文預測新進玩家是否會流失，將著重於 Recall，即為將所有可能流失的新進玩家預測出來，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流失的新進玩家全部預測出來，使得遊戲商能盡可能地保留住所有玩家。

β 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 3.3: β 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 3.2.1.1 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。

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基尼不純度 (G) (見式 3.1)，稱為基尼重要性 (Gini Importance, GI)，再將其正規化至區間 $[0,1]$ 中。

式 3.5 為計算樹中各節點之基尼重要性 (GI)，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} 之基尼不純度 (G)，並計算 D_{left} 及 D_{right} 之樣本數權重比例，最後將 D_p 之基尼不純度 (G) 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 之基尼重要性 (GI)，並加總 l 之基尼重要性 (GI)，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)。

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

參 考 文 獻

- [1] P. Miller, “Gdc 2012: How valve made team fortress 2 free-to-play,” *Gamasutra. Haettu*, vol. 7, 2012.
- [2] F. Reichheld and W. Sasser, “Zero defects: quality comes to service,” *Harvard Business Review*, September/October, pp. 105–111, 1990.
- [3] Swrve, “The april 2014 new players report.” <https://www.swrve.com/resources/weblog/the-april-2014-new-players-report>, 2014.
- [4] K. Mustač, K. Bačić, L. Skorin-Kapov, and M. Sužnjević, “Predicting player churn of a free-to-play mobile video game using supervised machine learning,” *MDPI*, 2022.
- [5] J. W. Tukey, *Exploratory data analysis*, vol. 2. Reading, MA, 1977.
- [6] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [7] D. Powers, “Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation,” *Mach. Learn. Technol.*, vol. 2, 01 2008.
- [8] C. Goutte and É. Gaussier, “A probabilistic interpretation of precision, recall and f-score, with implication for evaluation,” in *ECIR*, 2005.
- [9] 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].
- [10] 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, Á. 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.
- [11] 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.

- [12] B. Gregory, “Predicting customer churn: Extreme gradient boosting with temporal data,” *arXiv preprint arXiv: 1802.03396*, 2018.
- [13] 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.
- [14] 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.
- [15] 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.
- [16] 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.
- [17] H. Xie, S. Devlin, D. Kudenko, and P. Cowling, “Predicting player disengagement and first purchase with event-frequency based data representation,” in *2015 IEEE conference on Computational Intelligence and Games*, pp. 230–237, IEEE, 2015.
- [18] 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.
- [19] 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.
- [20] L. Breiman, “Bagging predictors,” *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.
- [21] L. Breiman, “Random forests,” *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [22] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and regression trees*. CRC press, 1984.

- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] M. Kubat, R. Holte, and S. Matwin, “Learning when negative examples abound,” in *European Conference on Machine Learning*, pp. 146–153, Springer, 1997.
- [31] 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.
- [32] N. V. Chawla, “Data mining for imbalanced datasets: An overview,” in *Data mining and knowledge discovery handbook*, pp. 875–886, Springer, 2009.

- [33] I. Tomek, “Two modifications of cnn,” *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, 1976.
- [34] 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.
- [35] 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.
- [36] 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.
- [37] J. Brownlee, *Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning*. Machine Learning Mastery, 2020.