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

中華民國一一一年七月二十一日

# 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、秝安、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

# 圖　　目　　錄

# 表　　目　　錄

# 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 |
| $\lfloor\ \rfloor$ | 地板函數 |
| $G(\ )$ | *Gini Impurity* |
| $GI(\ )$ | *Gini Importance* |
| $fi(\ )$ | 資料特徵重要性 ( *Feature Importance* ) |
| $D_p$ | 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

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」原本為買斷制型式遊戲，在 2012 年改為免費遊戲商業模式後，遊戲營收提高達 12 倍之多 [1]，但是，免費遊戲類型的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。此外，在遊戲中，獲得一個新玩家的成本也比留住一個玩家要昂貴許多。

因此，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。透過上述情境發想，若能透過巨量資料探勘訓練出一流失預測模型並運用於遊戲領域，利用其預測結果了解玩家流失的原因與動機，將可以交由運營人員作為後續玩家挽留的操作依據。

## 1.2　研究目標

隨著免費遊戲類型的遊戲客戶獲取成本不斷提高，留住玩家成為重要議題。另外，根據網站 Swrve 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% [2]，可以看出新進玩家的流失率極高。因此，若能運用巨量資料探勘框架，來協助預測新進玩家是否流失，將可以針對可能流失玩家作為挽留的目標並進行操作。

本論文的研究目標為預測新進玩家是否流失與了解玩家流失的原因。我們將運用一巨量資料探勘框架，包含對於資料集之前處理，透過特徵工程建立大量特徵，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，來預測出可能流失的新進玩家，最後依其機器學習預測結果，分析各個特徵中的突出性，來進行新進玩家流失原因的解釋說明。

## 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：玩家創帳號後前幾天，會將此時期的玩家遊戲軌跡作為資料特徵來訓練模型；(2) 挽留期：於觀察期之後，作為給運營操作的時間；(3) 表現期：於挽留期之後，主要決定玩家是否流失，如果玩家在此時期未登入則視為流失玩家，反之視為非流失玩家。

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所需資料，將觀察期視為資料特徵探勘期，並對其進行特徵工程，將原始特徵透過加總、平均等統計手法來建立新特徵，也額外以天為單位計算來獲得更多特徵，最後著手準備目標值，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 )，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 Gini Importance，並搭配決策樹作為代理人模型，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 Receiver Operating Characteristic ( ROC ) 曲線 [3] 與 Precision-Recall ( PR ) 曲線 [4] 來協助驗證學習模型之優劣，並同時利用 Weighted $F_\beta$ - Score [5] 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. 提出一新進玩家觀察期，排除舊有玩家，以利學習模型著重於新進玩家的資訊上。
2. 提出一新進玩家挽留期，預留時間以方便運營做挽留操作。
3. 提出一新進玩家表現期，配合新進玩家觀察期，以利學習模型預測玩家是否流失。
4. 透過資料特徵工程，以統計手法建立更多資料特徵。

# 參 考 文 獻


[1] P. Miller, “Gdc 2012: How valve made team fortress 2 free-to-play,” *Gamasutra. Haettu*, vol. 7, p. 2012, 2012.

[2] Swrve, “The april 2014 new players report.” `https://www.swrve.com/resources/weblog/the-april-2014-new-players-report`, 2014.

[3] T. Fawcett, “An introduction to roc analysis,” *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.

[4] D. Powers, “Evaluation: From precision, recall and f-factor to roc, informedness, markedness and correlation,” *Mach. Learn. Technol.*, vol. 2, 01 2008.

[5] C. Goutte and É. Gaussier, “A probabilistic interpretation of precision, recall and f-score, with implication for evaluation,” in *ECIR*, 2005.