Applied Soft Computing 118 (2022) 108491

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Churn prediction in digital game-based learning using data mining techniques: Logistic regression, decision tree, and random forest

Mai Kiguchi [a], Waddah Saeed [a,b,*], Imran Medi [a]

[a] School of Computing, Asia Pacific University of Technology & Innovation (APU), 57000 Bukit Jalil, Kuala Lumpur, Malaysia
[b] Center for Artificial Intelligence Research (CAIR), Department of ICT, University of Agder, Jon Lilletunsvei 9, 4879 Grimstad, Norway

ARTICLE INFO

Article history:
Received 4 March 2021
Received in revised form 11 January 2022
Accepted 16 January 2022
Available online 4 February 2022

Keywords:
Churn determination
Churn prediction
Machine learning
Digital game-based learning
Educational Technology

ABSTRACT

Educational Technology (EdTech) is an industry that integrates education and technology advances. Digital game-based learning (DGBL) is one of the narrowed-down categories of EdTech. One of the common issues in the EdTech market is the higher churn rate. However, because the DGBL market is still in the early stage, few studies related to marketing perspectives exist. Besides, the approach in education or online gaming industries can be only partially applicable to DGBL. A popular approach for addressing a higher churn rate is churn prediction. By using a dataset from a Japanese company providing DGBL services, this work proposes an approach for the combination of defining churn and churn prediction for DGBL. This work has three objectives. First, determining churn in DGBL by comparing the recency and the addition of average and two standard deviations of user inactive time. Second, clarifying the churn rate of the Japanese service, which became evident as 56.77% by using the newly created churn definition. Third, developing a churn prediction model by comparing logistic regression (LR), decision tree, and random forest models. Feature selection, dataset split ratio comparison, and hyperparameter tuning were conducted to achieve better predictions. Based on the results, LR scored the highest AUC of 0.9225 and an F1-score of 0.9194. These results are on the higher side comparing with the past churn prediction studies in online gaming and education industries. As a consequence, the results indicate the effectiveness of the proposed approach for churn determination and prediction in DGBL.

1. Introduction

Educational Technology (EdTech) is an industry that integrates education and technology advances. The education market is one of the vast markets where expenditure comes from governments, parents, corporates, and individuals [1]. The education sector is highly under digitized, with less than 3% of expenditure is allocated to technology [1]. Nevertheless, it is estimated that the amount spent on digital will grow to $342 billion by 2025 [1]. This is because of the growth of EdTech. Digital game-based learning (DGBL) is one of the narrowed-down categories of EdTech. The concept of the original non-tech game-based learning is having gameplay in a learning context, while DGBL is another specific terminology for game-based learning with the use of technology [2]. Because of the market expansion, the importance of the marketing approach in DGBL has been increasing.

Some of the important key indicators of the marketing approach are churn, retention, and churn prediction. Customer churn is the percentage of customers who stopped using a service or product in a certain period, while customer retention is the percentage of customer relationships, especially the ability to maintain customers using the service or product over time.

The churn rates in EdTech have been higher in common. The Washington Post reported about virtual education in 2019 [3], and it noted the difference in graduation rate between normal and virtual schools. The graduation rate of virtual schools was only 50.1%, whereas the overall graduation rate in the United States was 84%. In addition, the median churn rate in the education industry was 10.29%, which is the third-highest among nine categories as reported in [4].

To improve user retention, churn prediction is one of the important approaches in customer retention and management [5]. It allows companies to create better marketing strategies for churners. For example, sending push notifications or e-mails is one of the common ideas to prevent "will be churners" users based on the churn prediction. Therefore, successful churn prediction provides a better retention rate and reduces the cost, and it will benefit stakeholders such as game developers, advertisers, and platform operators.

* Corresponding author at: Center for Artificial Intelligence Research (CAIR), Department of ICT, University of Agder, Jon Lilletunsvei 9, 4879 Grimstad, Norway.
E-mail addresses: maiam24@gmail.com (M. Kiguchi), waddah.waheeb@uia.no, waddah.waheeb@gmail.com (W. Saeed), imran.medi@apu.edu.my (I. Medi).

https://doi.org/10.1016/j.asoc.2022.108491
1568-4946/

However, there is no research in DGBL about churn prediction and even churn determination. Research about online gaming and the education industry can be partly applicable to DGBL because DGBL is a combination of online gaming and education. The education industry has much research on prediction based on student demographics and academic information, but few with user behavior data. On the other hand, the online gaming industry has abundant research about churn prediction with user behavior, but there are significant differences with DGBL. First, the evaluation time frame is different. The time frame in the online gaming industry is often 7, 14, or 30 days [5], whereas education uses a longer time frame like the entire course period, which can be from a couple of weeks to one-year [6,7]. DGBL time frame is similar to education, and it requires waiting to define churn until the end, even if the courses are ongoing. This makes churn determination and prediction challenging to standardize. Social interaction is another difference in DGBL. Online gaming or application tends to have more social interaction between users, which affects user's churn [5]. DGBL products or services are designed for learning with gameplay, and there is less social interaction during the play than in the online gaming industry.

To summarize the points, there are no studies on churn determination and prediction in DGBL, and it is not easy to define and predict churners by simply applying education or online gaming industry knowledge due to the difference in learning periods, used data types, and social interaction. In this work, the combination of defining churn in DGBL and churn prediction was proposed. Accordingly, the key contributions of this work, which sets the foundation for further research, can be summarized as follows:

- Due to the lack of studies on churn determination and prediction in DGBL, this work presents an in-depth review of three main categories related to DGBL. The first category is the concepts and past studies of DGBL. Next is examining the definition of churn or retention, and the method and evaluation period of retention analysis from educational and online gaming perspectives. The last category is reviewing related research of churn prediction in education, online gaming, and DGBL. We believe this work will be a valuable resource for researchers working in churn determination and prediction in DGBL.
- Providing the definition of churns in DGBL and establishing the path for churn analysis and prediction for DGBL studies. Due to the lack of studies on churn determination and prediction in DGBL, this research contributes to churn determination and prediction in DGBL. This will help DGBL or gamification companies to conduct better marketing analyses on their services or products.
- To the best of our knowledge, there is no established model to address the higher churn by churn prediction in DGBL. Developing a churn prediction model in DGBL companies can help to provide an approach to reduce the higher churn rate in DGBL. Therefore, this work developed a churn prediction model by comparing three common models for churn prediction, namely logistic regression, decision tree, and random forest. A case study based on data from a Japanese company that provides online coding services was used to build and test the models.

The remainder of this paper is structured as follows: Section 2 is a review of the related works. The research method used with the implementation is explained in Section 3. Results and discussions are given in Section 4. There are four subsections in Section 4. The first one is about churn determination. The second subsection is about the descriptive and retention analysis based on the defined churn. The third one is hyperparameter optimization to obtain better prediction performance. The last subsection is about churn prediction and its performance results. Finally, the conclusion and future works are highlighted in Section 5.

2. Related works

The related works of three types of categories are discussed in this section. The first category is the concepts and past studies of DGBL. Next is the definition of churn or retention and the method and evaluation period of retention analysis. These were examined by industry which is education and online gaming. The last category is the churn prediction. DGBL, online gaming, and education researches related to churn prediction were reviewed in the last category.

2.1. Digital Game-Based Learning (DGBL)

The use of gameplay in education is not a new idea, but the use has been limited in nursery school for a long time [8]. With technological advancement, the use of DGBL was already employed in the 1970s [9]. However, the term "digital game-based learning (DGBL)" was made popular in the early 2000s by Gee [9] and Prensky [10]. DGBL is an approach to facilitate learning with digital games. Although DGBL is quite different from gamification, these two are sometimes being confused. The difference is that gamification uses game-like elements such as level, points, and rewards, whereas DGBL is learning through gameplay [11].

DGBL is gaining more attention from both educators and researchers [8]. The reason for widespread attention to DGBL is a combination of three aspects [12]. The three aspects are ongoing research by DGBL proponents, today's digital natives who are disengaged from traditional instruction, and increased popularity in the gaming industry.

In the previous research work, there are two main approaches to the DGBL category [8]. The first approach is based on the concept of Prensky, "Edutainment". Edutainment is entertainment that is designed for being educational [10]. With the idea of edutainment, DGBL can improve the motivation, or the confidence of students [13]. Another approach is based on the idea of constructivist and cognitivist [8]. With this, the educational value of DGBL is the acquired knowledge from decisions and behaviors made by the player to adjust to challenging situations on top of motivating.

In addition to the above two approaches, many researchers have evaluated the relationships between learning effectiveness, motivation, different courses, and different gaming design. These researches disclose that DGBL can enlarge the learning interests, enhance their motivation, and increase their performance [12].

There are various combinations of subjects, ages, and regions for the research focus; however, the current trend in DGBL can be provided by the three latest meta-analyses. The first meta-analysis focuses on serious game use in education [14]. The second meta-analysis scopes for kindergarten to 12th grade (K-12) mathematics education [15]. The third one focuses on the effect of DGBL on elementary science learning [16]. Based on these meta-analyses, the trend in DGBL is revealed. The publishing trend in DGBL has been increasing for more than a decade.

The work in [16] reported that all researches are fallen into two major categories: motivational and skills acquisitions, and knowledge construction and content understanding outcomes, which are the same approaches as mentioned by [8]. The work in [15] disclosed that the primary goal of all research for K-12 mathematics education is to understand how DGBL affects mathematics students' achievement. The work in [14] examined the meta-analysis in a wide range of research in DGBL. It reported

three findings which are influencing factors, positive effects, and negative effects. Influencing factors include gaming easiness and surprises, relationships between learning attributes and gaming mechanics, and types of games and age of learners. The positive effective findings include improving the learning outcomes and teaching, obtaining cognitive abilities, and facilitating a holistic understanding of concepts. On the other hand, negative effects are also reported from a couple of research as negative influences on the learning effectiveness from the aggravated mental workload.

As seen from the above results, both the past studies and the current trend of research in DGBL mainly focus on evaluating the relationship between learning effectiveness, motivation, different courses, and gaming design. The only research in DGBL for churn prediction was specialized for tutorial churners before registration. Even though the need for the marketing approach in DGBL will likely increase in the future, considering the current and expected future market expansion, there is no research on marketing-related topics in DGBL yet.

2.2. Definition of churn or retention, method and period of retention analysis

As explained earlier, the churn rate is higher in the EdTech industry in general. However, the definition of churn or retention can be varied and vague from two points of view. The first point is the period used for churn or retention calculation. According to [17], churn rate is measured by month, quarter, or year depending on the industry and products. The basic period is annual for most of the companies. Some services that are charged monthly use the churn rate by month or look at the churn rate monthly due to having a faster churn rate. Second, the meaning of churn. There is no doubt for the definition of churn rate, which is a percentage of customers who stopped using a service [18]. Nonetheless, what "churn" or "retention" means is different depending on the services and product.

Therefore, past studies should be reviewed for these perspectives. Due to the lack of research in DGBL, gaming and education research are chosen to be examined to see the definition of churn and how the churn period is chosen because DGBL is a combination of gaming and education.

2.2.1. Definition of churn, and period of retention analysis in the online gaming industry

There is no universally accepted definition of retention or churn because retention analysis is often used for internal use only [5,19]. Of course, game publishers use metrics to calculate but do not share sensitive detailed information with others.

Though, there is a rough definition in the gaming industry. A study claimed that online games should use a different definition of churn, not like other industries such as telecommunication or financial services [20]. Instead of using the withdrawal of membership, the inactivity period should be used. In the research, more than 13 weeks without any access was defined as churned. Using inactivity is one of the common approaches in the gaming industry, which is seen in other literature [21–23].

A work proposed a definition of churn in a different perspective for online casino game services [19]. Instead of using a specific time frame that is easy to understand and execute, using an arbitrary time frame for each user was suggested. Hence, it used three metrics to define churner: each player's last play, average days between play, and two standard deviations of the average days between play.

However, the common approach of retention analysis is a combination of the login frequency and evaluation time in the gaming industry [5]. The evaluation period often starts from the first day of the release date or the day a new user joined, and it is often 7, 14, or 30 days. There are four methods for retention calculation, and they are full retention, classic retention, accumulative retention, and return retention. The method of retention rate calculation is summarized in Table 1. In [5], full retention in 10 days was used to see the overall retention rate.

Classifying the method based on Table 1, some other research fit in the methods. For example, a study of prediction of churn of freemium mobile games used a return retention method and one month period for evaluation [24]. Another work also used return retention within 14 days after the last activity [25]. In general, the four retention methods with user login frequency or the inactive period are often used to define churn in the online gaming industry. Many of the approaches fall into one of the four methods, and the common churn evaluation period is 7, 14, or 30 days due to the higher churn in the first month. These methods are applied to both computer and mobile gaming.

2.2.2. Definition of churn, and period of retention analysis in the education industry

In the education field, a more simple retention analysis method is used compared to the gaming industry. For offline courses, it is clear that students stopped coming to the class means dropped out (churned). Nevertheless, same as the gaming industry, there was no clear definition of dropout for online courses. At an earlier stage of the dropout research, a study claimed that there is no clear definition of dropout, and it proposed the dropout as students that withdraw from the e-learning courses with financial penalties. This is because students can drop a course at the add/drop period with fully refunded [26]. Recent research on retention management for academic online courses seems to use the same definition of the student who stopped coming to the course and highly likely not including the add/drop period dropouts. However, the evaluation period is different depending on the research because the period is the same as the course duration. For example, one academic year was used for dropout analysis [27]. Another one used a fall term of three years at a university for dropout prediction [28].

This is the same with massive open online courses (MOOCs). MOOCs are a subset of educational technology, and it is a popular field of research these days. Retention analysis on MOOCs uses the same period, course duration, and the same definition of dropout, which means learners who are uncompleted the course. An analysis of dropout causes used one academic year of three semesters of MIT and Harvard courses, and the same definition of dropout [29]. Another research analyzed factors affecting retention, also used course duration, which is six weeks as a period and the same dropout [30]. Some courses with multiple weeks used multiple periods for evaluating dropouts. For instance, a study used eight weeks and checked every week's dropout rate [31]. Few studies defined and used a different definition of dropout. For example, a study of analysis on dropout reasons for MOOCs used the entire course period, and dropout was determined as "not logged in the course for more than 14 days" without finishing the course [32]. Another work tried a different approach to retention analysis of MOOCs [7]. It used one semester period data and compared the number of samples based on their defined engagement indicators between chapters. This is a similar concept as dropout or churn.

Overall, the popular period is course duration, but there are variations from a couple of weeks to a couple of years which makes it difficult to determine a specific cutoff like the gaming industry. On the other hand, the dropout definition is more straightforward because if the student did not complete the course, the student is defined as a churner in the educational industry.

Table 1
Retention rate calculation methods.
Source: Adopted from [5].

Method	Detail
Full retention	Count in if players return every single day during the evaluation period. It is extremely restrictive and not so widespread.
Classic retention	Count in if players return on the evaluation day. It is the easiest way to calculate, and it is the most widely used.
Accumulative retention	Count if players return days during the evaluation period and higher than the predefined threshold. This method is flexible but with a high computational cost.
Return retention	Count in if players return at least once during the evaluation period. It is the least restrictive and often produces relatively promising results.

Based on the above studies, the most common way to define churn is using a fixed evaluation period in online gaming industries, whereas course duration in education. This means the churn or not can be finally clear at the course end in education. The variety of course period makes this more challenging to find a standard approach to define churn. Because of the educational characteristic of course duration, it is also difficult to apply using a fixed evaluation period to define churn from a common online gaming approach. Thus, this work tried to define churn by examining inactive periods or arbitrary timeframes based on user behavior.

2.3. Churn prediction

Churn prediction has been studied because it is an important topic in various domains [5]. With the definition of churn, it becomes possible to predict customer churn. For churn prediction in general, common techniques used are machine learning with algorithms such as logistic regression, decision tree, and random forest [18].

The details of the studies are summarized in Tables 2 and 3. Table 2 contains DGBL and gaming literatures, while Table 3 contains literatures in education. Compared to Table 2, Table 3 has one extra column which is data source because literature in education often uses different data sources. Details about the studies are given in the following subsections.

2.3.1. Retention prediction in digital game-based learning

Although, to the best of our knowledge, there is no research on churn or retention prediction in DGBL, there is one research that predicts abandonment in online coding tutorials. The work is categorized in the gaming sector, but the chosen product for the research teaches programming concepts, which is considered in the DGBL category. The research focused on the prediction of learners who are likely to complete the next lesson of a tutorial instead of using the existing churn prediction approach [33]. This is because leaving open does not fit in the definition of the end of the course, and the tutorial is before registration of the membership, and it does not fit in the definition of withdrawal of membership. Cumulative features such as idle time, number of execution button clicks, and time to spend on reading, and learner features such as age, gender, registration status, and programming experience were used for the prediction. The used machine learning classifiers were logistic regression, random forest, and gradient boosting decision tree, and predicted 61% to 76% of learners who did not complete the next level with an average AUC of 0.68. However, this approach is too specialized for a tutorial before registration and is not generally applicable to DGBL products.

2.3.2. Churn and retention prediction in online and mobile gaming

The churn rate for online gaming industries is relatively high because players do not play the game until the content is exhausted [5]. There are many approaches or a combination of methods, selected data features, and time duration. However, the common approach is using machine learning with user behavior data. For example, logistic regression (LR), decision tree (DT), random forest (RF), and support vector machines (SVM) [25,34].

Earlier works often used the above common machine learning models. For example, LR, DT, Naive Bayes (NB), and artificial neural networks (ANN) were used to predict churn for freemium games [21]. SVM, DT, and LR were used to predict with general event frequency data for the versatility of any games [22]. Hidden Markov model was used to predict a major game, which is a shooter and massively multiplayer online game [23]. For a shorter period of prediction, a one-day churn prediction algorithm was introduced because more than 70% of new users would only play a game for one day and stop using it the next day [25]. The study used user activity features, monetization features, and gameplay style features (e.g., auction usage, spend on training), and five algorithms were used for prediction. The algorithms were LR, DT, RF, Gaussian NB, and gradient boosting. For a different approach, a work proposed using an arbitrary time frame for the definition of churn used the E-CHAID decision tree algorithm, and 60 features of seven categories were fed to the algorithm [19]. The data categories were recency, frequency, monetary value, length of relationship (e.g., number of days between the first and last playdates), inter-play (e.g., the average number of days between plays), bonuses/reward, and demographic. With technology advancement, another approach with a deep neural network (DNN) was proposed. A study proposed a DNN architecture of the inductive semi-supervised embedding model [34]. The semi-supervised model was used for the prediction and compared with state-of-the-art models such as LR, DT, RF, and SVM. Three categories of data features were used for this study; play history (e.g., game title, timestamp of play, wifi connection status, screen brightness), game profile (e.g., game genre, developer, number of downloads, rating), and user information (e.g., device model, region, OS version). Another study compared ANN and RF models for both classification and regression to predict churn by using in-game activities data [35]. The result showed that RF outperformed ANN model, and the RF classification and regression results were close. It also reported that time investments were the key indicators to predict churn. Nonetheless, the above studies focus on the improvement of prediction by changing models or the definition of calculation.

On the other hand, some recent studies proposed another different approach to customer segmentation. When mass-market strategy became more challenging to succeed due to not being

Table 2
Summary of churn prediction literature in DGBL and gaming industries.

Reference	Category	Purpose	Variables used for prediction	Model	AUC/F1 results
[33]	DGBL	Prediction of abandonment in tutorial	Cumulative features (e.g. idle time, execution button clicks), learner features (e.g. age, gender, registration status)	LR, RF, DGBT	AUC = 0 .68
[21]	Online /mobile game	Prediction of churn	Number of sessions, number of days, average playtime per session, average playtime between sessions, etc.	LR, DT, NB, NN	F1 = 0 .916
[22]	Online /mobile game	Prediction of disengagement	Event frequency	SVM, DT, LR	AUC = 0 .7
[23]	Online game	Prediction of churn	Activities, performance, achievements	Hidden Markov model	AUC = 0 .77
[19]	Online game	Prediction of churn	Recency, frequency, monetary value, length of relationships, inter-play, bonuses/rewards, demographic	E-CHAID DT	AUC = 0 .88
[5]	Online game	Prediction of player	Lifetime engagement features (e.g. login frequency, average playtime), performance features (e.g. level, coins), social interaction features (e.g. number of in-game friends)	Stickiness based FCM for clustering NB, RBF	–
[20]	Online game	Prediction of churn	In-game activities (e.g. number of days user play, playtime)	RF, XG boost, generalized boosting regression	AUC = 0 .9358
[25]	Mobile games	Prediction of early churn	Activity features (e.g. playtime, session count), monetization features (e.g. in-game money spend), gameplay style features (e.g. auction usage)	LR, DT, RF, NB, gradient boosting	AUC = 0 .83
[34]	Mobile game	Prediction of churn	Play history (e.g. timestamp of play), game profiles (e.g. game genre, developer, rating), user information (e.g. device model, OS version)	semi-supervised deep NN, LR, RS, DT, RF, SVM	AUC = 0 .82
[24]	Mobile game	Prediction of churn	Activity indicators, Activity time, Engagement indicators	Proposed joint model	–
[35]	Gamification/Online game	Prediction of churn	In-game activities (e.g. playtime, frequency of usage, game actions and participation)	ANN, RF (Classification and Regression)	AUC = 0 .77

Notes: DGBT, Gradient Boosting Decision Tree; DT, Decision Tree; LR, Logistic Regression; NB, Naive Bayes; NN, Neural Networks; RBF, Radial Basis Function; RF, Random Forest; SVM, Support vector machines.

able to satisfy all customers in the market, customer segmentation was proposed to address the issue [5]. It divided a market into distinct customer groups by similarities to create marketing strategies more efficiently [40]. This is considered one of the most effective tools for marketing to uncover market opportunities, to determine customer needs and wants, and to disclose potential markets [41]. Therefore, retention analysis with customer segments can illuminate risks or opportunities more clearly. For example, research proposed a churn prediction method by selecting the prediction target and by setting a threshold of the target [20]. The proposed reason is that the high churn rate and customer lifetime value are skewed, so focusing on churn prediction for loyal customers is more cost-effective. In addition, setting a threshold to maximize the expected profit rather than maximizing the accuracy. Hence, the prediction performance results include an expected profit for each model. This study used data of in-game activities (e.g., the number of days user play, playtime, achievement points, amount of money user gain). Similar to the above study, some research suggested calculating the churn rate for user clusters. Another study proposed a framework with the original joint model, and one part of the framework is predicting dropout probabilities. With the model, it also used k-mean clustering to segment users and then predicted dropout probabilities [24]. Another segmentation with a churn prediction model was proposed by [5]. It first created user segments by stickiness-based fuzzy C-means, then evaluated the retention trend for each cluster. Engagement, performance, and social features were used.

Thus, there are many perspectives and approaches to churn prediction in the gaming industry. The common approach is that all studies use user behavior data for prediction and the different game category (online or mobile) does not affect the approach that much.

2.3.3. Churn and retention prediction in education

In the education industry, the prediction of student failure or dropout has been a common topic. Several studies of prediction with machine learning have been reported. Despite having the same aim, the data used for prediction is different from the gaming industry. The common data collection is time-consuming questionnaires [36].

Earlier research proposed a prediction of academic failure of on-campus with educational data mining, and it used DT and induction rules [6]. It used two survey results and a dataset from the school. Therefore, there are many variables such as demographic data, family information, and academic grades and scores. This study concluded that prediction models performed with relevant accuracy. Another study to predict student failure

Table 3
Summary of churn prediction literatures in education.

Reference	Category	Purpose	Data source	Variables used for prediction	Model	AUC/F1 results
[36]	Online/On-campus	Early prediction of failure	Online course and school database	On-campus data (e.g. age, gender, civil status, exam performance), distance learning data (e.g. access frequency, participation in the forum)	SVM, DT via J48, NN, NB	F1 = 0 .82
[37]	Online	Prediction of dropout	Survey	Demographic variables, self-efficacy, readiness, prior knowledge, and locus of control	KNN, DT, NB, and NN	AUC = 0 .866
[28]	On-campus	Prediction of retention rate	School database	Gender, residency status, ACT composite score, high school class rank, race/ethnicity, student, etc.	LR	–
[6]	On-campus	Prediction of dropout and failure	Survey, School database	Scores of each subject, level of motivation, GPA, smoking habits, physical disability, etc.	JRip, NNge, OneR, Prism, Ridor, ADTree, J48, RandomTree, REPTree, SimpleCart	–
[38]	On-campus	Prediction of failure	Survey	Personal and family-related (e.g. age, parents occupation), previous education (e.g. scores of multiple subjects of previous education), academic results (e.g. scores)	NNge, OneR, SimpleCart, Random Tree, NB	–
[31]	MOOCs	Temporal prediction of dropout	MOOCs database	Clickstream (which pages students visited and when or how many times students clicked on certain sources (e.g., syllabus, modules, quizzes, etc.)), quiz scores and discussion forum data	general Bayesian network, decision tree (C4.5), Stacking	AUC = 90.7
[7]	MOOCs	Prediction of decrease of engagement	MOOCs database	Video engagement, exercise engagement, assignment engagement	LR, stochastic gradient descent, RF and SVM	AUC = 0 .914
[39]	MOOCs	Prediction of dropout	MOOCs database	Clickstream data divided into 7 categories (e.g. video, access, wiki, problem, navigate, discussion and page close)	Proposed deep learning model named CLSA, LR, SVM, CNN, LSTM, CNN-LSTM, and DP-CNN	F1 = 0 .869

also used three survey data [38]. One is for personal and family-related information (e.g., parents' occupation, income, number of family members), the second is for previous education (scores of multiple subjects from past education), and the third is for academic factors. The highest accuracy of 87.12% was achieved with NB.

As for the non-survey data use, a work used non-survey data to predict retention rate from student enrollment data of a university [28]. It focused on using demographic data such as gender, race and ethnicity, high school rank instead of academic and social engagement, and LR was chosen as an algorithm. The prediction model performance was proven with an accuracy of 83.2%.

However, these predictions are for on-campus students and not for online courses. A work proposed prediction of online program dropout [37]. It conducted five surveys for data collection and collected demographic variables such as age, gender, previous online experience, self-efficacy, readiness for online learning, prior knowledge about online program courses, and locus of control. It then used DT, NB, ANN, K-nearest neighbor (KNN) algorithms for prediction. Both AUC and accuracy were the highest with the KNN algorithm, 0.866 AUC and 87% accuracy. A study approached an evaluation of the effectiveness of student failure prediction [36]. It used non-survey data from two datasets, which were from distance learning and on-campus. On-campus data contained age, gender, civil status, exam performance, number of correct exercises, and amount of exercise performed. Distance learning data contained age, gender, civil status, the performance of exams, and assignments. This is different from past research because the data contained user behaviors such as access frequency, participation in the forum, and blog use.

Nevertheless, this is difficult to apply in DGBL services and products because it uses on-campus data, which is unavailable for commercial services and products.

The arrival of MOOCs gave a different approach to predicting churn or dropout by using more user behavior data like the gaming industry. The work in [31] focused on the early prediction of dropout of MOOCs and used clickstream data, quiz scores, and discussion forum data. C4.5 decision tree, general Bayesian network (GBN), and the ensemble learning method called stacking generalization were compared. The stacking of C4.5 and GBN outperformed the base algorithm alone in average precision and

AUC, which was 91.7% and 90.7%. Other research tried using clickstream data with a deep learning model called CLSA [39]. It reported that CLSA model gave the best results with an F1-score of 0.869 compared to the models reported in the past studies (LR, SVM, CNN, long short-term memory network (LSTM), CNN-LSTM, and time-series CNN-based).

Another study proposed a prediction model on the decrease of engagement with three different engagement features for MOOCs: video, exercise, and assignment engagements [7]. Video engagement is the average percentage of videos watched, including partially. The averaging percentage of exercises calculates exercise engagement. Assignment engagement is computed by the averaging percentage of the assignments submitted. LR, RF, SVM, and stochastic gradient descent (SGD) were used as prediction algorithms, and for all engagement, SGD outperformed other algorithms.

Even though there are many approaches in education, the different data sources (survey, school database, and MOOCs database) have different purposes and churn definitions. Survey data or on-campus studies are not suitable for DGBL services or products.

By studies in both in education and gaming industries, the combination of MOOCs and the gaming industry seems the best approach for DGBL because of the data perspectives. Using user activity data for machine learning models is the popular approach, and the common machine learning models in MOOCs and the gaming industry are LR and tree-based models. Therefore, this work applied the three common machine learning models (i.e., LR, RF, and DT) in MOOCs and the gaming industry.

3. Research method and implementation

In this section, the research method used with the implementation is explained in detail. It is good to note that all the programming code was written in Python, and some graphical plots were generated by Tableau.

3.1. Data collection

User data was provided by a Japanese company which is primary data. The company provides an online coding service which is like a role-playing game (RPG). A player moves through the story by talking to characters, moving areas, solving puzzles, and learning to code to cast magic. Two types of data were available from the service: structured data from the customer relationship management (CRM) database and semi-structured user log data. The number of participants in the data was 3,557. The gender distribution is about 56% Female and 44% Male. About 50% of the players fit in between age 17 and 33. The 25% of players are under 17, and another 25% are above 33.

The content of the data is shown in Table 4. The learning contents of the Japanese service are split into seven chapters, and the next chapter is accessible after about seven days to a couple of weeks after finishing the previous chapter. Therefore, each player has a different chapter release, start and finish dates. The chapter-related dates for each user are stored in two files (ID 1 and 2). ID 1 holds the timestamp of release, start, and finish dates. ID 2 contains chapter 7 finish dates which are stored in a different table from ID 1. In addition, each chapter contains many small lessons, and players acquire coins and experience points by finishing lessons. This obtaining history of experience points and coins are stored in ID 3. If a player repeats the same chapter, the history of repetition data, such as repeating chapter number and timestamp, are stored in ID 4. ID 5 file contains user demographic information from CRM. All user activity logs except for learning lessons are found in ID 6. For instance, logs are recorded when the user move to a different area, gains a new item or gets access to a new area. The lesson log contains the user activity log related to learning, such as the started and finished time of a lesson.

Table 4
The content of the data.

ID	Number of observations	Description
1	6,973	Chapter related dates such as release, start and finish dates.
2	514	Chapter 7 finish date.
3	478,700	Historical experience points and coins acquirement data.
4	1,496	Historical user replay data.
5	3,982	CRM data. Contains demographic information such as user name, email, address, birth date, and gender.
6	8,587,940	User log of all activities except for learning contents related.
7	562,581	User log in the learning contents such as start and finish date-time of a lesson.

3.2. Data preparation

In this section, the steps used for data preparation before modeling are explained.

3.2.1. Data selection

Based on past churn prediction studies that used user behavior data, there are some common categories of variables. Because DGBL has user behavior data similar to the gaming industry, most of the data selection is influenced by the gaming industry, and some are referred from MOOCs in education.

The majority of works in the gaming industry use the cumulative number of users such as the number of logins, plays, or clicks for engagement [5,19–21,24,25,34]. The average values or duration are common to be used as well [19,21]. In this work, to calculate these values, user activity log variables such as user id, timestamp, and contents were selected from the user activity log. In addition, performance features were used in the literature [5,7,19,31]. Levels and the number of coins were selected as performance features in this work. Demographic information is common to be used in education and few in the gaming industry [19,33,36,37]. So, gender, date of birth, resident area were selected from CRM data.

3.2.2. Data aggregation, transformation and integration

Several tasks were done at this step. Below are the details of these tasks.

User activity and lesson log integration. User logs contain user behavior with the action and timestamp. To calculate more aggregated values such as playtime, inactive time, and the number of logins, these two files were merged by account ID.

Experience points, coins, and replay data. The total value for each user for experience points, coins, and the number of replays was calculated.

CRM data transformation. The transformation process at this stage was done by altering gender from Japanese to English (i.e., 男性 to Male), calculating age from the birth date, and retrieving English prefecture from postal code.

Chapter progress data. The chapter status in the two collected files was combined to grasp how many users have finished playing the entire content. The data obtained after the combination process were timestamp when chapters were available to the user, the timestamp when the user started and finished the chapters, the period between open and start the chapters (wait), and the average period between open and start, which was calculated by dividing the total wait by the number of chapters played.

Table 5
The descriptive statistics of all variables.

Variable	Count	Mean	Std	Min	25%	50%	75%	Max
total_playtime (min)	3701	1640.15	2207.54	0	339.55	885.82	2213.1	38390.98
total_login	3701	36.01	50.27	1	700	2000	4700	742
total_inactive (min)	3701	307397.91	255321.92	0	66484.59	251926.17	496751.22	928384.54
playtime_average (min)	3701	52.09	28.87	0	33.93	46.74	63.15	298
inactive_average (min)	3701	16507.71	25549.83	0	4375.9	9074.17	18022.08	379752.63
first_login	3701	NaN	NaN	NaN	NaN	NaN	NaN	NaN
last_login	3701	NaN	NaN	NaN	NaN	NaN	NaN	NaN
entire_period (days)	3701	211.51	178.95	0	42	173	344	645
churn_status	3701	NaN	NaN	NaN	NaN	NaN	NaN	NaN
*chapter*1*_playtime* (min)	3701	404.72	373.23	0	243.65	344.1	479.33	7442.73
*chapter*2*_playtime* (min)	3701	479.59	535.10	0	0	386.45	804.52	5842.65
*chapter*3*_playtime* (min)	3701	210.84	399.96	0	0	0	383.18	8615.63
*chapter*4*_playtime* (min)	3701	234.51	642.30	0	0	0	274.37	22511.83
*chapter*5*_playtime* (min)	3701	117.44	360.91	0	0	0	0	7105.2
*chapter*6*_playtime* (min)	3701	99.97	331.14	0	0	0	0	4787.72
*chapter*7*_playtime* (min)	3701	93.08	492.92	0	0	0	0	14452.2
exp	3521	13753.37	14314.31	150	4570	8670	17700	135010
coins	3521	15485.90	15841.23	200	4700	9920	21760	155850
replay	3521	0.37	4.1	0	0	0	0	177
wait_average (days)	3701	11.43	37.92	0	0	1.48	6.46	644.54
gender	3557	NaN	NaN	NaN	NaN	NaN	NaN	NaN
age	3557	29.63	70.44	0	17	26	33	2010
prefecture	3533	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Player data. The last aggregation was conducted on the player log data created using user activity and lesson log integration steps. To calculate playtime and number of logins, two timestamps were compared, and the difference was aggregated as playtime or inactive time. The following rules were applied to determine that the user is still playing or has left:

- If the difference between two timestamps is equal to or less than 60 min, consider the user is still playing. The difference is treated as playtime.
- If the difference between the two timestamps is greater than 60 min, consider the player has left. The difference is treated as inactive time, and increment the number of the login by one.

At the end of this step, these are the created data for each user: total playtime in minutes, the total number of logins, the total inactive time between logins, average playtime per login, the average inactive time between logins, the first login timestamp, the last login timestamp, the engagement period, the subtraction of the last login from the first login, playtime of each chapter, and churn status either True or False.

Data integration. After the player data aggregation, all aggregated data and CRM data were merged. Consequently, this integrated data contains 3,701 players, all types of data including playtime, obtained score, chapter progress, and demographic-related information.

3.2.3. Exploratory data analysis (EDA) and data cleaning

The dataset is ready for EDA and data cleaning. First, the descriptive statistics of all variables were calculated as shown in Table 5.

3.2.4. Data cleaning

Since a company provided the data, the selected data was expected to be dirty. Data cleaning is necessary to enhance data quality [42].

Imputation of missing values was the first step. There are six variables with missing values; *exp*, *coins*, *replay*, *gender*, *age*, and *prefecture*. The observations with missing values were checked manually because there should not be data with missing values, according to the Japanese company. The observations with missing values have the same patterns with no demographic information such as *age*, *gender*, *prefecture*, and no play log after chapter 1. However, some accounts seemed to play *replay*, which should be available after chapter 3. Having a replay log but no log after chapter 1 should not exist. It turned out that there were 144 trial accounts created in terms of marketing for schools and campaigns. These trial accounts lacked more than 70% of data and did not have reliable values because they used debug features for test accounts so the player could skip chapters or lessons to see other components. Therefore, these 144 accounts that did not have a *gender* variable were removed from the dataset. *exp*, *coins*, and *replay* values exist only for players who played until a certain area in chapter 1, so newer players do not have this data. Hence, these missing values were filled as 0. For the prefecture, "Tokyo" is the most frequent prefecture based on the statistical results. NaN was replaced with "Tokyo". The *prefecture* is categorical values, so it was then label-encoded. The value was set in alphabetical order. After imputation of missing values, *gender* and *churn_status* were also label-encoded to the binary values. Female was set to 0 and Male to 1 for *gender* and for *churn_status* False was set to 0 and True to 1.

Next, highly impossible values in *age* were handled. A highly impossible age was found based on the describe output and box plot. The minimum age was 0, and the maximum was 2010. For the younger age, research about reading ability age was reviewed to determine the lowest possible age. According to Japanese research, 60% of four years old children can read 80% of Hiragana even though the learning level varies depending on the family or day-care environment [43]. 80% of the reading ability of Hiragana is not perfect for playing the service because there are two other more types of characters in Japanese (Katakana and Kanji). Nevertheless, it is possible to start playing with their parents' help. This implies that the ages below 4 should have less reading ability, making it more difficult to play the service. Thus, the ages below four were replaced by the median of 26. On the other hand, the ages above 100 are 119, 825, 972, 2002, 2004, 2009, 2010. The oldest age in January 2020 in Japan was 117 years old [44], so the ages above 100 are also replaced with the median 26.

Lastly, outliers need to be handled. The found outliers should not be deleted or replaced because these are real values. Nonetheless, outliers have a strong impact on the modeling, so standardization should reduce the impact. The standardization of the variables with outliers was conducted for these variables.

Table 6
The variables in *features* dataset.

Variable name	Category	Details
total_login	Engagement	Total number of logins of the user.
entire_period (days)		The engagement period. Subtract *first_login* from *last_login*.
avr_ch_wait (days)		The average period between open and start. Total wait divide by the number of chapters played.
replay		Total number of replay per user.
total_playtime (min)		Total playtime of the user in minutes.
total_inactive (min)		Total inactive time between logins.
average_playtime (min)		Average playtime per login. Calculated by *total_playtime* divided by *total_login*.
average_inactive (min)		The average inactive time between logins. Calculated by *total_inactive* divided by *total_login*.
*ch*1*_playtime* (min)	Performance	Playtime of chapter 1
*ch*2*_playtime* (min)		Playtime of chapter 2
*ch*3*_playtime* (min)		Playtime of chapter 3
*ch*4*_playtime* (min)		Playtime of chapter 4
*ch*5*_playtime* (min)		Playtime of chapter 5
*ch*6*_playtime* (min)		Playtime of chapter 6
*ch*7*_playtime* (min)		Playtime of chapter 7
exp		Total exp points per user
coins		Total coins per user
gender	Demographic	Gender in binary 0 or 1 (0 = Female and 1 = Male)
age		Age of the player
prefecture		Prefecture from 0 to 47
churn_status	Target	Churn status in binary. 0 or 1 (0 = False and 1 = True)

3.2.5. Feature selection

The variables in the dataset were classified into three categories of demographic, engagement, and performance features as shown in Table 6. We called this dataset as *features* dataset. The target variable is *churn_status*.

Feature selection may improve model performance. Therefore, except for the *churn_status*, 20 variables were processed for feature selection. Many past studies related to churn prediction in gaming and education industries used manual feature selection, such as using their experience and original calculations depending on the variables. This kind of approach is not widely applicable. Nevertheless, two past studies, which are the prediction of dropout of MOOCs [31] and prediction of the player lifetime [5] used Principal Component Analysis (PCA). Thus, in this work, the PCA with automatic selection for the number of components was used [45]. The generated principal components were 19 named from "PC-1" to "PC-19". The explained variance of these principal components are shown in Table A.1. These generated principal components were selected and stored in a new dataset, called *features_pc*. Both datasets, *features* and *features_pc*, were used for modeling to compare the influence in the performance.

3.3. Definition of churn

The majority of the definition of churn is course in-completion in education. Many academic online course durations are fixed weeks and have assignment submission or exams at intervals [31, 36]. In this case, churners are users who did not complete the work in the week or by the end of the course. However, commercial educational content is often composed of a sequence of chapters or levels, and it is common in MOOCs [7]. A study in DGBL conducted an engagement analysis for each level [33]. This also implies the generality of chapters (levels) in DGBL. The difficulty of using chapters is that there is no time limitation, and this means it is difficult to define if the player really dropped out or not.

To address this issue, the gaming industry churn definition can be more applicable, which uses user inactivity duration after the last login. By using inactivity duration as a cutoff, churn can be defined. A "churn window (C)" approach proposed by [23] can be used to define the cutoff. If there is no active data for a player during at least C weeks, the player is considered churned. The churn window is set based on the scattered plot of total playtime (sec) vs. average absence from the game (in days). In this work, C was set to 4 because the scattered plot density is much higher when the average absence is less or equal to 28 days (which means four weeks). Note that with this calculation, the defined churned users can return to the game.

If the above approach does not work well with the provided dataset, such as the evenly scattered data or no distinctive cutoff, an arbitrary time frame for each user is calculated to define churners. Research showed that recency-related variables are the most important variables to predict churn for online gambling over the other variables such as frequency, monetary, or length of relationship-related variables [46]. In addition, the approach of calculating arbitrary time frames by using the recency and standard deviation has been used to define churn for the online gaming industry [19]. The defined churners were used to predict future churners by E-CHAID decision tree, and the results showed the valuableness of the approach. Therefore, the following equation from the research was used to determine churners:

$$recency > avg\_between + 2 * std\_between \quad (1)$$

where *recency* is the number of inactive days from the last play, $avg\_between$ is average days between play, and $std\_between$ is the standard deviation of days between play.

3.4. Modeling and evaluation

After the feature selection step, the data for both datasets were split into training and test sets in three patterns of data splitting. The first pattern is 90% for training and 10% for test sets. The second one is 85% for training and 15% for test sets.

Table 7
AUC results to find out the best dataset and splitting percentage for modeling.

Model	Dataset	AUC		
		80%–20%	85%–15%	90-10%
Decision tree	*features*	0.7106	0.7213	0.7081
	features_pc	0.8345	0.8410	0.8492
Logistic regression	*features*	0.7150	0.7207	0.7222
	features_pc	0.7832	0.7708	0.7808
Random forest	*features*	0.8617	0.8571	**0.9605**
	features_pc	0.9582	0.9584	**0.9605**

Notes: The best result for each model is underlined.
The best result in the table is in boldface.

The last one is 80% for training and 20% for test sets. Stratified 10-fold cross-validation for performance comparison of datasets and hyperparameter tuning was used with the training set.

As for algorithms, the three common algorithms for churn prediction were selected based on education and gaming industries literature. Because DGBL is a combination of these industries and there is no research in DGBL churn prediction, the common algorithms from these two industries were chosen to apply for the DGBL dataset. These algorithms are logistic regression [7,21,22,25,33,34], decision tree [21,22,25,31,34], and random forest [7,25,33,34].

Logistic regression (LR) is a probabilistic algorithm for binary classification [25]. The mathematically clear output is one of the reasons for the popularity. The problem of LR is the difficulty with interpretation [22]. On the other hand, decision tree (DT) is one of the most interpretable models by a human. DT is a classification algorithm that is used for supervised learning problems and generates a rule-based hierarchical tree [25]. The tree displays what features are connected to the terminal nodes, so the interpretation is simpler. Random forest (RF) is another tree-based classification algorithm that consists of many decision trees. RF constructs multiple decision trees by randomly sampling, and each tree conduct classification and the result. RF then collects the outcomes from the trees and selects the best result as a final result by voting [47].

The area under the receiver operating characteristics (ROC) curve (AUC) was used as a performance metric. ROC curve is a probability curve of sensitivity and specificity, and performance measurement for classification [48]. AUC is the area under the curve, and a higher AUC close to 1 means a larger area, which represents a better classifier. Many studies used AUC as a metrics to compare the prediction models in both gaming and education industries [7,19,22,23,25,31,33,34]. Because the proportion of the target value *churn_status* (0 and 1) are expected to be imbalanced due to having more churners in general, AUC was used because the imbalanced distribution does not influence it [48].

AUC results to find out the best dataset and splitting percentage for modeling is summarized in Table 7. The results are using the stratified 10-fold cross-validation using the training set only. It can be seen from this table that using *features_pc* produced better performance with all models compared to *features*. However, both produced the same best performance with random forest model with a data split ratio of 90% training and 10% test. Thus, this data pattern was used for the modeling and hyperparameter optimization for all three machine learning algorithms, and the 10% test dataset was put aside until the last modeling evaluation.

4. Results and discussions

There are four subsections in this section. The first subsection is about churn determination. This is based on the case study product. The second one is the descriptive and retention analysis based on the defined churn. These are specific for the case study services, and it helps the company creates new marketing strategies and improves curriculum to increase the retention rate. The third subsection is hyperparameter optimization. DT, RF, and LR models were created using the default setting in Section 3.4. The goal of the hyperparameters optimization step is to find better performance for the models. The last subsection is about churn prediction and its performance results. There are three machine learning models which are DT, RF, and LR. The results of these models with the best hyperparameters are compared.

4.1. Churn determination

In Section 3.3, we explained how to determine churn. The first plan is checking a scatter plot of total playtime and average inactive time. If the first plan does not indicate the distinctive churn spot, then the second plan calculates an arbitrary time frame for each user to define churn.

Two variables *total_playtime* in minutes and *average_inactive* in minutes were used. *average_inactive* was converted from minutes to days. The generated scatter plot is illustrated in Fig. 1. The plot is less dense around 100 days, or a more obvious spot is between 150 and 180. However, this is not a convincing cutoff value due to the volume of churners. Even if the cutoff is set at 100 days, the number of churned players is 28, which only covers 0.75% of the population. Furthermore, there is no strong reason to set a cutoff on other days.

Next, another calculation of the average inactive was applied to see the difference. The following equation calculates the average inactive in a day:

$$average\_inactive(Hours) = (60 * 24 - total\_playtime/total\_login)/60 \quad (2)$$

Fig. 2 displays the scatter plot based on the calculated average. The second plot is harder to find less dense areas after the higher dense hours. The potential reason for the issue is that DGBL has a long inactive period, and the spread-out plot makes it less recognizable than the online gaming industry data. These scatter plot approaches can be concluded that they may not be suitable for DGBL.

Therefore, the second approach of calculation of arbitrary time frame was conducted. To compute the time frame, three values are required. Recency (*recency*), average inactive time between logins (*avr_inactive*), and standard deviation of inactive (*std_inactive*). Because the average inactive had already been calculated, computation of *recency*, and standard deviation should be added. The *recency* is the time between the last login and the current day. Because the date of the dataset given is January 27th, 2020, the current day is set as January 28th, 2020. With the current date, the *recency* is calculated. For the standard deviation for each user, a list named *inactive_list* was used. The *inactive_list* contains inactive time in seconds between logins for the user for the calculation. The *personal_cutoff* has been determined from the addition of *avr_inactive* and double of *std_inactive*. If the *recency* is bigger than the *personal_cutoff*, the user is defined as churned and *churn_status* is set to True.

The defined churners are plotted to see the percentage of churners as shown in Fig. 3. Out of 3,557 observations, 56.77% of the population are churners and 43.23% are non-churners based on the value in *churn_status*. The class imbalance here is likely considered as a slight imbalance (4 churners: 3 non-churners). The calculation and the proportion are acceptable compared to the first approach and by considering the higher churn rate in EdTech as explained earlier. Hence, it is concluded that the retention rate is 43.23% for the product.

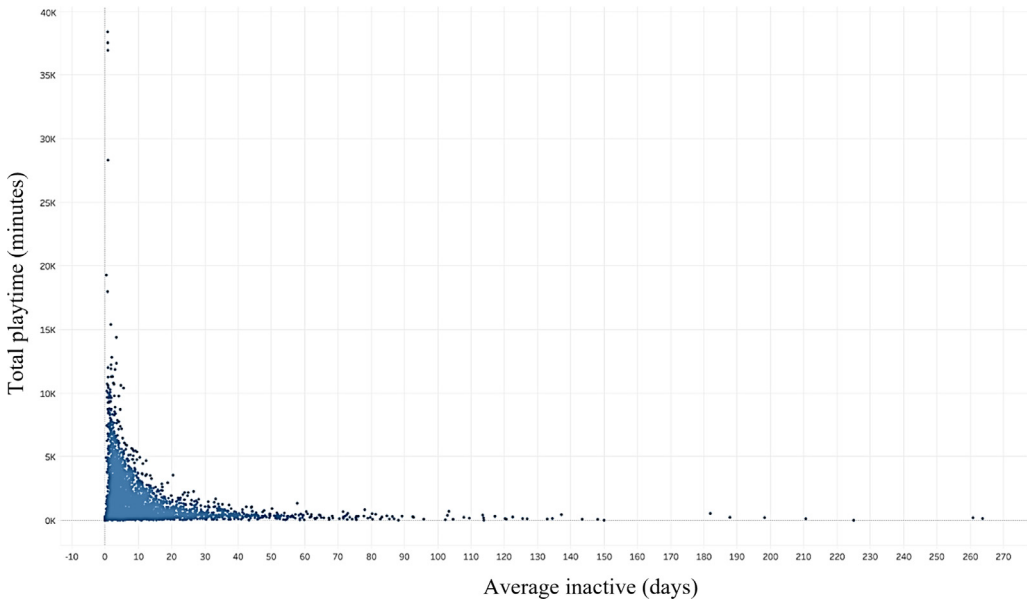


Fig. 1. Scatter plot 1 with average inactive (days) and total playtime.

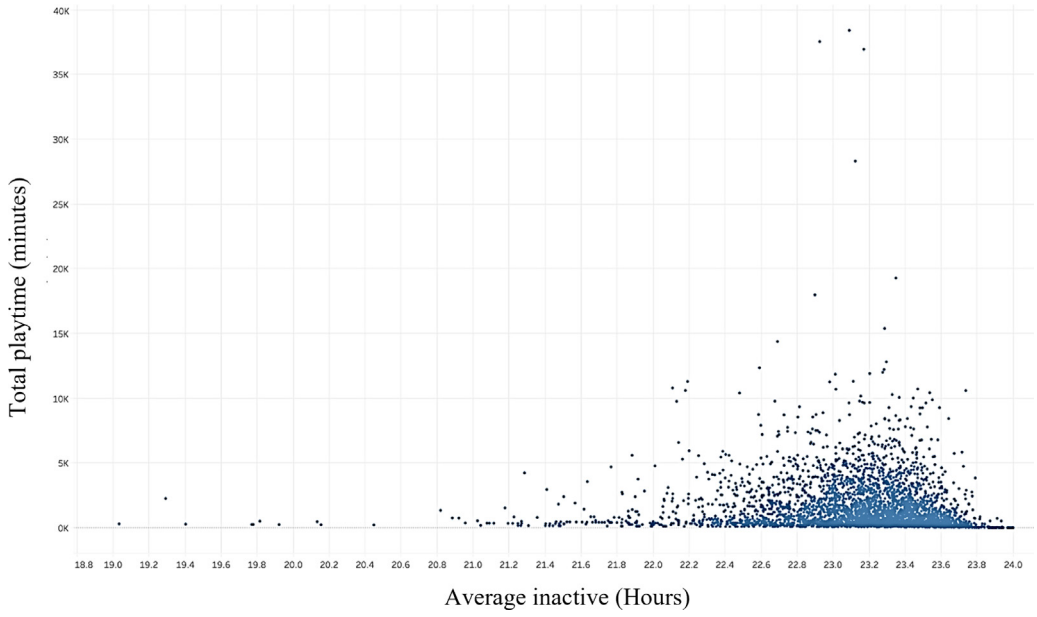


Fig. 2. Scatter plot 2 with average inactive in a day (hours) and total playtime.

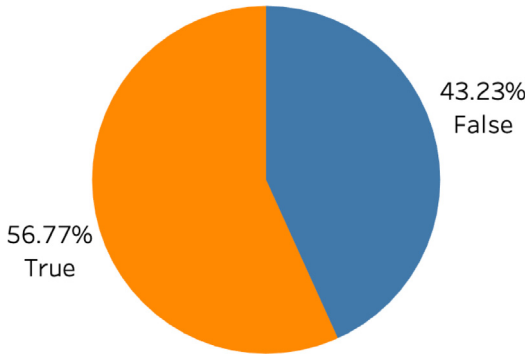


Fig. 3. The proportion of churners (True) and non-churners (False).

4.2. Retention analysis

In this subsection, descriptive analysis for overview and retention analysis by chapter was conducted. The descriptive analysis is to grasp the overview of user activities by using the company's data.

4.2.1. Descriptive analysis of the user logs

To comprehend the current situation of the case study service, monthly active users (MAU) and the popular hour and day of the week are plotted as shown in Fig. 4 and Fig. 5, respectively. The MAU had generally been increasing during the first year, and then it settled around 1,000. Since the number of users has increased since the service launch, the MAU settlement should be concerned.

Regarding the popular hour and day of the week, the expected popular hour was after office hours until midnight as many web

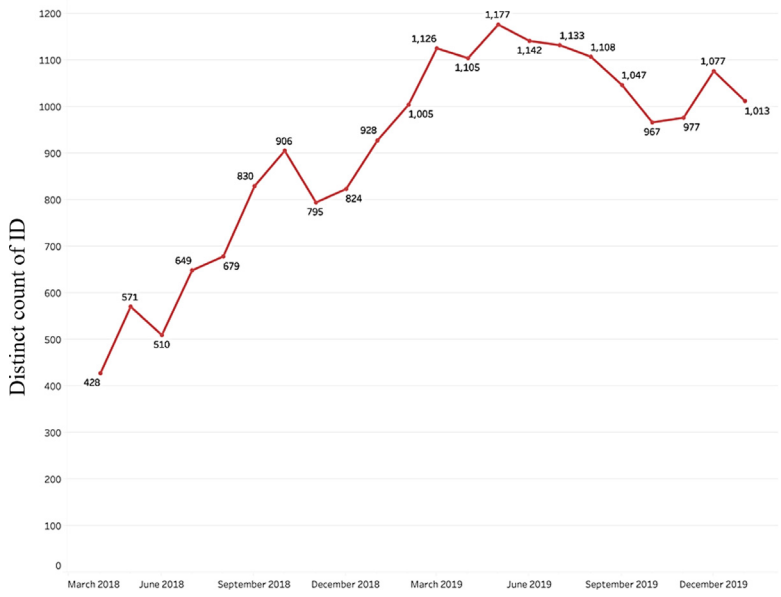


Fig. 4. Monthly active users (MAU).

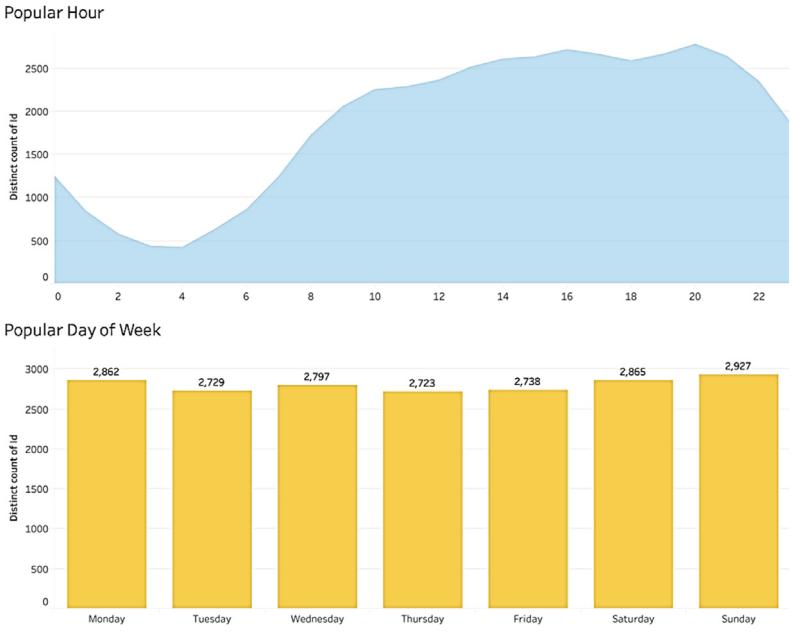


Fig. 5. Popular hour and day of the week.

services, and the day of the week is on the weekend. Nevertheless, as shown in Fig. 5, between 1 PM to 9 PM has bigger access than other hours, and the number of access drops after 8 PM. In addition, there is no difference between the weekdays and weekends unexpectedly. The assumption for the reason for the popular hour and day of the week is that there are more students than adults' workers. They can access the service earlier after school, and bedtime is also earlier, so the access starts decreasing after 8 PM. By considering the mean age is 29.6 and the median is 26 as shown in Table 5, a certain part of the registered age could be the player's parent's age.

4.2.2. Retention analysis by chapter using the aggregated data

After understanding the overview of user behavior, the chapter's retention analysis was conducted using the newly defined churn status and aggregated data. First, the determination of how long the users spend each chapter on average was checked. The median of *playtime* (hour) per chapter is plotted to grasp the general playtime in Fig. 6. The reason for using the median is to reduce the impact of the outliers, which have an extremely long playtime. The shortest playtime is in chapter 1, and the longest is in chapter 2 in general. Considering the 56.77% are churners and 43.23% are non-churners based on the churn determination, there is a recognizable difference between churners and non-churners. Overall the churners play less than non-churners, but the playtime gap depends on the chapters. Table 8 shows the difference of median playtime (hour) for each chapter.

Next, to see the completion rate of each chapter, the number of finished users for each chapter was calculated. Then, the completion percentage of the current chapter from the previous chapter completion was calculated. The completion percentage for non-churners is generally high, but not 100%. This means that

Table 8
Median playtime (hour) difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-churners	5.96	12.46	8.15	11.79	8.17	10.20	11.27
Churners	5.70	10.52	7.13	9.94	6.86	7.00	6.62
Difference	0.26	1.94	1.02	1.85	1.31	3.20	4.65

Table 9
Chapter completion rate difference by churn status.

	Chapter 1	Chapter 2	Chapter 3	Chapter 4	Chapter 5	Chapter 6	Chapter 7
Non-Churners	71.92	67.01	82.93	76.53	81.11	80.82	89.49
Churners	75.22	50.35	65.32	62.40	58.39	38.83	0
Difference	−3.3	16.66	17.61	14.13	22.72	41.99	N/A

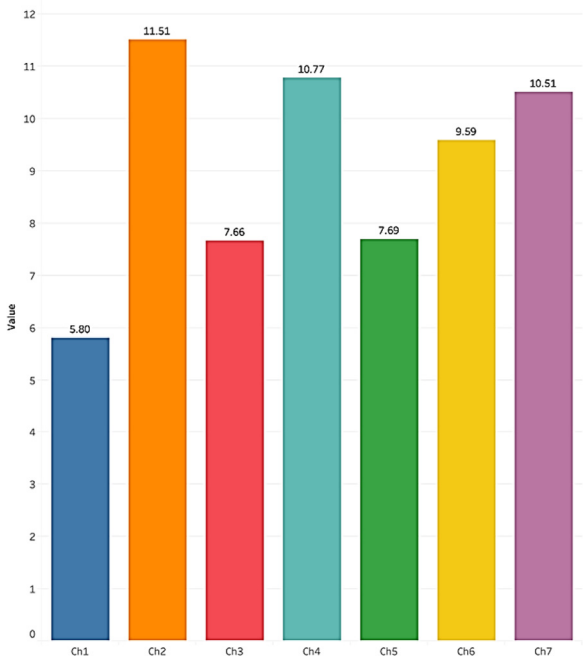


Fig. 6. Median playtime per chapter (hour).

there is a certain percentage of users who are currently playing the chapter. This percentage can be used as a benchmark for churners. The difference in completion rate is summarized in Table 9. By comparing the results in Tables 8 and 9, the following five interesting points are discovered.

1. Chapter 1 does not have a big difference between churners and non-churners. So this could be that the churners have almost played the entire contents of chapter 1 and stopped coming back.
2. Chapter 2 has the longest playtime for both churners and non-churners. However, the playtime difference between them is almost two hours. Also, the completion rate for churners decreased from 75.22% of chapter 1 to 50.35% of chapter 2 as shown in Table 9. This means many users had dropped in chapter 2. Therefore, chapter 2 may improve by shortening or simplify its lesson content.
3. Chapter 3 has only one hour difference between churners and non-churners, but the gap of completion rates is huge, which is 17.61%. This implies the churners possibly dropped out at the later of the chapter lesson. Therefore, the later contents for this chapter may need to be examined the necessity of update.
4. Chapter 4 has the second-highest playtime for churners and non-churners. Furthermore, the difference in playtime is larger. The gap indicates that the churners had stopped coming back at the earlier lessons in chapter 4, and the churners may have lost interest by longer playtime. Same as chapter 2, shortening or simplifying lesson content may boost the retention rate.
5. Chapter 5 and 6 have a bigger difference in completion rate, and chapter 6 and 7 has a bigger playtime difference between churners and non-churners. This may be caused by the smaller number of records that can be enhanced by adding the latest data.

4.3. Hyperparameter optimization

Hyperparameters are the parameters set before learning, and tuning hyperparameters affects the performance outcome of the model. Default settings were used for the previous data selection. However, it is important to tune hyperparameters for each model to possibly create a better performance model. For the categorical options of the hyperparameters, AUC results were compared. For the numeric hyperparameter options, there were two steps. First, the AUC results were plotted with a variety of values to make a broad estimation. Second, checking more specific values the same as the categorical values to discover the best hyperparameter values. The results reported in this section used the stratified 10-fold cross-validation using the training set only.

4.3.1. Decision Tree (DT)

For the optimization of DT, five hyperparameters were modified to see the performance difference. These hyperparameters are shown in Table 10.

The results of AUC with different criteria and splitter values are revealed in Fig. 7. The “entropy" value for *criterion* has better AUC, and the “best" value for *splitter* is better than “random" value. The performance using different tree depth values is illustrated in Fig. 8(a). After around six tree depths, the model performance clearly decreases, and after ten tree depths, the performance is about the same. This means that a bigger tree depth causes lower performance even though the computing cost is higher. So, between 4 and 8 are set as *max_depth* and tested again to see the AUC. Because the *criterion* with “entropy" value has better AUC, the *criterion* hyperparameter is set to “entropy". The outcome is summarized in Table 11. The highest AUC is at max depth = 6, so the best value for *max_depth* is set to 6.

As for *minimum_samples_split* and *minimum_samples_leaf*, they have a similar pattern as shown in Fig. 8(b) and Fig. 8(c), respectively. Generally, the more the hyperparameter value goes up, the lower AUC is output. This is understandable because more percentage of data is used for *minimum_samples_split* or *minimum_samples_leaf*; the split should be rougher. For the *minimum_samples_leaf*, the best performance is the smallest value, so the default should be used, which is equal to 1. Nevertheless, Fig. 8(b) is not convincing because the AUC reaches a peak of

Table 10
Decision tree hyperparameters settings.

Hyperparameter	Description	Values
criterion	It measures the quality of a split.	Gini and entropy
splitter	A strategy that is used to select the split at each node.	best and random
max_depth	The maximum depth of the tree.	[1 - 32]
minimum_samples_split	The minimum number of samples required to split.	The 0.01 to 0.5 are set for the comparison. 30 evenly spaced values between 0.01 and 0.5 are created and evaluated.
minimum_samples_leaf	The minimum number of samples required to be a leaf node.	1 and 30 values between 0.01 and 0.5 are set.

```
Decision Tree ==============================================
Gini Criteria AUC with Validation:  0.8492491456309462
Entropy Criteria AUC with Validation:  0.8533239233455893
Best Splitter AUC with Validation:  0.8492491456309462
Random Splitter AUC with Validation:  0.8183504809576065
```

Fig. 7. The result of AUC with different criteria and splitter values.

Table 11
The result of AUC for the decision tree with different *max_depth* values.

max_depth	4	5	6	7	8
AUC	0.9182	0.9286	**0.9290**	0.9286	0.9289

around 5%. The 5% of data is 142 by considering the number of observations of the training dataset for cross-validation is 2,846. Thus, from 25 to 200 increases by 25 are tested with cross-validation. The result is shown in Table 12. The best AUC is when *minimum_samples_split* equals 50, so the best value is set to 50. The summary of the best settings for DT is *criterion* = "entropy", *splitter* = best, *max_depth* = 6, *minimum_samples_split* = 50, and *minimum_samples_leaf* = 1.

4.3.2. Random Forest (RF)

Tuning RF is similar to DT because there are many same parameters. The *max_depth*, *minimum_samples_split*, and *minimum_samples_leaf* are the same, and the different hyperparameter is the number of trees in the forest (*n_estimators*). The result of AUC with different estimators is illustrated in Fig. 9(a). The AUC reaches closest to 1 about 100 estimators. To see the precise value of AUC, the values 50, 75, 100, and 120 were tested again for estimators. The result is shown in Table 13. The AUC is the highest with 100 estimators, and it does not improve after that according to the plot. Having more trees requires more computing costs. Therefore, the best value for *n_estimators* is 100.

Next, the performance of *max_depth* values is illustrated in Fig. 9(b). The AUC reaches the maximum of around 10. So the *max_depth* value is set from 7 until 13 and the *n_estimators* is set to 100. The outcome is in Table 14. The *max_depth* at 10 produces the highest AUC which is acceptable by considering the earlier AUC plot. So, 10 is the best value for *max_depth*. The other two hyperparameters have a similar pattern as with DT. The results are displayed in Fig. 9(c) and Fig. 9(d). Both the *minimum_samples_split* and *minimum_samples_leaf* have the same pattern in which the AUC decreases as the value increases. Thus, the best value is the smallest value. The summary of the best settings for RF is: *n_estimators* = 100, *splitter* = best, *max_depth* = 10, *minimum_samples_split* = 2, and *minimum_samples_leaf* = 1.

4.3.3. Logistic Regression (LR)

Two hyperparameters were optimized for LR. The first one is the penalty which is a type of normalization used in penalization. The second hyperparameter is the inverse of regularization strength (C); a smaller value means stronger regularization. First, the performance with different types of penalties was checked, and the results are displayed in Fig. 10. As shown in Fig. 10, "L1" has the highest AUC value. Therefore, "L1" was chosen.

For C, thirty evenly spaced values between 0.01 to 1.5 were set to C as shown in Fig. 11. The AUC reaches the highest at around 1, and the AUC does not change after that. To make sure what value is the best with precise value, the AUC value at 0.9, 1.0, 1.1, and 1.2 were evaluated, and the results are summarized in Table 15. The AUC is the same for all C values, so the value chosen for C is 0.9. The hyperparameters and the final choices are shown in Table 15. The summary of the best settings for LR is: penalty = "L1" and C = 0.9.

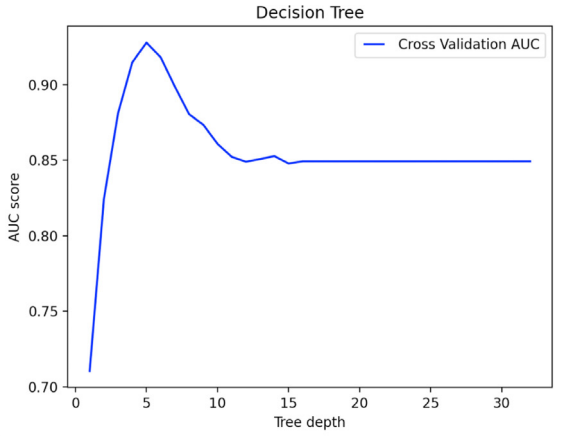
4.4. Churn prediction and model evaluation

The final model of each algorithm was created with the best hyperparameters reported in the previous subsections. For the last evaluation, the dataset used for modeling was the same split ratio of 90% training and 10% test. Also, the principal components were used. The difference is that the evaluation was conducted to output AUC, confusion matrix, and classification report, which outputs precision, recall, and F1-score. The reason for checking the other metrics is that the AUC does not cover the whole factors behind. Precision means the ratio of correctly predicted positive cases to the total of predicted positive cases, and recall is the proportion of actual positive observations that are correctly predicted positive observations. Generally, precision and recall are in tension. So, improving one metric causes reducing the other. F1-score is a metric to see the balance between precision and recall.

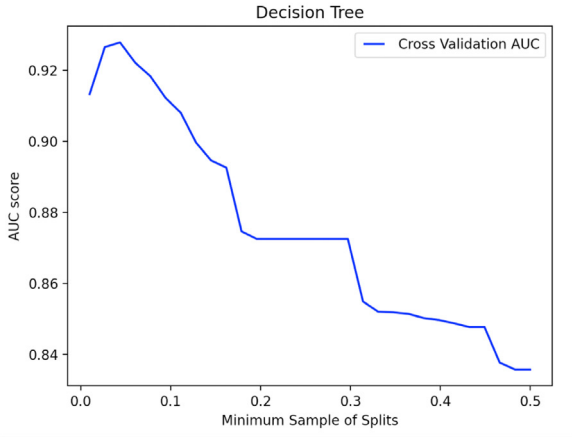
First, the confusion matrix of each model is plotted as shown in Fig. 12. True Positive (TP) is correctly predicted an actual churner as a churner, True Negative (TN) is correctly predicted an actual non-churner as a non-churner. The overall prediction results of the three models can be said that predictions are mostly correct because the number of TP and TN are more than the majority without checking other metrics. The number of TP is

Table 12
The result of AUC for the decision tree with different *minimum_samples_split* values.

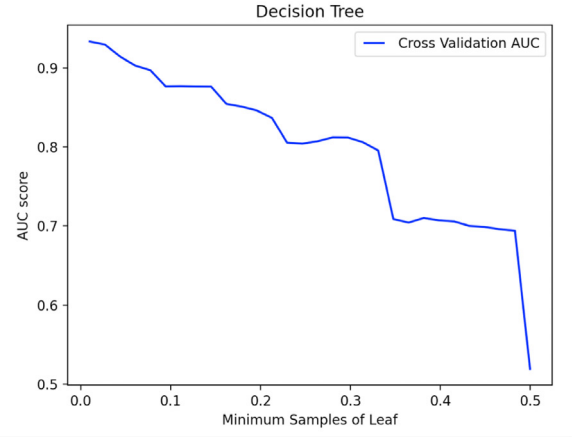
Minimum split	25	50	75	100	125	150	175	200
AUC	0.9317	**0.9333**	0.9320	0.9329	0.9290	0.9284	0.9258	0.9262



(a) With different max_depth values.



(b) With different $minimum\_samples\_split$ values.



(c) With different $minimum\_samples\_leaf$ values.

Fig. 8. Hyperparameter optimization results for decision tree model.

greater than TN, which is due to the imbalanced target value. This applies to the relationship of False Negative (FN) and False Positive (FP).

Next, by using the classification report, specific values were checked. The reports are shown in Fig. 13. According to the result,

Table 13
The result of AUC for the random forest with different estimators.

n_estimators	50	75	100	125
AUC	0.9570	0.9590	**0.9605**	0.9603

the total observation is 356, the churners (1 in the report) are 232, non-churners (0 in the report) are 124. This means 65.16% of test data are churners which are higher than the actual percentage of the whole data of 56.7%. Therefore, the results will be more biased. There are "macro avr" and "weighted avr" in the classification report. The "weighted avr considers the imbalanced proportion and calculates the precision, recall, and F1-score with weight. Since the proportion of the target is not balanced, the "weighted avr" is chosen to be compared for evaluation.

The summary of the metrics of the three models is presented in Table 16. The precision, recall, and F1-score were taken from "weighted avr". The LR model produced the highest AUC and F1-score. The F1-score of 0.9194 is high, which means the recall and precision are well balanced. Although DT and RF have good performance of AUC, which is 0.8452 and 0.8493, their prediction performances are lower than the LR model. Tree-based algorithm characteristics may cause the similarity of results of DT and RF.

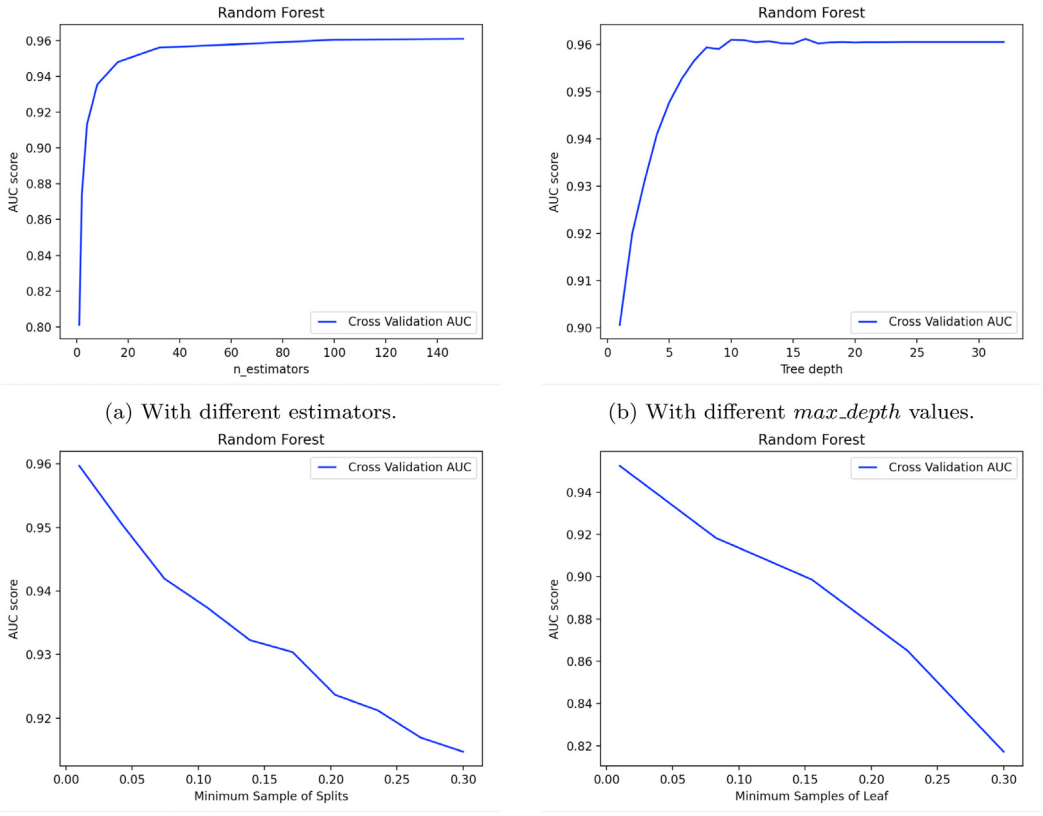
Considering the prediction performances of past studies, the AUC and F1-score are in a higher range in this work. Most studies used AUC for the performance metrics, and the popular range is between the late seventies to early nineties, which are summarized in Tables 2 and 3. As for the F1-score, a study conducted churn prediction for freemium games measured F1-score as a performance metric, and the highest model of F1-score was DT with 0.916 [21]. Thus, the LR model has an excellent prediction performance even considering the other studies in the gaming and education industries.

The reasons LR showed the best performance could be considered as follows based on the advantages or disadvantages of LR, DT, and RF [49]. First, LR tends to overfit with high dimensional data, but a low dimensional dataset was used for this examination. Also, LR works well with linearly separatable datasets, whereas DT has less efficiency with continuous numerical data. Many of the used dataset variables are continuous numerical variables (such as minutes and days) with a wide-range and more significant gap between mean and standard deviation, so this could affect the tree-based models, and worked better with LR model than DT and RF.

4.5. Discussion

The combination of churn determination and prediction presented above showed the effectiveness of the proposed approach in DGBL. First, defining churn based on the arbitrary time frame for each user by comparing the recency and the addition of average and two standard deviations of user inactive time was concluded as the best approach. Then, the churn prediction with the best performance was using LR model with penalty set to "L1" and C set to "0.9".

The applicability for DGBL services is also considered for the above approach. The defined churn calculation can be helpful to many DGBL services even if there are various course duration or no specified duration. This is because the recency and inactive time between logins of each user should be available or calculable



(a) With different estimators.

(b) With different max_depth values.

(c) With different $minimum\_samples\_split$ values.

(d) With different $minimum\_samples\_leaf$ values.

Fig. 9. Hyperparameter optimization results for random forest model.

Table 14
The result of AUC for the random forest with different *max_depth* values.

max_depth	7	8	9	10	11	12	13
AUC	0.9565	0.9594	0.9591	**0.9610**	0.9609	0.9605	0.9607

```
Logistic Regression ====================
Penalty L1 AUC :  0.9745084802343669
Penalty L2 AUC :  0.7807881974415437
Penalty None AUC :  0.7807922014455476
```

Fig. 10. The result of AUC with different penalty values.

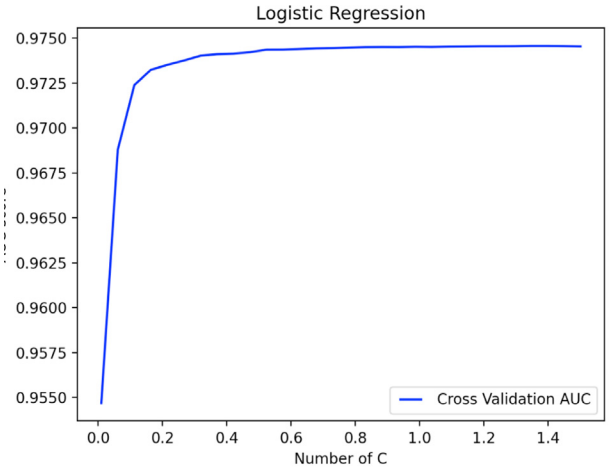


Fig. 11. The result of AUC with different C values.

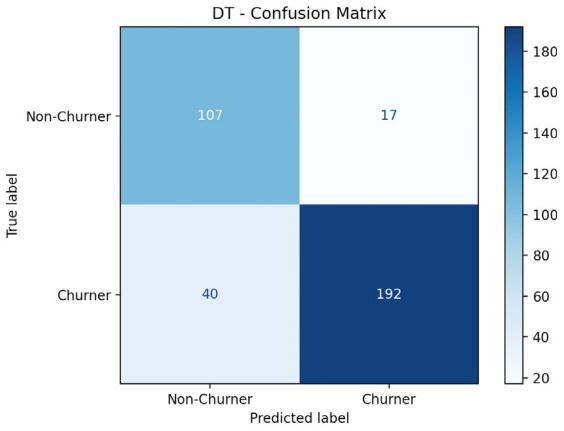
Table 15
The result of AUC for the logistic regression with different C values.

C	0.9	1	1.1	1.2
AUC	**0.9745**	**0.9745**	**0.9745**	**0.9745**

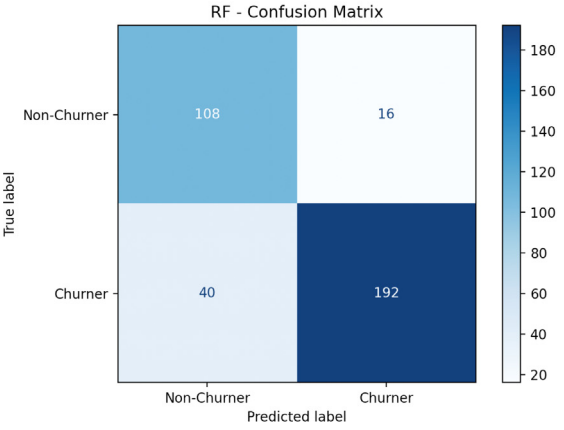
Table 16
Results summary.

Metrics/Algorithms	Decision tree	Random forest	Logistic regression
Precision	0.8522	0.8557	**0.9228**
Recall	0.8399	0.8427	**0.9185**
F1-Score	0.8425	0.8453	**0.9194**
AUC	0.8452	0.8493	**0.9225**

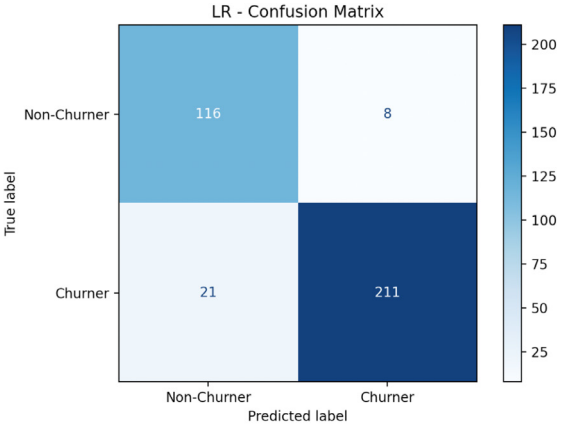
for most of the services. In addition, this approach allows us to have the flexibility to define churners even if the users are playing in mid-course. Since having a fixed evaluation time and cutoff like the online gaming industry or having a clear definition of a drop-off with various evaluation periods was challenging in DGBL, this applicability and flexibility of the proposed churn determination can address the issues. Also, for the input of the LR model, three categories of input variables were used. The categories are demographic, engagement, and performance, which are common values for DGBL, online gaming, and MOOCs. Although the detailed variables and hyperparameters can be different from service to service, the input categories for better churn prediction



(a) Decision tree.



(b) Random forest.



(c) Logistic regression.

Fig. 12. Confusion matrix results.

```
Decision Tree ========================================
[Test] ROC AUC of DT:  0.8452447163515017
▼▼▼ [Test] Classification Report of DT ▼▼▼
              precision    recall  f1-score   support

           0     0.7279    0.8629    0.7897       124
           1     0.9187    0.8276    0.8707       232

    accuracy                         0.8399       356
   macro avg     0.8233    0.8452    0.8302       356
weighted avg     0.8522    0.8399    0.8425       356
```

(a) Decision tree.

```
Random Forest ========================================
[Test] ROC AUC of RF:  0.8492769744160178
▼▼▼ [Test] Classification Report of RF ▼▼▼
              precision    recall  f1-score   support

           0     0.7297    0.8710    0.7941       124
           1     0.9231    0.8276    0.8727       232

    accuracy                         0.8427       356
   macro avg     0.8264    0.8493    0.8334       356
weighted avg     0.8557    0.8427    0.8453       356
```

(b) Random forest.

```
Logistic Regression ==================================
[Test] ROC AUC of LR:  0.9224833147942157
▼▼▼ [Test] Classification Report of LR ▼▼▼
              precision    recall  f1-score   support

           0     0.8467    0.9355    0.8889       124
           1     0.9635    0.9095    0.9357       232

    accuracy                         0.9185       356
   macro avg     0.9051    0.9225    0.9123       356
weighted avg     0.9228    0.9185    0.9194       356
```

(c) Logistic regression.

Fig. 13. Prediction results. Non-churner class is represented by (0) and the churner class by (1). The values 124 and 232 in the support column represents the number of samples for non-churner and churner, respectively.

and indicators of hyperparameters in DGBL were proposed. Thus, the proposed methodology of churn determination and prediction can be implemented with any DGBL company.

5. Conclusion and future work

DGBL is a narrowed-down category of EdTech and a digital version of game-based learning, an approach to facilitate learning with gameplay. One of the common issues in the EdTech market is the higher churn rate. Retention and churn rate are some of the common performance key indicators for business. Keeping customers brings more profit because it is less expensive than acquiring new customers, and loyal customers tend to spend more money on the service. Therefore, understanding and analyzing retention and churn is important for a business.

To address the issue, a marketing approach is becoming more important. However, because the market is still in the early stage, there are few studies related to marketing perspectives even though there are abundant companies in the EdTech and DGBL industries. In addition, the approach in education or online gaming industries can be only partially applicable to DGBL because of the difference in the definition of churn, evaluation time frame for churners, and social interaction.

One of the popular approaches for retention management is churn prediction. Churn prediction allows companies to create better marketing strategies to improve user retention. Using a dataset from a Japanese company providing DGBL service as a case study, an approach for churn prediction for DGBL is proposed. There are three objectives for this work. The first one is the determination of churn, which is defined by applying an arbitrary time frame for each user. The calculation compares the

Table A.1
Explained variance of principal components.

	total_login	*total_playtime* (min)	*total_inactive* (min)	*average_playtime* (min)	*average_inactive* (min)	*entire_period* (days)	*ch1_playtime* (min)	*ch2_playtime* (min)	*ch3_playtime* (min)
PC-1	2.00E−06	1.00E−06	1.00E+00	−6.69E−07	1.00E−06	0.0007	4.46E−07	0.000002	1.00E−06
PC-2	−5.00E−04	0.00023	−7.88E−06	1.23E−03	−0.002	0.0099	2.32E−03	0.002199	7.00E−05
PC-3	0.0102	0.01323	−2.06E−06	5.76E−03	−0.001	0.0061	5.67E−03	0.012157	0.0116
PC-4	0.1663	0.18799	−6.05E−04	1.75E−02	−0.096	0.8606	8.42E−02	0.116752	0.1384
PC-5	0.2719	0.30871	3.52E−04	3.62E−02	−0.129	−0.506	7.84E−02	0.161545	0.2256
PC-6	−0.029	0.03238	−2.08E−05	1.11E−01	0.4419	0.0295	1.32E−01	−0.382636	−0.242
PC-7	0.1508	−0.03262	−7.21E−06	−7.75E−01	−0.159	0.011	−5.39E−02	−0.169805	−0.134
PC-8	0.0694	0.0559	1.86E−05	7.53E−02	−0.246	−0.026	7.83E−01	0.17535	0.0091
PC-9	0.1422	0.02436	7.13E−06	−4.70E−01	−0.128	−0.012	−9.13E−02	0.15487	0.1345
PC-10	−0.069	−0.02935	−2.39E−05	1.63E−01	0.0029	0.0344	−4.78E−01	0.11993	0.2582
PC-11	0.1205	0.08686	8.32E−07	−2.45E−01	0.7706	−0.004	2.10E−01	0.093908	0.4173
PC-12	0.056	0.05772	7.58E−07	2.64E−02	0.098	−0.001	−1.33E−01	0.331262	0.1194
PC-13	−0.14	−0.0956	−3.09E−06	−4.76E−02	−0.119	0.005	1.27E−01	−0.655647	0.2889
PC-14	−0.081	0.01833	2.38E−06	6.26E−02	−0.208	−0.002	−2.28E−02	−0.229848	0.6743
PC-15	−0.047	−0.01299	2.87E−06	5.21E−03	−0.034	−0.004	1.01E−02	0.106602	0.0718
PC-16	0.1724	0.04301	−1.46E−06	1.73E−02	−0.034	0.0022	2.07E−02	−0.059972	−0.038
PC-17	0.4638	0.08592	8.15E−06	2.14E−01	−0.102	−0.011	−2.69E−02	−0.170329	0.0664
PC-18	0.7444	−0.25832	−5.03E−06	1.41E−01	0.0578	0.0068	−1.21E−01	−0.16183	−0.076
PC-19	−0.007	0.00207	2.78E−07	−3.74E−04	0.0051	−5.00E−04	−5.72E−03	0.008591	0.003

	ch4_playtime (min)	*ch5_playtime* (min)	*ch6_playtime* (min)	*ch7_playtime* (min)	*avr_ch_wait* (days)	*exp*	*coins*	*replay*
PC-1	0.000001	1.00E−06	8.98E−07	5.97E−07	7.69E−07	2.00E−06	2.00E−06	3.49E−07
PC-2	−0.001188	−0.001	4.60E−04	−1.01E−03	−9.65E−04	0.001	0.0012	−1.27E−03
PC-3	0.007957	0.0091	1.01E−02	8.35E−03	1.47E−03	0.0128	0.012	−1.08E−03
PC-4	0.147524	0.1502	1.50E−01	1.33E−01	−3.16E−02	0.1635	0.1639	8.41E−02
PC-5	0.281479	0.2694	2.48E−01	2.29E−01	−7.15E−02	0.2839	0.2834	2.03E−01
PC-6	0.168011	0.1533	6.46E−02	2.74E−01	2.91E−01	−0.157	−0.149	5.52E−01
PC-7	0.072236	0.0241	−1.11E−02	8.13E−02	−4.48E−01	−0.139	−0.129	2.24E−01
PC-8	0.024768	−0.171	−3.29E−01	−2.21E−01	−2.78E−02	−0.111	−0.1	2.66E−01
PC-9	0.062961	−0.065	−7.69E−02	−8.02E−02	8.19E−01	−0.007	−0.01	2.76E−03
PC-10	0.440438	−0.146	−3.06E−01	−3.69E−01	−1.23E−01	−0.041	−0.04	4.40E−01
PC-11	0.02843	−0.056	−8.19E−03	−2.00E−01	−1.33E−01	0.0109	0.0045	−1.87E−01
PC-12	−0.211524	−0.392	−3.56E−01	7.06E−01	−5.77E−02	−0.041	−0.047	9.85E−02
PC-13	0.161819	−0.402	−1.43E−01	1.26E−01	3.37E−02	0.3078	0.3135	−8.56E−02
PC-14	−0.39824	0.3272	4.34E−02	4.93E−02	7.15E−03	−0.269	−0.268	1.36E−01
PC-15	−0.140912	−0.587	7.19E−01	−1.07E−01	−4.31E−03	−0.063	−0.058	2.69E−01
PC-16	0.202304	−0.021	−1.09E−02	2.98E−02	−1.19E−02	−0.113	−0.12	−1.88E−01
PC-17	0.377125	−0.183	9.63E−02	1.09E−01	7.23E−03	−0.358	−0.347	−3.43E−01
PC-18	−0.409904	−0.036	−1.08E−01	−1.91E−01	8.90E−04	0.1577	0.1674	1.94E−01
PC-19	−0.001526	0.0035	−9.50E−04	1.98E−03	2.27E−03	−0.704	0.7096	−7.52E−03

	gender	*age*	*prefecture*
PC-1	−3.25E−08	−2.00E−06	−6.73E−07
PC-2	1.20E−03	−0.104	−9.94E−01
PC-3	−2.74E−03	0.9939	−1.04E−01
PC-4	7.86E−03	−0.021	1.14E−02
PC-5	−2.73E−04	−0.025	−1.88E−03
PC-6	−1.61E−02	0.0043	−3.57E−03
PC-7	4.50E−02	0.0098	−2.40E−03
PC-8	1.07E−02	0.001	2.21E−03
PC-9	1.06E−02	−0.002	−9.19E−04
PC-10	−2.61E−02	0.0037	−1.34E−03
PC-11	1.46E−02	−0.005	−2.80E−04
PC-12	−3.76E−02	−0.002	2.31E−05
PC-13	−4.20E−03	0.0016	−1.18E−04
PC-14	−7.22E−02	0.0009	−8.36E−04

(*continued on next page*)

Table A.1 (*continued*).

	gender	*age*	*prefecture*
PC-15	−8.61E−02	0.0002	1.04E−03
PC-16	−9.26E−01	−0.003	−1.11E−03
PC-17	3.53E−01	0.0003	−5.30E−04
PC-18	−3.58E−02	0.0009	−3.28E−04
PC-19	−7.05E−03	0.0004	1.07E−04

recency and the addition of average and two standard deviations of user inactive time. The second objective is to clarify the churn rate of the Japanese service. Using the defined churn definition, the churn rate became evident as 56.77% with the case study data. In addition, the descriptive analysis and retention analysis were conducted. The descriptive analysis uncovered the MAU and popular dates and times of the users. The retention analysis reveals the different behavior of churners and non-churners per chapter. The third objective is to develop a churn prediction model by comparing LR, DT, and RF models. Feature selection, dataset split ratio comparison, and hyperparameter tuning were conducted to build the best performance models. The LR model has a distinctively highest AUC of 0.9225 and an F1-score of 0.9194. The higher F1-score means the recall and precision are well balanced. The 0.9225 of AUC is on the higher side comparing with the AUC of past churn prediction studies in online gaming and education industries.

There was no common determination of churn and churn prediction in DGBL because of lack of research. The variety of course duration makes it difficult to apply the common gaming churn determination approach, and the commonly used data types are non-user behavior data such as on-campus and demographics data in education. However, the proposed approach can address the issue. The values used to determine churn are using only users' recency and inactive time between logins. This can be available or calculable in most of the DGBL services. Also, the churn determination approach is applicable to the variety of course duration and even new or ongoing courses. The model's input values are common three categories: demographic, engagement, and performance in DGBL. Thus, this research establishes the path of churn determination and prediction in DGBL and even gamification companies to conduct better marketing strategies. As a consequence, the results indicate the effectiveness of the proposed approach of determination of churn and churn prediction in DGBL.

A limitation of the proposed approach is that the defining churn is not suitable for the users who just have started playing. The calculation requires the mean inactive time and standard deviation inactive time, so users with a couple of logins might have less reliable results. The prediction of early churn will not be suitable for this approach. Another limitation is data size. The number of observations used for the churn prediction model is 3,557, and this is considered a smaller dataset. The LR was the best performance model with the case study, but other tree-based algorithms had good results. This implies the other models have the possibility of improving the performance with different dataset sizes and, of course, input variables. Hence, the proposed prediction model may produce different results if the dataset size is huge.

Considering the above two limitations, the procedure can be further developed. Defining the early churn determination and prediction for a shorter course is one possible future work. As for the larger dataset, RF may have better performance. Alternatively, different models such as neural networks and deep neural networks can be tested. Another approach with a bigger dataset is to try chapter-based churn prediction, which could help create more specific marketing strategies for each chapter churner.

CRediT authorship contribution statement

Mai Kiguchi: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Waddah Saeed:** Conceptualization, Methodology, Writing – review & editing, Supervision. **Imran Medi:** Conceptualization, Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix. Explained variance of principal components

The explained variance of principal components is shown in Table A.1, which represents the amount of variance explained by each of the selected components.

References

[1] H. HolonIQ, 10 Charts that explain the global education technology market, 2019, URL: https://www.holoniq.com/edtech/10-charts-that-explain-the-global-education-technology-market/.
[2] S. Perini, R. Luglietti, M. Margoudi, M. Oliveira, M. Taisch, Learning and motivational effects of digital game-based learning (DGBL) for manufacturing education –The life cycle assessment (LCA) game, Comput. Ind. 102 (2018) 40–49.
[3] V. Strauss, New report on virtual education: 'it sure sounds good. As it turns out, it's too good to be true, 2019, URL: https://www.washingtonpost.com/education/2019/05/29/new-report-virtual-education-it-sure-sounds-good-it-turns-out-its-too-good-be-true.
[4] Recurly, Benchmarks for subscription E-commerce, 2019, URL: https://info.recurly.com/research/benchmarks-for-subscription-ecommerce.
[5] X. Fu, X. Chen, Y.-T. Shi, I. Bose, S. Cai, User segmentation for retention management in online social games, Decis. Support Syst. 101 (2017) 51–68.
[6] C. Marquez-Vera, C.R. Morales, S.V. Soto, Predicting school failure and dropout by using data mining techniques, IEEE Revista Iberoamericana De Tecnologias Del Aprendizaje 8 (1) (2013) 7–14.
[7] M.L. Bote-Lorenzo, E. Gómez-Sánchez, Predicting the decrease of engagement indicators in a MOOC, in: Proceedings of the Seventh International Learning Analytics & Knowledge Conference, in: LAK, vol. 17, Association for Computing Machinery, New York, NY, USA, 2017, pp. 143–147.
[8] E. Sanchez, Game-based learning, in: A. Tatnall (Ed.), Encyclopedia of Educ. Inf. Technol., Springer International Publishing, Cham, 2019, pp. 1–9.
[9] J.P. Gee, What video games have to teach us about learning and literacy, Comput. Entertain. (CIE) 1 (1) (2003) 20.
[10] M. Prensky, Digital game-based learning, Comput. Entertain. (CIE) 1 (1) (2003) 21.
[11] A. Khan, F.H. Ahmad, M.M. Malik, Use of digital game based learning and gamification in secondary school science: The effect on student engagement, learning and gender difference, Educ. Inf. Technol. 22 (6) (2017) 2767–2804.
[12] I.R.M. Association, et al., Gamification: Concepts, Methodologies, Tools, and Applications, IGI Global, 2015.
[13] S. Egenfeldt-Nielsen, B. Meyer, B.H. Soerensen, Serious Games in Education: a Global Perspective, ISD LLC, 2011.
[14] Y. Zhonggen, A meta-analysis of use of serious games in education over a decade, Int. J. Comput. Games Technol. 2019 (2019).
[15] J. Byun, E. Joung, Digital game-based learning for K–12 mathematics education: A meta-analysis, Sch. Sci. Math. 118 (3–4) (2018) 113–126.

[16] M.H. Hussein, S.H. Ow, L.S. Cheong, M. Thong, N. Ale Ebrahim, Effects of digital game-based learning on elementary science learning: A systematic review, IEEE Access 7 (2019) 62465–62478.
[17] A. Gallo, The value of keeping the right customers, Harv. Bus. Rev. 29 (2014) 2014.
[18] C. Yang, X. Shi, L. Jie, J. Han, I know you'll be back: Interpretable new user clustering and churn prediction on a mobile social application, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, in: KDD, vol. 18, Association for Computing Machinery, New York, NY, USA, 2018, pp. 914–922.
[19] E. Suh, M. Alhaery, Customer retention: Reducing online casino player churn through the application of predictive modeling, UNLV Gaming Res. Rev. J. 20 (2) (2016) 6.
[20] E. Lee, B. Kim, S. Kang, B. Kang, Y. Jang, H.K. Kim, Profit optimizing churn prediction for long-term loyal customer in online games, IEEE Trans. Games (2018).
[21] F. Hadiji, R. Sifa, A. Drachen, C. Thurau, K. Kersting, C. Bauckhage, Predicting player churn in the wild, in: 2014 IEEE Conference on Computational Intelligence and Games, IEEE, 2014, pp. 1–8.
[22] H. Xie, S. Devlin, D. Kudenko, P. Cowling, Predicting player disengagement and first purchase with event-frequency based data representation, in: 2015 IEEE Conference on Computational Intelligence and Games, CIG, IEEE, 2015, pp. 230–237.
[23] M. Tamassia, W. Raffe, R. Sifa, A. Drachen, F. Zambetta, 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, IEEE, 2016, pp. 1–8.
[24] T. Banerjee, G. Mukherjee, S. Dutta, P. Ghosh, A large-scale constrained joint modeling approach for predicting user activity, engagement, and churn with application to freemium mobile games, J. Am. Stat. Assoc. 115 (530) (2020) 538–554.
[25] M. Milošević, N. Živić, I. Andjelković, Early churn prediction with personalized targeting in mobile social games, Expert Syst. Appl. 83 (2017) 326–332.
[26] Y. Levy, Comparing dropouts and persistence in e-learning courses, Comput. Educ. 48 (2) (2007) 185–204.
[27] C. Sorensen, J. Donovan, An examination of factors that impact the retention of online students at a for-profit university, Online Learning 21 (3) (2017) 206–221.
[28] M.A. Bingham, N.W. Solverson, Using enrollment data to predict retention rate, J. Stud. Aff. Res. Practice 53 (1) (2016) 51–64.
[29] S. Gupta, A.S. Sabitha, Deciphering the attributes of student retention in massive open online courses using data mining techniques, Educ. Inf. Technol. 24 (3) (2019) 1973–1994.
[30] K.S. Hone, G.R. El Said, Exploring the factors affecting MOOC retention: A survey study, Comput. Educ. 98 (2016) 157–168.
[31] W. Xing, X. Chen, J. Stein, M. Marcinkowski, Temporal predication of dropouts in MOOCs: Reaching the low hanging fruit through stacking generalization, Comput. Hum. Behav. 58 (2016) 119–129.
[32] Z. Niu, W. Li, X. Yan, N. Wu, Exploring causes for the dropout on massive open online courses, in: Proceedings of ACM Turing Celebration Conference-China, 2018, pp. 47–52.
[33] A. Yan, M.J. Lee, A.J. Ko, Predicting abandonment in online coding tutorials, in: 2017 IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC, IEEE, 2017, pp. 191–199.
[34] X. Liu, M. Xie, X. Wen, R. Chen, Y. Ge, N. Duffield, N. Wang, A semi-supervised and inductive embedding model for churn prediction of large-scale mobile games, in: 2018 IEEE International Conference on Data Mining, ICDM, IEEE, 2018, pp. 277–286.
[35] E. Loria, A. Marconi, Exploiting limited players' behavioral data to predict churn in gamification, Electron. Commer. Res. Appl. 47 (2021) 101057, http://dx.doi.org/10.1016/j.elerap.2021.101057.
[36] E.B. Costa, B. Fonseca, M.A. Santana, F.F. de Araújo, J. Rego, Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses, Comput. Hum. Behav. 73 (2017) 247–256.
[37] E. Yukselturk, S. Ozekes, Y.K. Türel, Predicting dropout student: an application of data mining methods in an online education program, Eur. J. Open Distance E-Learning 17 (1) (2014) 118–133.
[38] L.P. Khobragade, P. Mahadik, Students' academic failure prediction using data mining, Int. J. Adv. Res. Comput. Commun. Eng. 4 (11) (2015) 290–298.
[39] Q. Fu, Z. Gao, J. Zhou, Y. Zheng, CLSA: A novel deep learning model for MOOC dropout prediction, Comput. Electr. Eng. 94 (2021) 107315, http://dx.doi.org/10.1016/j.compeleceng.2021.107315.
[40] W.R. Smith, Product differentiation and market segmentation as alternative marketing strategies, J. Mark. 21 (1) (1956) 3–8.
[41] Z.G. Yi, Marketing Services and Resources in Information Organizations, Chandos Publishing, 2017.
[42] J. Han, J. Pei, M. Kamber, Data Mining: Concepts and Techniques, Elsevier, 2011.
[43] M. Utashiro, S. Hashimoto, A. Hayashi, Research on acquisition of hiragana in normally developing children : Focused on current acquisition state and instability, Bull. Tokyo Gakugei Univ. Educ. Sci. 66 (2) (2015) 397–402.
[44] L. Beachum, World's oldest person breaks her own record by turning 117, The Washington post, 2020, URL: https://www.washingtonpost.com/world/2020/01/06/worlds-oldest-woman-breaks-her-own-record-by-turning/.
[45] T.P. Minka, Automatic choice of dimensionality for PCA, in: Advances in Neural Information Processing Systems, 2001, pp. 598–604.
[46] K. Coussement, K.W. De Bock, Customer churn prediction in the online gambling industry: The beneficial effect of ensemble learning, J. Bus. Res. 66 (9) (2013) 1629–1636, http://dx.doi.org/10.1016/j.jbusres.2012.12.008, Advancing Research Methods in Marketing.
[47] W. Mao, F. Wang, Cultural modeling for behavior analysis and prediction, in: New Advances in Intelligence and Security Informatics, first ed., Academic Press, Waltham, MA, USA, 2012, pp. 91–102.
[48] L.A. Jeni, J.F. Cohn, F. De La Torre, Facing imbalanced data–recommendations for the use of performance metrics, in: 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction, IEEE, 2013, pp. 245–251.
[49] A. Khanna, D. Gupta, N. Dey, Applications of Big Data in Healthcare, Academic Press, 2021.