Proceedings of the Seventeenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment
(AIIDE 2021)

Optimizing Profit by Mitigating Recurrent Churn Labeling Issues: Analysis from the Game Domain

João Felipe Humenhuk, Luiz Bernardo Martins Kummer, Emerson Cabrera Paraiso
Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná
R. Imac. Conceição, 1155
Curitiba, Paraná, Brazil 80215-901
{joaofh, luiz.kummer, paraiso}@ppgia.pucpr.br

Abstract

Churn can be interpreted as customer defection and can be considered one of the most critical challenges in the Game Analytics domain because of its impact on the game industry's profit. When predicting churn, the first step is defining what is considered churn, which can change depending on the players' behaviors and approaches. This work studied related works and revealed two recurrent issues in the labeling process: limitations on the adopted labeling approaches (1) and the static definition of churn (2). To mitigate the first issue, an individualized labeling approach was deployed. To address the second one, a novel evaluation method, based on the impact of a change in the churn definition, was proposed. This method allowed the proposition of two new labeling approaches, which were included in the analysis. By comparing the labeling approaches in two games using a profit perspective, it was identified that the new ones present statistically significant benefits compared to the traditional ones. Regarding the evaluation method, its usage can justify when the redefinition of churn and the classifier's retraining should happen to improve profit. The results are valuable for the game context, potentially extended to other contexts by delivering more reliable labels and more validated classification performance.

Introduction

The game management evolved from a perspective in which game development is considered finished after its release to a continuous development perspective without an apparent dead end, the idea of Games as Services (GaaS) (Clark 2014). The GaaS management approach considers that the players' motivations can be identified and maintained over time through Game Analytics analysis associated with the release of new game content. Such new management can keep the active players playing longer and motivate new ones to start playing, resulting in a profit increase (El-Nasr, Drachen, and Canossa 2016).

The profitability of the GaaS policy has many sources, such as in-game purchases, subscriptions, and the game purchase itself. The game purchase source relies on the players' initial interest in playing the game, whereas the other sources are related to the continuous enjoyment in playing the game (Kummer, Nievola, and Paraiso 2017). Furthermore, the GaaS approaches carry additional game management challenges, such as user profiling, game upgrades management, and churn prediction. Churn is described as customer defection. It can be considered one of the most critical challenges when using such approaches because the industry profit comes from the players playing the game (Hadiji et al. 2014). Besides, it is six times cheaper to maintain active players than acquire new ones (Yuan et al. 2017), highlighting the importance of retaining possible churners.

Although churn is an old problem present in different businesses (e.g., telecom, TV, banking, or games), the first models to predict it using machine learning started at the end of the twentieth century (Ahn et al. 2020). Churn in entertainment games differs from other contexts due to its nature of voluntary usage, where the act of playing is not attached to any responsibility. This characteristic provides an additional challenge, as it implies that a player can stop or return to play at any time and without notice (a non-contractual bond) (Tamaddoni, Stakhovych, and Ewing 2016). This characteristic adds complexity to the classification task and the labeling. Instead of proposing a new algorithm to classify churning better, this article aims to question the way datasets are labeled regarding churn. Possible issues in the labeling process can mask bad results in churn prevention campaigns behind a good classifier's performance. Churn retaining campaign encompasses numerous ways the producers reach the players and try to maintain them playing. Since a campaign's cost (e.g., marketing, design, and time) is related to the number of target players, wrong labeling can lead to money loss. Even though some approaches, such as in-game rewards donation, do not have a cost for the producers, they represent a profit loss if they start giving them to non-churning players that would instead buy them.

To understand and mitigate the existing problems in players' churn labeling, we propose two Research Questions (RQs). The first step is to identify the challenges, represented by RQ1: "What are the recurrent issues present on the churn labeling task?". This question is answered based on related works, including methods, procedures, encountered challenges, and future works. This analysis identified two possible problems: limitations of the approaches used and a static Churn Definition (CD) (e.g., assuming the definition "after three days absent, a player is considered a churner" for a whole game's life cycle). The second step aims at verifying if they are indeed issues through the following RQ2: "How to measure the need to redefine churn?". A novel evaluation method is proposed to answer it, capable of quantitatively comparing the labeling approaches regarding the impact of a change in the CD over time. Also, an individualized labeling approach is deployed, and two novel labeling approaches are proposed to cover the weaknesses identified by RQ1.

From the insights of this paper, other researchers from industry or academia can apply, explore, and take advantage of the proposed evaluation method and labeling approaches on their acting contexts (such as telecom, fast-food, finance, among others). On the one hand, the new labeling approaches provide more reliable results for the churn prediction task than commonly used ones, which has the potential to increase the profit of churn retaining campaigns. On the other hand, the evaluation method can assess the need to relabel the dataset and retrain the churn prediction classifier. The proposed evaluation method can also be used over new churn labeling approaches or other games that were not approached in this work.

This article's structure starts with the related works and the most common labeling approaches analysis, followed by the descriptions of an individualized labeling approach and the considered datasets. Next, the evaluation method, experimental protocol, novel labeling approaches, and results are presented. Later, the results are shown and discussed, and the conclusions are given together with future works.

Related Works

Numerous works predict churn in various domains (Ahn et al. 2020). Focusing on games, some utilize only the time spent playing (Milošević, Živić, and Andjelković 2017; Kummer, Nievola, and Paraiso 2018), others use social aspects (Liu et al. 2019), some see the problem as time-series (Yang et al. 2019), others utilize Natural Language Processing (Kilimci, Yörük, and Akyokus 2020). They use data from different game genres such as Multiplayer Online Battle Arena (MOBA) and Massively Multiplayer Online Role-Playing Game (MMORPG), from different platforms, like desktop and mobile, and target diverse types of players. All of them provide useful information about algorithms and techniques to classify the players as churners or non-churners. Still, they do not pay special attention to the labeling process, using a static definition of who should be considered a churner for the whole dataset. This issue raises some concerns about the reliability of the performances obtained because the definition of churn, and consequently, the labels used for inducing the classifiers, could change over time, but it is assumed that they remain the same.

Focusing on labeling, two works in the games domain studied their behavior using different techniques. Approaching with an economic view, (Clemente-Císcar, San Matías, and Giner-Bosch 2014) utilizes the idea of loyal customers proposed by (Buckinx and Van den Poel 2005) to calculate the usefulness of the CDs by analyzing the economic loss of the churn preventing campaigns. Although it is an exciting approach and directly impacts the game producers' profit, its financial data are necessary, which is generally unavailable, invalidating its usage in most contexts. The work of Rothmeier and colleagues (Rothmeier et al. 2020) provide insights into the different approaches used to label the players. They explain and test four techniques by comparing the final results obtained from various algorithms in the churn prediction task. The four approaches are divided into two categories, the ones that utilize the players' log history, namely, Naive and Sliding Window, and the other that use the idea of disengagement. The latter assumes that a significant reduction in playtime characterizes disengagement, later implicating in churn (Xie et al. 2015). The disengagement-based approaches have an exciting concept, but since they utilize time spent playing to label the players, data usually used as a feature in the churn classification task, we choose to exclude these approaches because they could implicate bias. For instance, the model's rules could disregard other features and consider only the one used to generate the labels.

Even though they tested different churn labeling approaches, the experiments focused on the classifiers' results, not the labeling itself. The problem of evaluating the labeling process resides in the nonexistence of true labels, resulting in the lack of values to be used for comparison. True labels only exist in the cases of games that finished their usage life cycles (Kummer, Nievola, and Paraiso 2017), where the notion of churn has a final form for each player. Note that as each game can have distinct players' behaviors, transferring learning from one game to another without a possible bias is impossible. In sum, true labels cannot be considered since they only exist when the churn prediction is not a need anymore, as the game operation was finished. Meaning that churn labeling approaches must encompass the ability to adapt to the churn volatility during a game life cycle, which is firmly attached to the players' behaviors that change over time (Cook 2007; Zhu, Li, and Zhao 2010). This fact highlights that when a game adopts a static definition of churn, it ignores the changes in players' behavior and keeps predicting churn based on a possible no coherent concept according to current data. An entailed problem is that a good accuracy of a classifier can hide this situation, as what is predicted with high confidence could not be linked to the actual notion of churn, leading to poor churn management.

Identified Labeling Approaches

There are three approaches used in related works to label players as churners or non-churners, each one will be explained, and their weaknesses will be discussed below. Later, an approach that overcomes the identified limitations is presented.

Fixed Value

One of the most common ways to label the players as churners or non-churners is to define a number of days or a Fixed Value (FV). If a player has not played consecutively for this amount of days in the most recent data, named its Last Absence (LA), he/she is considered a churner, as demonstrated by Eq. 1. This value can be defined empirically (Rothenbuehler et al. 2015), but some authors utilized the players' history to achieve an FV that encompasses the behaviors of the player base majority, as performed by (Periáñez et al. 2016), (Runge et al. 2014), and (Yang et al. 2019). Following this idea of a data-driven approach, Eq. 2 represents how the FV calculations were performed in this article, where an Absence With a Return (AWR) is the number of consecutive days not played followed by a day played, the n represents the number of AWRs considering all players and i the ith AWR.

$$Label = \begin{cases} \text{Churner,} & \text{if } LA > FV \\ \text{Non-Churner,} & \text{otherwise} \end{cases} \tag{1}$$

$$FV = \frac{\sum_{i=1}^{n} AWR_i}{n} \tag{2}$$

Naive

The Naive approach consists of dividing the dataset into two roughly equal parts and verifying if the players were present in both parts (non-churner), only on the first (churner) or second (beginner). As stated by (Rothmeier et al. 2020), the addition of the third class provides the intention of improving class balance, but the data splitting technique has several drawbacks. For example, it can be challenging to choose a specific timestamp. The chosen one can bring lots of data from some players but almost none from another. Furthermore, the beginners' exclusion from the non-churners class can conceal particular insights, and the adopted split can lead to important information loss about behavior changes in the second part. A usage example can be seen in (Drachen et al. 2016), where the Naive approach was utilized on the first and second weeks of the players' log history to label the players.

Aside from the drawbacks mentioned earlier, the Naive approach can take, in the worst case, double the size of the dataset to identify churners because it would take the same size of the dataset as the number of days not played. Another problem can be identified if the CD changes because, in the Naive, the rule is always the same. This characteristic restricts a comparison between old and new data and invalidates the identification of a change in the CD. The inability to identify this change excludes the possibility of evaluating the approach's correctness, which is a considerable drawback.

Sliding Window

Like the Naive, the Sliding Window (SW) approach follows the same rules, but the difference occurs when separating the dataset into two parts. The splitting does not need to divide the dataset equally, enabling fine-tuning to specific games and solving the issue of taking too long to identify churners. Since the SW approach follows the same rules of the Naive, accounting for the players' presence in two separate windows, it also suffers from the same problem of not capturing the CD changes.

Individualized Fixed Value

Considering the presented scenarios and the two drawbacks of the commonly used approaches, namely, (1) the inability to identify the changes in the CD and (2) the use of the same definition for all players, the Individualized Fixed Value (IFV) is presented. It follows the idea of defining a value as a threshold, as done by the FV approach but focusing on each player separately. This correlation to the FV enables capturing changes in the CD (solving problem 1). Using an individualized value solves the issue of not fitting every player's behavior (solving problem 2). This labeling approach was applied in the fast-food industry by Bayrak and colleagues (Bayrak et al. 2021) to personalize the churn prevention system according to each customer's behaviors. Given the game domain, to calculate each player's IFV in this work, Eq. 3 was used, where j is the jth player, n_j represents the number of individual AWRs of this player, and i_j is the ith AWR of this player.

Labelling Approach	Labelling Strategy	Individualized Analysis	Allows CD Changes
FV	Absence Average		X
IFV	Absence Average	X	X
Naive, SW	Presence	X	

Table 1: Labelling approaches' characteristics

$$IFV_j = \frac{\sum_{i_j=1}^{n_j} AbsenceWithReturn_{i_J}}{n_j} \tag{3}$$

Given these approaches, there are two strategies to label the players: one considering a fixed value (the average of AWRs) and another that uses the players' presence. Also, these strategies can encompass individual analysis or not, as well as allowing or not identifying changes in the definition of churn. Tab. 1 summarizes such aspects for each approach. Note that the presented IFV is the only one that copes with both individualized analysis and the identification of CD changes.

Game Datasets

The data from two games were used during the experiments: League of Legends (LOL) and World of Warcraft (WOW). LOL is a MOBA game developed by Riot Games[1], which consists of five versus five battles, where each group can be formed by acquaintances or players selected by the matchmaking system of the game. The game's goal is to destroy the enemy's Nexus, a structure located near its base. It is necessary to play cooperatively, conquer objectives, and win recurrent team fights to achieve it. After the end of a match, a player can choose to play again. In this case, a new game will start where all the resources gathered in the previous game are forgotten, meaning that each match is isolated, and the main goal is collective among the players on the same team.

WOW, developed by Blizzard[2] is an MMORPG where

[1] https://www.riotgames.com

[2] https://www.blizzard.com

Game	# Players	# Months	Period
LOL	2,400	23	Oct. 2018 Sep. 2020
WOW	91,064	37	Jan. 2006 Jan. 2009

Table 2: Characteristics of the datasets

the main objective of each player is to get stronger. To conquer higher levels and better equipment, the player can defeat monsters and complete quests or missions with other players. In WOW's gameplay, a player can quit the game anytime he/she wants and return at the same point. These aspects show that the two games diverge in how the games are played, the goal, cooperativeness, and other game design choices. The differences make them suitable for comparing the different churn labeling approaches because they could indicate that the CD changes differently among games.

The datasets containing players' history logs from the game WOW and LOL were used to calculate this change. Tab. 2 presents their characteristics, considering the number of unique players, amount of months, and periods. The only information that the datasets contain are the players' IDs, unique to each player, and series of zeros, ones, and minus ones representing, respectively, days not played, days played, and days antecedent to the first game played. The minus ones are necessary to remove the bias caused by accounts not created before the first date in the dataset, which could be misinterpreted as an absence. The WOW dataset was created by (Lee et al. 2011) and modified to only encompass the previously described information. The LOL dataset was downloaded and organized by the authors utilizing the producers' Application Programming Interface[3] with randomly selected players. Both datasets are available at https://www.ppgia.pucpr.br/~paraiso/Projects/GameAnalytics/DataBases/PlayersLogHistory/.

Proposed Evaluation Method

A way to evaluate the labels could be done by separating the players' log data into two windows (past and current) and deploying a chosen labeling approach in both windows, resulting in two definitions of churn. It is essential to notice that the final result of this process is not the players' labeling but to define, in each window, what churn is (e.g., after n consecutive days a player did not enter the game, he/she is considered a churner). After splitting the data and acquiring a CD for each window, it is possible to apply both definitions in the current window, which contains the most recent data, and compare the obtained labels. Considering the nature of the players' behavior, the labels are believed to change in a given moment because, as observed and described by (Cook 2007; Zhu, Li, and Zhao 2010; Rothenbuehler et al. 2015), the players' life cycle travels a linear path represented by different motivational stages. These behavioral dynamics cause a change in the CD, measured by disagreements between the resultant labels.

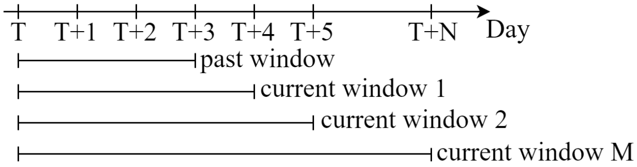

Figure 1: Evaluation method windows comparison

An example of the deployment of the proposed evaluation method is illustrated in Fig. 1. First, we deploy a chosen labeling approach in the past and the current window (i.e., current window 1) to calculate two FVs (individualized per player or not, depending on the labeling approach). Then, using the FVs acquired, we label the players in the current window and compare the obtained labels. By comparing the new and the former definitions of churn in the latest data, it is possible to capture which players had their labels changed, a warning that the labels should be revised. In the comparison, the most recent data labels are considered the true labels, and the F1 score can be utilized to measure the agreement between the past and current CDs, keeping in view the unbalanced nature of the data. Finished the first comparison, the current window is incremented by one day, always containing the elements of the previous window. This process continues until the last day of the dataset.

Since the value representing a change in the definition is the disagreement between the two sets of labels and not the agreement, we utilize equation 4 to achieve the Churn Definition Change Rate (CDCR), ranging from 0 to 1. The higher the value of the CDCR, the more influence a change in the churn definition have in the labeling process, resulting in less reliable classification performance. When this value reaches a certain threshold, user-defined, it is advised to revise the churn definition and retrain the classification model using more accurate labels.

$$ChurnDefinitionChangeRate = 1 – F1Score \tag{4}$$

Experimental Protocol and Results

In the experiments, since the only approaches that can be used to calculate a change in the churn definition are the FV and IFV, the Naive and the SW approaches were excluded. To clarify, in the FV approach, the value utilized in the labeling process was chosen following (Runge et al. 2014), averaging the AWRs of all players in the window used in the evaluation. The IFV follows the same process, but the average is calculated separately for each player. When performing all experiments, eight window sizes were used for the past windows (i.e., 7, 14, 21, 30, 60, 90, 180, and 270 days) to encompass various scenarios regarding the amount of initial players' log data.

Before calculating any CD, an experiment was made to verify the need for an individualized approach, accomplished by calculating the players' AWRs standard deviation. The calculation was performed in both datasets using the same concept of past and current windows of the proposed evaluation method. To better illustrate, Eq. 5 and Eq.

[3] https://developer.riotgames.com/apis

Game	Window Size	FV	SDA
LOL	7	1.61	10.12
LOL	14	2.06	10.04
LOL	21	2.35	10.02
LOL	30	2.56	10.05
LOL	60	3.00	10.19
LOL	90	3.34	10.35
LOL	180	3.91	10.63
LOL	270	4.22	10.46
WOW	7	2.17	105.77
WOW	14	3.35	106.20
WOW	21	4.08	106.74
WOW	30	4.83	107.47
WOW	60	6.79	110.11
WOW	90	7.97	112.98
WOW	180	11.07	122.28
WOW	270	12.85	133.04

Table 3: Standard deviation average

6 demonstrate how the calculations were made for each window size, where SD is the standard deviation, n is the number of IFVs at the current window, IFV_i is the ith IFV, FV is the FV calculated in the past window, SDA is the standard deviation average, m is the number of SDs, and SD_j is the jth SD. The results obtained are presented in Tab. 3.

$$SD = \sqrt{\frac{\sum_{i=1}^{n} (IFV_i - FV)^2}{n-1}} \tag{5}$$

$$SDA = \frac{\sum_{j=1}^{m} SD_j}{m} \tag{6}$$

It is possible to notice that, at the worst case, although the FV (all players AWRs average) is 12.85, indicating that the players typically stay approximately 13 days without playing and then returns, the SDA shows that to encompass 68% of the player base this value should vary between 0 and 146 days. This high range of days indicates that an individual analysis could present benefits compared to the general one because each player is analyzed considering his/her behavior. Since the individualized approach can better identify each player's behavior regarding its AWRs, in the subsequent experiments, instead of comparing FV of the past window against FV of the current window, we will use the IFV in the current window as the true labels.

At this point, all the experiments and solutions are focused on mitigating one of the two identified issues (i.e., limitations on the labeling approaches used in related works). To address the second issue regarding the static definition of churn, we propose two novel labeling approaches based on the FV and IFV approaches. The difference from the original concepts is that instead of calculating the IFVs or FV at the past window and fixing this CD, we propose updating the IFVs or FV when the CDCR reaches a certain threshold, allowing a fine-tuning of the model's retraining frequency. We named these approaches Fixed Value with Redefinition (FVR) and Individualized Fixed Value with Redefinition (IFVR). The value used for the threshold can change depending on the game's characteristics, but for this work, it was set as 0.05 to represent a significant CD change impact. Algorithms 1 and 2 better illustrate, respectively, the FVR and IFVR calculations.

Algorithm 1: FVR Running Example

```
windowSize = 7;
threshold = 0.05;
index = 1;
pastWindowData = data[0:windowSize];
pastFV = average(pastWindowData["AWRs"]);
for player in pastWindowData["players"] do
    if player["lastAbsence"] $\leq$ pastFV then
        player["label"] = "Non-Churner";
    else
        player["label"] = "Churner";
    end
end
while windowSize + index $\leq$ data.length do
    currentWindowData = data[0:windowSize + index];
    currentFV = average(currentWindowData["AWRs"]);
    for player in currentWindowData["players"] do
        if player["lastAbsence"] $\leq$ currentFV then
            player["label"] = "Non-Churner";
        else
            player["label"] = "Churner";
        end
    end
    cdcr = calculateCDCR(pastWindowData, currentWindowData);
    if cdcr $\geq$ threshold then
        pastWindowData = currentWindowData;
        pastFV = currentFV;
        retrainModel();
    end
    index += 1;
end
```

We utilized the Precision metric, ranging from 0 to 1, for the approaches comparison, which is calculated in the same way as the F1 score (i.e., using the past and current concepts). Since it consists of the number of True Positives (TPs) and False Positives (FPs), this metric was chosen to represent a profit increase in churn retaining campaigns. As proposed by (Lee et al. 2018), the profit can be calculated using an equation similar to Eq. 7, where CLV is the expected customer lifetime value, γ is the rate of retained players in the churn retaining campaign, and C is the campaign's cost. The threshold t was excluded from the original formula because we act on the labeling task, which has no influence on the classifiers' threshold that defines the number of TPs and FPs. Analyzing this formula is possible to conclude that disregarding the retention rate and cost of the campaigns, which depend on various factors outside the labeling, the amount of TPs and FPs can be easily observed. Considering

Algorithm 2: IFVR Calculation

```
windowSize = 7;
threshold = 0.05;
index = 1;
pastWindowData = data[0:windowSize];
for player in pastWindowData["players"] do
    player["IFV"] = average(player["AWRs"]);
    if player["lastAbsence"] $\leq$ player["IFV"] then
        player["label"] = "Non-Churner";
    else
        player["label"] = "Churner";
    end
end
while windowSize + index $\leq$ data.length do
    currentWindowData = data[0:windowSize + index];
    for player in currentWindowData["players"] do
        player["IFV"] = average(player["AWRs"]);
        if player["lastAbsence"] $\leq$ player["IFV"] then
            player["label"] = "Non-Churner";
        else
            player["label"] = "Churner";
        end
    end
    cdcr = calculateCDCR(pastWindowData, currentWindowData);
    if cdcr $\geq$ threshold then
        pastWindowData = currentWindowData;
        retrainModel();
    end
    index += 1;
end
```

that the greater the number of TPs and the lower the number of FPs, the higher the Precision is, the same applies to the profit. Therefore, by increasing the Precision, the producer's profit is positively impacted, regardless of CLV, γ, and C.

$$profit = CLV(\gamma TP) - C(TP + FP) \tag{7}$$

Fig. 2 and 3 illustrate the results obtained from comparing all four approaches using the Precision. It can be observed that contrary to what was hypothesized, the IFV approach had, in general, a smaller Precision than the FV, excluding the cases with larger amounts of initial data in the game LOL. As for the two novel approaches that utilize the redefinition concept, they had, in general, better results compared to the ones that maintain a static CD. The only case that this was not true was comparing the FVR and the IFV with a window size of 270 in the LOL data. Even though the IFV does not present benefits when maintaining the CD static, it is possible to see that it obtains the highest Precision among all approaches when the CD is redefined. This result suggests that the IFVR better models the churn behavior. At last, all comparisons presented a significant difference using the Student's t-test with a significance of 0.05.

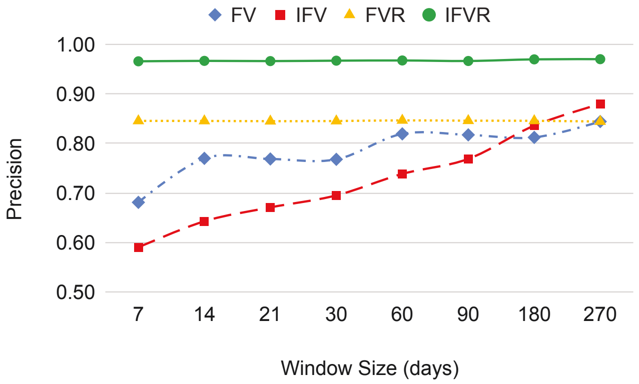

Figure 2: Labeling approaches comparison in the game LOL

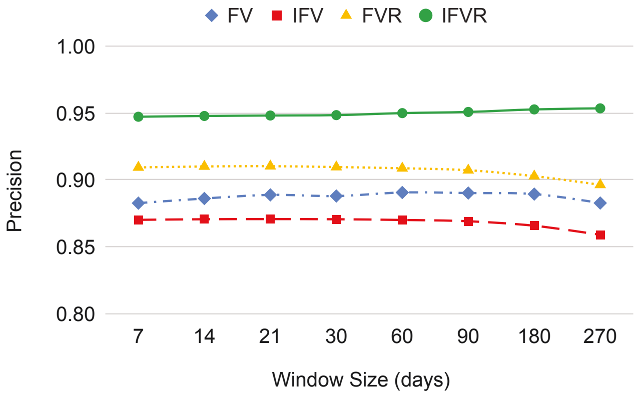

Figure 3: Labeling approaches comparison in the game WOW

Discussion

The experimental results obtained from the datasets comparison show us that the players' behavior contained in the WOW dataset is dissimilar to the LOL because of the differences in the standard deviations and the labeling approaches results. Despite the distinct player base behavior, if we focus on the Precision averages obtained by each approach, independent of the dataset, the IFVR obtained the highest Precision values. It means that it better represents the players, maximizing the amount of TPs and minimizing the FPs, consequently improving the profit in the churn retaining campaigns. Furthermore, contrary to what was supposed, the IFV seems to provide benefits only with the redefinition or in some games, like LOL, when having considerable amounts of frequency information from the players (e.g., at least 180 days for the LOL data). Thus, it is justifiable to advise producers and other authors to utilize the IFVR approach if a redefinition policy is utilized. Otherwise, the FV would be a good option, as commonly approached in the literature. Finally, regardless of the dataset, the IFVR provides more consistent results because it maintains almost the same Precision across all window sizes.

Conclusions

This work focused on the players' labeling process, which is crucial for a reliable churn prediction task. As far as our knowledge goes, this study is the first to analyze this process from a quantitative and profit perspective. We investigated related works to find what could be improved and found two issues, answering RQ1: "What are the recurrent issues present on the churn labeling task?". When defining churn, the problems are related to the limitations on the adopted labeling approaches and the static definition of churn. Focusing on the labeling approaches, two of them use the idea of presence, which can bring many drawbacks, like defining an acceptable window and the window split. Excluding both approaches that use presence, the most common technique uses the idea of calculating a fixed value used to label the players and can be seen as the average behavior of the players, based on their absences. It does not suffer from the same drawbacks when presence is considered, but the idea of using the same value for different players raises doubts about its reliability. Bearing in mind that players' can have different behaviors among themselves, and a unique value could not be sufficient to represent them all, an individualized approach, named IFV, that calculates individual AWR values was deployed. Even though this approach can be seen as simple, adding the redefinition concept allowed the improvement of the traditional churn labeling approaches, optimizing the profit from possible churn retaining campaigns.

Still analyzing the methods used in the related works, it was identified that the churn definition is decided as the first step and never changes, which could be a problem, raising the RQ2: 'How to measure the need to redefine churn?". Wishing to have a way to compare the labels produced by different definitions of churn, we proposed a method capable of quantitatively calculating the influence a change of churn definition has on the resulting labels. Using the proposed evaluation method, it was possible to verify the need to redefine churn, and two new labeling approaches were proposed, the FVR and the IFVR. The redefinition showed great improvements in the Precision for both approaches, surpassing the techniques' Precision without redefinition.

When focusing on the Precision results from the two datasets and their standard deviations, it is possible to conclude that a game can have players with different behaviors, and each approach, parameter, and decision when predicting churn should be specific to the game in question. This work compares two games, LOL and WOW, and highlights some general rules when choosing the best approach. Generally, the IFVR is preferable, but if the redefinition is not implemented, it is advised to utilize the FV instead of the IFV.

Researchers from academia or industry can use this work to improve their churn prediction systems with more reliable labels (by redefining churn and calculating it individually) and can evaluate the classifier's reliability using the proposed evaluation method together with the novel metric. Aside from more reliable labels, the industry can implement the IFVR to optimize the profit in churn retaining campaigns.

Even though the encountered issues were mitigated, this work has some limitations. The most notable regards the threshold used, which was set static but could be fine-tuned to each game, and even an automated process can be proposed in future works to fit any game. The second limitation regards the window sizes utilized. The concept of the IFV and the IFVR is to model each player individually. Still, the size of the windows used to calculate the CD is not customized, opening the possibility of future work encompassing this change. Lastly, more studies can perform the same experiments in other domains to prove or disprove its applicability and importance in different areas that deal with the same churn labeling challenge.

Acknowledgments

We want to thank Siemens Ltd, CAPES-Brazil (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), and Fundação Araucária (CP 09/2016) for their support in this research. We also thank Riot Games for providing access to the game data.

References

Ahn, J.; Hwang, J.; Kim, D.; Choi, H.; and Kang, S. 2020. A Survey on Churn Analysis in Various Business Domains. *IEEE Access* 8: 220816–220839.

Bayrak, A. T.; Aktaş, A. A.; Tunalı, O.; Susuz, O.; and Abbak, N. 2021. Personalized Customer Churn Analysis with Long Short-Term Memory. In *2021 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 79–82. IEEE.

Buckinx, W.; and Van den Poel, D. 2005. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European journal of operational research* 164(1): 252–268.

Clark, O. 2014. *Games as a service: How free to play design can make better games*. London: Focal Press.

Clemente-Císcar, M.; San Matías, S.; and Giner-Bosch, V. 2014. A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings. *European Journal of Operational Research* 239(1): 276–285.

Cook, D. 2007. The Circle of Life: An Analysis of the Game Product Lifecycle. https://www.gamasutra.com/view/feature/129880/the_circle_of_life_an_analysis_of_.php. Accessed: 2021-08-03.

Drachen, A.; Lundquist, E. T.; Kung, Y.; Rao, P.; Sifa, R.; Runge, J.; and Klabjan, D. 2016. Rapid prediction of player retention in free-to-play mobile games. In *Twelfth artificial intelligence and interactive digital entertainment conference*, 23–29. San Francisco: AAAI Press.

El-Nasr, M. S.; Drachen, A.; and Canossa, A. 2016. *Game analytics*. London: Springer.

Hadiji, F.; Sifa, R.; Drachen, A.; Thurau, C.; Kersting, K.; and Bauckhage, C. 2014. Predicting player churn in the wild. In *2014 IEEE Conference on Computational Intelligence and Games*, 1–8. Dortmund: IEEE.

Kilimci, Z. H.; Yörük, H.; and Akyokus, S. 2020. Sentiment Analysis Based Churn Prediction in Mobile Games using Word Embedding Models and Deep Learning Algorithms. In *2020 International Conference on INnovations in Intelligent SysTems and Applications (INISTA)*, 1–7. Novi Sad: IEEE.

Kummer, L. B. M.; Nievola, J. C.; and Paraiso, E. C. 2017. Digital Game Usage Lifecycle: a systematic literature review. In *Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, 1163–1172. Curitiba: SBC.

Kummer, L. B. M.; Nievola, J. C.; and Paraiso, E. C. 2018. Applying Commitment to Churn and Remaining Players Lifetime Prediction. In *Computational Intelligence and Games (CIG), 2018 IEEE Conference on*, 213–220. Maastricht: IEEE.

Lee, E.; Kim, B.; Kang, S.; Kang, B.; Jang, Y.; and Kim, H. K. 2018. Profit optimizing churn prediction for long-term loyal customers in online games. *IEEE Transactions on Games* 12(1): 41–53.

Lee, Y.-T.; Chen, K.-T.; Cheng, Y.-M.; and Lei, C.-L. 2011. World of Warcraft avatar history dataset. In *Proceedings of the second annual ACM conference on Multimedia systems*, 123–128. San Jose: ACM.

Liu, D.-R.; Liao, H.-Y.; Chen, K.-Y.; and Chiu, Y.-L. 2019. Churn prediction and social neighbour influences for different types of user groups in virtual worlds. *Expert Systems* 36(3): e12384.

Milošević, M.; Živić, N.; and Andjelković, I. 2017. Early churn prediction with personalized targeting in mobile social games. *Expert Systems with Applications* 83: 326–332.

Periáñez, Á.; Saas, A.; Guitart, A.; and Magne, C. 2016. Churn prediction in mobile social games: towards a complete assessment using survival ensembles. In *Data Science and Advanced Analytics (DSAA), 2016 IEEE International Conference on*, 564–573. Montreal: IEEE.

Rothenbuehler, P.; Runge, J.; Garcin, F.; and Faltings, B. 2015. Hidden markov models for churn prediction. In *2015 SAI Intelligent Systems Conference (IntelliSys)*, 723–730. London: IEEE.

Rothmeier, K.; Pflanzl, N.; Hüllmann, J. A.; and Preuss, M. 2020. Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game. *IEEE Transactions on Games* 13(1): 78–88.

Runge, J.; Gao, P.; Garcin, F.; and Faltings, B. 2014. Churn prediction for high-value players in casual social games. In *Computational Intelligence and Games (CIG), 2014 IEEE Conference on*, 1–8. Dortmund: IEEE.

Tamaddoni, A.; Stakhovych, S.; and Ewing, M. 2016. Comparing churn prediction techniques and assessing their performance: a contingent perspective. *Journal of service research* 19(2): 123–141.

Xie, H.; Devlin, S.; Kudenko, D.; and Cowling, P. 2015. Predicting player disengagement and first purchase with event-frequency based data representation. In *2015 IEEE Conference on Computational Intelligence and Games (CIG)*, 230–237. Tainan: IEEE.

Yang, W.; Huang, T.; Zeng, J.; Yang, G.; Cai, J.; Chen, L.; Mishra, S.; and Liu, Y. E. 2019. Mining Player In-game Time Spending Regularity for Churn Prediction in Free Online Games. In *2019 IEEE Conference on Games (CoG)*, 1–8. London: IEEE.

Yuan, S.; Bai, S.; Song, M.; and Zhou, Z. 2017. Customer churn prediction in the online new media platform: a case study on juzi entertainment. In *2017 International Conference on Platform Technology and Service (PlatCon)*, 1–5. IEEE.

Zhu, L.; Li, Y.; and Zhao, G. 2010. Exploring the Online-Game Life Cycle Stages. In *E-Business and E-Government (ICEE), 2010 International Conference on*, 2436–2438. Guangzhou: IEEE.