Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling

Guan-Yuan Wang

▶ **To cite this version:**

Guan-Yuan Wang. Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling. Statistika: Statistics and Economy Journal, 2022, 102 (4), pp.443-453. ⟨10.54694/stat.2022.18⟩. ⟨hal-04632443⟩

HAL Id: hal-04632443

https://hal.science/hal-04632443v1

Submitted on 2 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

HAL Authorization

Churn Prediction for High-Value Players in Freemium Mobile Games: Using Random Under-Sampling

Guan-Yuan Wang[1] | *Vilnius University, Vilnius, Lithuania*

Received 7.4.2022, **Accepted** (reviewed) 8.6.2022, **Published** 16.12.2022

Abstract

Many game development companies use game data analysis for mining insights about users' behaviour and possible product growth. One of the most important analysis tasks for game development is user churn prediction. Effective churn prediction can help hold users in the game by initiating additional actions for their engagement. We focused on high-value user churn prediction as it is of particular interest for any business to keep paying customers satisfied and engaged. We consider the churn prediction problem as a classification problem and conduct the random under-sampling approach to address imbalanced class distribution between churners and active users. Based on our real-life data from a freemium casual mobile game, although the best model was chosen as the final classification algorithm for extracted data, we can definitely say there is no general solution to the stated problem. Model performance highly depends on the churn definition, user segmentation and feature engineering, it is therefore necessary to have a custom approach to churn analysis in each specific case.

Keywords	***DOI***	***JEL code***
Churn prediction, mobile games, classification models, resamlpling methods, imbalanced class distribution, machine learning	*https://doi.org/10.54694/stat.2022.18*	*C10, C53, M20, M30*

INTRODUCTION

In the mobile gaming industry, data mining and machine learning methods are widely applied to solve various business problems. The two primary reasons for this are the availability of vast amounts of records on players' in-game behaviour (essentially, every action players make is being recorded) and a limited amount of direct communication with players, which requires understanding their needs by other methods. The majority of mobile games nowadays operate under the so-called "freemium" business model, which implies that players can download the game and play completely free of charge but are offered a range

[1] Faculty of Mathematics and Informatics, Vilnius University, Universiteto g. 3, Vilnius 01513, Lithuania. E-mail: guan-yuan.wang@tprs.stud.vu.lt, phone: (+44)7766387130.

of available in-game purchases that are supposed to enhance gaming experiences. In such a context, player retention is one of the core business problems for game developers, since abandoning the game is costless and effortless for a player, while typically it is more costly for the company to acquire new players than to maintain the current ones. Therefore, one of the popular applications of machine learning in the industry is player churn prediction, which can be generalized as predicting which players, when and with what probability will abandon the game. On a general level, such predictions help a company to develop better business plans and forecasts and grasp a better understanding of its business, while a more specific application would be to directly address the potential churners identified early on and try to retain them in the game.

In this paper, we use a dataset from a casual freemium mobile game to develop a churn probability prediction model for high-value players.

1 LITERATURE REVIEW

1.1 Churn definition

The first issue that is crucial for churn detection in a freemium environment is how churn is actually determined. The churn prediction problem heavily depends on the definition of churn, so different labelling approaches for churn yield different results. Since in the games operating under the freemium business model there is no formal event of discontinuing participation, there is no clear signal that a player has churned. Therefore, a typical approach is to consider a player churned if he does not login into the game for a prespecified period of time. The periods considered for identifying churners range vastly among studies. While some researchers focused on predicting player retention on a specific day (Drachen et al., 2016; Fu et al., 2017), others considered periods of inactivity from 9 days (Bertens et al., 2017) through 14 days (Kristensen and Burelli, 2019) to 13 weeks (Lee et al., 2020) as churn signals. There seems to be a trade-off between the desire to identify churners as early as possible and the risk of misclassification. Hence, several approaches were suggested for empirical identification of the appropriate period taking this trade-off into account.

Bertens et al. (2017) and Guitart et al. (2019) tuned the churn signal period by calculating the percentages of false churners and missed sales (revenue from false churners) for different periods and choosing a period which resulted in less than 5% false churners and 1% of missed sales. A simpler approach was suggested by Runge et al. (2014), who computed a frequency distribution of gaps between logins and chose a period corresponding to the 98% percentile of this distribution. Lee et al. (2020) used the game development cycle (26 weeks) as a starting point to calculate the period for when a user is considered to have churned, arguing that a user who does not respond to the major updates should be considered lost. The final churn period (13 weeks) was based on a compromise between certainty and profit; the compromise was that among the players defined as churners, 25% are not going to churn according to the definition of the game development cycle (26 weeks). Rothmeier et al. (2020) discussed four disparate approaches: naive approach, sliding windows approach, quartile approach and trend over varying dates approach on churn detection issues.

1.2 Segmentation

Another issue highlighted by several researchers is that the player base usually is not homogeneous, so segmentation might be needed to achieve better business outcomes and prediction accuracy. From a business perspective, several researchers highlight the fact that not all customers are equally valuable to the company and hence, it is reasonable to focus the retention management and churn prediction modelling effort specifically on the high-value customer segment (Liu et al., 2019; Runge et al., 2014; Lee et al., 2020, e.g.). Typically, high-value customers are defined as those bringing most of the revenue (Runge et al., 2014); however, an important point was made by Liu et al. (2019) that in the social gaming

context customer's experience with the game and his/her social influence in it might serve as additional measures of customer's worth that should not be neglected. Moreover, segmenting customers based on their in-game behaviour and modelling churn for different groups separately might improve the accuracy of the predictions (Liu et al., 2019; Fu et al., 2017).

1.3 Feature selection

Careful feature variable selection is another way considered by researchers to improve the effect of customer churn prediction. The more powerful feature variables are, the better the customer churn prediction effect is. Ascarza et al. (2018) provided an overview of customer retention management research and identified usage behaviour, user characteristics, satisfaction and social connectivity as the most common predictors in churn modelling. However, since there is often little information available about the players and direct marketing surveys, which could provide data about players' satisfaction are not widespread, in the gaming industry churn modelling mostly has to rely on user behaviour and social connectivity data alone. The user behaviour domain can be considered as consisting of two primary subcategories: in-game activity data (what players do in the game) and engagement data (how frequently and intensively players do). Thus, most of the research we considered used a combination of in-game activity data, engagement data and in-game social interaction data.

In-game activity (also called user performance (Fu et al., 2017) was operationalized with such variables as rounds played, in-game currency balance (Runge et al., 2014), level, number of quests completed, coins collected (Lee et al., 2020), experience points gained, number of quests, number of characters controlled, average character level, number of levels advanced etc. (Borbora and Srivastava, 2012).

User engagement was operationalized as days in the game, time series of logins, last purchase, days since last purchase (Runge et al., 2014), total inter-session length (Borbora and Srivastava, 2012), login frequency, length of login time and average playtime (Fu et al., 2017). In regards to time series data of user activity, Migueis et al. (2012) measured the similarity of the sequence of customers' first purchases to model customer churn, exploring the predictive power of the likelihood of the first product category purchase sequence made by a new customer to identify whether they are churners or not. Although this was done in a retail setting, a similar approach can be introduced in the gaming industry as well to specifically predict the churn of new users.

The last group of features is in-game social interaction features, which include, but are not limited to such variables as invites sent (Runge et al., 2014), friend quitting a game, reduction of the legion (Lee et al., 2020), rate of group interactions, number of churners interacted with (Borbora and Srivastava, 2012), number of in-game friends, whether or not player joined a guild and guild role (Fu et al., 2017). Considering social interaction features, it is worth mentioning that besides directly using players' social interaction data as predictors, it can be used to identify interconnected players and model the cross-effects they might have. For instance, Liu et al. (2019) used players' in-game interaction to identify the most influential social neighbours for each player and showed that introducing the effect of neighbours' features into the player's churn prediction model might improve the model accuracy in some cases.

The last group of features is in-game social interaction features, which include, but are not limited to such variables as invites sent (Runge et al., 2014), friends quitting a game, reduction of the legion (Lee et al., 2020), rate of group interactions, number of churners interacted with (Borbora and Srivastava, 2012), number of in-game friends, whether or not player joined a guild and guild role (Fu et al., 2017). Considering social interaction features, it is worth mentioning that besides directly using players' social interaction data as predictors, it can be used to identify interconnected players and model the cross-effects they might have. For instance, Liu et al. (2019) used players' in-game interaction to identify the most influential social neighbours for each player and showed that introducing the effect of neighbours' features into the player churn prediction model might improve the model accuracy in some cases.

Another group of predictors that emerges from the research is RFM-based (Yeh et al., 2009; Liu et al., 2019; Rahim et al., 2021). This group of predictors is somewhat similar to user engagement as it consists of login and purchase data: recency and frequency of the usage and monetary value spent in the game. Besides just selecting the features that seem most meaningful out of the data available, some researchers engage in designing complex predictors of their own out of the initial data. An example of such an approach is provided by Liu et al. (2019), who designed a complex single measure of what they called "Activity Energy" of the player out of the data available and used the trend in it as the main predictor in their models. Interestingly, Borbora and Srivastava (2012) compared data-driven and theory-driven models with different approaches to variable selection and suggested that even though theory-driven models may be less accurate, they can be better interpretable and, therefore, more preferable over complex data-driven models.

1.4 Churn modelling

Up to this point, we outlined several important issues that might be useful to consider prior to churn modelling identified within the prior research. Next, we briefly consider modelling approaches that are being used for churn prediction. The majority of studies treat churn prediction as a binary classification problem (Hadiji et al., 2014; Runge et al., 2014; Xie et al., 2015; Fathian et al., 2016; Kim et al., 2017; Liu et al., 2019; Lee et al., 2020). In other words, their goal was to predict whether a player will churn at some particular point in time or, equivalently, to identify those players who will churn on a specific day. A wide variety of classification models and algorithms were considered for this purpose, including KNN, ADTreesLogit, Random Forest, Naive Bayes, Neural Networks, Logistic Regression, Decision Trees and Support Vector Machines. Binary classification output makes it easy to compare performances of the models using such metrics as Area under the ROC Curve (AUC) or F-score, so typically several models are compared. However, for business applications, it might be more useful to predict specific churn probabilities in order to distinguish 'on-the-edge' churners who might be responsive to additional marketing efforts from the definite ones (Ascarza et al., 2018).

Despite its popularity, binary classification is not the only way to frame the churn prediction problem. As an alternative, it might be approached with time-series modelling methods (del Río et al., 2021) or survival analysis techniques (Bertens et al., 2017; Guitart et al., 2018). Time-series modelling provides a more general take on the issue compared to the classification and survival analysis approaches. Del Río et al. (2021) used time-series state-space models (ARIMA and Unobserved Components) to model the transition and retention time series for churned, paying, and non-paying user groups. Such an approach allows predicting how many users are expected to churn in a given period of time but does not provide any insights on which users it might be. On the opposite side, survival analysis approach might be seen as the most detailed out of three, since it models a survival curve for each player using such methods as survival ensembles, Cox regression or Kaplan-Meier Model (Guitart et al., 2018). Another aspect worth mentioning is that very few studies considered ensemble learning methods. Fathian et al. (2016) found that classifier ensembles perform better than single classifiers, and boosting methods are superior to bagging, while Guitart et al. (2018) showed that survival ensembles outperform single Cox regression in a survival analysis context.

2 DATA

In this research, we are modelling player churn in a mobile farm game context. The game is available on Android and iOS platforms and has more than 10 000 000+ installs on Android alone. It operates under the freemium business model, which implies that the game is technically free, but users have an opportunity to purchase certain things within the game that are supposed to enhance the gaming experience.

2.1 Sample selection and main definitions

Since we have access to all the user's log data collected in the game, our first goal is to collect a dataset for modelling. The goal of this research is to predict churn of the high-value users, so these concepts have to be operationalized. High-value users were defined as top-paying users, who cumulatively contributed 50% of total revenue in the 90 days from the observation date. Setting the observation date on August 30, 2021, we identified 34 769 high-value users. In order to identify an appropriate period of inactivity to be used as a signal of churn, the distributions of gaps between logins for these high-value users were investigated. Two particular distributions were considered: the distribution of all the gaps between logins for these users and the distribution of maximum gaps between logins. The upper quantiles are presented in Table 1.

Table 1 Distribution of gaps between logins

Quantile	50%	61%	90%	95%	98%	99.5%
All gaps (days)	0	0	1	2	4	4
Max gaps (days)	8	14	71	116	215	468

Source: Own construction

These two approaches demonstrate rather different pictures. Overall, the results suggest that high-value users typically do not take considerable breaks from the game: when they play it, they play it almost every day, as demonstrated by all gaps distribution. On the other hand, some high-value users seem to abandon the game for a long time but come back afterwards. For example, the 90% quantile of max gaps of 71 days suggests that 10% of users under consideration returned to the game after taking a break for more than two months. These results highlight an interesting dilemma for the churn definition in freemium mobile games. On one hand, the interval of inactivity used for churn labelling should be highly unlikely in order to keep the number of misclassified users low. On the other hand, taking a period of months would be rather meaningless from both business and modelling perspectives, since for such a period, it first would be too long to act upon and for the second, we would end up with too few potential churners labelled. Therefore, we chose a 14-day period corresponding to the 99.5% quantile of all gaps and 61% quantile of max gaps as a churn labelling inactivity period. In other words, a user who does not log into the game for 14 consecutive days is considered to be churned. While 39% of users would actually log into the game at least one more time after being labelled churned according to such a definition, the probability of such a gap is only 0.5% as suggested by all gaps distribution. These statements may appear contradictory at first, but this contradiction might be resolved by accepting a less strict churn interpretation and the concept of "returnees" or the users, who churn at one point, but come back into the game later on. Following this definition of churn, we included into the final sample 22 547 high-value users who would be considered active on the observation date and labelled 2 676 (11.87%) of those who churned right afterwards as churned.

2.2 Features extraction

For the selected sample of users, we extract the set of predictors to be used for modelling from the raw players' log data. In order to capture the trend in the user's behaviour, the features are constructed for weekly time intervals 12 weeks before the observation date. In other words, a single feature, for example, the total number of logins is split into 12 predictors: number of logins in the week prior to the observation date, two weeks prior to the observation date, etc. The idea behind such an approach is that the dynamics of a user's past behaviour, rather than the behaviour itself, is expected to hold most of the necessary signals to predict player churn and we would like to capture it in our models. The final set of features is extracted including information about:

- User's logins in the previous 12 weeks: total number of logins, average session duration, total time spent in the game.
- User's payments in the previous 12 weeks: sum of payments, number of payments.
- User's in-game activity in the previous 12 weeks: number of tasks completed, number of achievements received, average time spent on completing a task.
- User's progress in the game in the previous 12 weeks: maximum level achieved, additional levels achieved
- User's profile data: platform (device OS), days since installed, source of installation.

Overall, 141 predictors were extracted.

3 METHODOLOGY

In order to evaluate the performance of the models trained, the dataset is split into training and test sets by allocating 70% of observations to the training set and the rest for testing. Since the prevalence of churners is relatively low, we ensure that they are equally represented in both sets. Training dataset consist of 15 782 observations, including 1 872 (11.86%) churners and 13 910 active users, while the test dataset consist of 6 765 observations: 804 (11.88%) churners and 5 961 active users. It bespeaks that there is an imbalanced class distribution in our dataset. Specifically, the sample size of churners s far smaller than that of non-churners, which can cause high overall classification accuracy but low classification accuracy of churners. In practice, this misclassification of churners usually causes heavier economic losses, especially in our case study of high-value players.

In order to address the imbalanced data issue (11.87% of users are labelled as churners), we employ the Random Under-Sampling (RUS) approach. By undersampling without replacement of the majority class, we end up with a sample of size 3 733 (1 872 churners and 1 861 active users). Undersampling is selected due to higher prediction accuracy and sensitivity compared to oversampling. As a secondary benefit, undersampling of the training set allows us to lower the computational resources needed for training models. Van Hulse et al. (2007), Seiffert et al. (2014), Bauder et al. (2018) and Xiao et al. (2021) show that random undersampling significantly leads to improving the models' performance for datasets with the issue of severe imbalanced class.

3.1 Modelling

3.1.1 Regularized Discriminant Analysis (RDA)

Regularized Discriminant Analysis is a flexible classification technique serving as a bridge between the stricter Linear and Quadratic Discriminant Analyses. However, it imposes certain requirements on the initial data for the model to fit. One such requirement is the invertibility of the covariance matrices for each class of the response variable. In our case, the initial dataset violated this requirement, however, the issue was solved by Principal Component Analysis (PCA) pre-processing of the scaled and centred data. Given that RDA performs the implicit feature selection and PCA is a dimension reduction technique, in this case, the two-step predictor selection was performed. The two RDA parameters: lambda and gamma were tuned over the range between the extreme values of 0 and 1 with step 0.1. The optimal parameters based on the AUC were chosen to be lambda = 1 (essentially assuming the QDA covariance matrices) and gamma = 0.1. Monte-Carlo cross-validation with 25 groups was used to ensure the model's reliability.

3.1.2 Neural Networks (NN)

Neural Networks are powerful and rather universal models; however, they are also computationally demanding to train. In this research, the final NN model was selected in two steps. First, the multilayer structure was investigated by comparing the models with 1 to 3 layers, 1 to 5 nodes in the first layer and 0 to 5 in the 2nd and 3rd and 0 to 2 decay parameters. Since many tuning parameters were considered

in this step, a simple 3-fold cross-validation was used in this step in order to decrease the computational burden and the models were trained on the under-sampled training set. The best model was chosen based on the AUC performance metric and turned out to be a single-layer network with a single node and 0 decay. Therefore, in the second step, just the single-layer models were evaluated over the same ranges of the number of nodes (1 to 5) and decay (0 to 2), but with much more rigorous Monte-Carlo 25 group cross-validation. The final model was chosen based on the AUC and similarly to the previous step included just one node and small decay of 0.1. In both steps, the models were trained on the scaled, centred and spatial sign-transformed data.

3.1.3 K-Nearest Neighbours (KNN)

KNN is one of the simplest classification models, so it is often used as a benchmark for other models to be compared against. The KNN was trained on the centred and scaled under-sampled training set using Monte-Carlo Cross-Validation with 25 groups. The tuning parameter K (number of neighbours) was tuned based on the ROC curve as a performance metric and was set to 43 in the final model.

3.1.4 Logistic regression

Logistic regression is a simple statistical model that uses the logistic function to model a binary dependent variable, although more complex extensions exist. The logarithm of the odds for positive class (churn in this case) is a linear combination of one or more categorical or continuous variables. Here we trained the model on the centred and scaled under-sampled training set.

3.1.5 Random forest

Random forest is a machine learning algorithm that combines multiple decision trees to reach a single result. The basic idea of the algorithm is to gather information from multiple decision trees and select variables and threshold values that maximize classification accuracy. Decision trees, on the other hand, are built by recursively evaluating different features and at each node using the feature that best splits the data. Random forest is a tree-based model and hence does not require feature scaling (contrary to distance-based models). In addition, it automatically detects variables that are important and ignores less important ones; therefore, all predictors were used for modelling. The model was trained on the under-sampled training set. The final model had 500 trees and used 11 predictors for splitting at each node.

3.1.6 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a linear model for classification and regression problems. The idea of SVM is simple: the algorithm creates a line or a hyperplane which separates the data into classes - class boundaries. In this work, we used non-linear SVM. Boundaries in this kind of SVM don't have to be a straight line. It allows capturing more complex relationships between data points, but training time for such a kind of SVM is longer as it's much more computationally intensive. SVM has two main parameters: C (regularization parameter) and Gamma. C parameter controls the tradeoff between smooth decision boundary and classifying training points correctly. Gamma, in its turn, defines how far the influence of a single training example reaches. If it has a low value it means that every point has a far reach and conversely high value of gamma means that every point has a close reach. In the final model, C = 1 and Gamma is inversely proportional to the number of features.

3.1.7 XGBoost (XGB)

XGBoost (Extreme Gradient Boosting) is a decision-tree-based ensemble algorithm that uses gradient boosting. Gradient boosting is a supervised learning algorithm, which attempts to accurately predict a target variable by combining the estimates of a set of simpler, weaker models. The training

of the model proceeds iteratively, adding new trees that predict the residuals or errors of prior trees that are then combined with previous trees to make the final prediction. It uses a gradient descent algorithm to minimize the loss when adding new models. In the model were used 300 trees (n estimators = 300) with maximum depth equals 6 (max depth = 6).

3.1.8 Classifier Ensemble

Besides training single learning algorithms, we considered the Ensemble model which consists of KNN, RDA, NN, Random Forest and logistic regression models. Three types of ensembles are used: majority voting, probability average and weighted probability average. For the weighted probability average, since RDA with the highest sensitivity and random forest with the highest AUC score, churn probabilities for these two models were multiplied by two times compared to the other models, which is by 0.286 and the rest by 0.143.

4 RESULTS

Before analysing the results, let us know how to evaluate and choose our model. In this study, we present five criteria: AUC, accuracy, sensitivity, specificity, and F1-score. Accuracy measures how many observations, both positive and negative, are correctly classified. AUC means the area under the ROC, which is defined as a trade-off between the true positive rate and the false positive rate, characterized as a curve corresponding to each threshold. Sensitivity is the ability of a test to correctly identify an observation with a positive result (in our case, churner). Specificity is the ability of a test to correctly identify people with a negative result (non-churner). F1-score combines sensitivity and specificity into one metric by calculating the harmonic mean between those two criteria. AUC and F1-score both are robust evaluation metrics but researchers are more commonly using AUC as the criterion. With that said, we also presented the F1-score into our list for interested readers. While the AUC is often used as a single comparison measure, it does not account for the cost of any particular type of misclassification. At the same time, in the context of churn modelling, the cost of the missed churner is often much higher than the cost of incorrectly treating an active user as a potential churner. Therefore, the second metric we should pay attention to is sensitivity, or a model's ability to correctly classify the churners. Lastly, we

Table 2 Model evaluation metrics

Model	AUC	Accuracy	Sensitivity	Specificity	F1-score
KNN	0.844	0.6922	0.8756	0.6675	0.4034
RDA	0.895	0.7418	0.8955	0.7210	0.4518
NN	0.906	0.8306	0.8545	0.8274	0.5452
Random forest	0.930	0.8661	0.8632	0.8665	0.6051
Logistic regression	0.882	0.7938	0.8408	0.7874	0.4921
SVM	0.912	0.8017	0.8462	0.4181	0.7315
XGB	0.915	0.9178	0.6405	0.6585	0.6494
Ensemble (majority voting)		0.8085	0.8939	0.7969	0.5259
Ensemble (probability average)	0.917	0.8142	0.8914	0.8038	0.5328
Ensemble (weighted probability average)	0.920	0.8194	0.8989	0.8087	0.5420
Random forest (imbalanced data)	0.929	0.9172	0.6019	0.9597	0.6335

Source: Own construction

should not use accuracy on imbalanced problems. It is easy to obtain a high accuracy score by simply classifying all observations as the majority class, which is shown in our result.

Table 2 presents the comparison of the main classifiers' performance metrics for the trained models. The sensitivity and specificity are computed for the standard threshold of 50% and treating predicted churn as a positive case. In addition, ROC relies on the concept of the adjustable threshold in order to adjust the trade-off between sensitivity and specificity. Majority voting is just used for classification by considering all models which are already constructed. There is no parameter we can adjust; therefore, we cannot calculate the AUC for this method. Considering both metrics, AUC and sensitivity, the top-performing models are Random Forest, Neural Network and the Classifier Ensembles, especially the weighted probability average one. The Classifier Ensembles managed to improve slightly in sensitivity over the Random Forest, but at the cost of specificity. Comparing the classifiers' performance in the context of highly unbalanced classes might not be exactly straightforward. We showed that if we classify the imbalanced data, we will obtain relatively low sensitivity even using the best model, Random Forest. As we discussed earlier, that would bring inevitably massive economic losses, but the random under-sampling approach indeed helps us address the imbalanced class problem.

CONCLUSION

In this paper, we considered a problem of user churn prediction in the mobile freemium game context, provided an example of such predictive modelling and compared the performance of different classification models. While in general, the concept of user churn is relatively straightforward, its operationalization becomes rather challenging and nuanced in a highly unstructured environment of modern mobile gaming. The issue lying at the core of the problem is when the user should be considered churned in the first place. The approach utilized in the current paper is based on the analysis of the distribution of the intervals between logins of the target users provides an objective foundation for churn definition; however, it still does not result in a single correct answer and the modelling results might differ substantially based on the definition chosen. Thus, we conclude that in the practical application a variety of churn definitions should be investigated and the one which produces the best business results should be used.

Another important point that should be highlighted is that it is often reasonable to model churn separately for different segments of users. The first part of this idea is that from the business perspective not all users are equally important, so it is reasonable to focus on the high-value ones and tailor the models towards the specifics of the behaviour of this particular group as it was done in the current research. The second part worth mentioning although it was not directly utilized in the practical part of this research is that further user segmentation based on some user's behavioural profiles (for instance, user's level of social connectedness in the game) might also improve the churn prediction results as it was highlighted in the literature review.

From the modelling side, several conclusions can be drawn from the current research. First of all, we would like to stress that class imbalance is an inherent issue in churn modelling because at any point in time only a small portion of users are expected to churn. At the same time, from the business perspective, the sensitivity of the model or, in other words, its ability to correctly classify soon churners, is more important than general model accuracy, since it is typically more expensive for the company to miss the churner than to spend some extra resources on encouraging an active user. Thus, resampling the training set might be utilized to balance the classes. In the case of current research, undersampling of the majority class (active users) produced better results than oversampling of the underrepresented class of churners, while also lowering the computational burden by decreasing the size of the training set, which might be desirable in some cases.

As for the models' performance, the Random Forest algorithm yielded the best results in the current research, while the simplest KNN expectedly showed the worst performance. The combination of different

models into the classifier ensemble did not yield significant performance improvement over the single Random Forest as was expected. Perhaps, the reason for that was that to achieve performance improvement by combining several different classifiers, they should perform approximately equally in the first place.

Finally, this work certainly has lots of limitations and is not a final say in the churn modelling. It provides an overview of the key issues one should consider for the practical churn modelling in the mobile game environment, but the practical results will highly depend on the specific choices of the churn definition, user segmentation approach, features extracted, algorithms used, etc. Thus, it is a careful and consistent choice and adjustment of all the factors discussed that shall lead to the practically applicable model.

References

ASCARZA, E., NESLIN, S. A., NETZER, O., ANDERSON, Z., FADER, P. S., GUPTA, S., HARDIE, B. G. S., LEMMENS, A., LIBAI, B., NEAL, D., PROVOST, F., SCHRIFT, R. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions [online]. *Customer Needs and Solutions*, 5(1–2): 6–81. <http://doi.org/10.1007/s40547-017-0080-0>.

BAUDER, R. A., KHOSHGOFTAAR, T. M., HASANIN, T. (2018). Data sampling approaches with severely imbalanced big data for medicare fraud detection [online]. *2018 IEEE 30th international conference on tools with artificial intelligence (ICTAI)*, 137–142. <http://doi.org/10.1109/ICTAI.2018.00030>.

BERTENS, P., GUITART, A., PERIANEZ, A. (2017). Games and big data: a scalable multidimensional churn prediction model [online]. *2017 IEEE Conference on Computational Intelligence and Games, CIG*, 33–36. <http://doi.org/10.1109/CIG.2017.8080412>.

BORBORA, Z. H., SRIVASTAVA, J. (2012). User Behavior Modelling Approach for Churn Prediction in Online Games [online]. *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, 51–60. <http://doi.org/10.1109/SocialCom-PASSAT.2012.84>.

DEL RÍO, A. F., GUITART, A., PERIANEZ, A. (2021). A time series approach to player churn and conversion in videogames [online]. *Intelligent Data Analysis*, 25(1): 177–203. <http://doi.org/10.3233/IDA-194940>.

DRACHEN, A., LUNDQUIST, E. T., KUNG, Y., RAO, P., SIFA, R., RUNGE, J., KLABJAN, D. (2016). Rapid prediction of player retention in free-to-play mobile games. *Twelfth artificial intelligence and interactive digital entertainment conference.*

FATHIAN, M., HOSEINPOOR, Y., MINAEI-BIDGOLI, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods [online]. *Kybernetes*, 45(5): 732–743. <http://doi.org/10.1108/K-07-2015-0172>.

FU, X., CHEN, X., SHI, Y. T., BOSE, I., CAI, S. (2017). User segmentation for retention management in online social games [online]. *Decision Support Systems*, 101: 51–68.<http://doi.org/10.1016/j.dss.2017.05.015>.

GUITART, A., CHEN, P., PERIANEZ, A. (2018). The Winning Solution to the IEEE CIG 2017 Game Data Mining Competition [online]. *Machine Learning and Knowledge Extraction*, 1(1): 252–264. <http://doi.org/10.3390/make1010016>.

GUITART, A., DEL RIO, A. F., PERIANEZ, A. (2019). Understanding player engagement and in-game purchasing behavior with ensemble learning [online]. *20th International Conference on Intelligent Games and Simulation, GAME-ON 2019*, 1: 78–85. <https://doi.org/10.48550/arxiv.1907.03947>.

HADIJI, F., SIFA, R., DRACHEN, A., THURAU, C., KERSTING, K., BAUCKHAGE, C. (2014). Predicting player churn in the wild [online]. *2014 IEEE Conference on Computational Intelligence and Games*, 1–8. <https://doi.org/10.1109/CIG.2014.6932876>.

KIM, S., CHOI, D., LEE, E., RHEE, W. (2017). Churn prediction of mobile and online casual games using play log data [online]. *PLoS ONE*, 12(7): 1–20. <http://doi.org/10.1371/journal.pone.0180735>.

KRISTENSEN, J. T., BURELLI, P. (2019). Combining sequential and aggregated data for churn prediction in casual freemium games [online]. *2019 IEEE Conference on Games (CoG)*, 1–8. <http://doi.org/10.1109/CIG.2019.8848106>.

LEE, E., KIM, B., KANG, S., KANG, B., JANG, Y., KIM, H. K. (2020). Profit optimizing churn prediction for long-term loyal customers in online games [online]. *IEEE Transactions on Games*, 12(1): 41–53. <http://doi.org/10.1109/TG.2018.2871215>.

LIU, D. R., LIAO, H. Y., CHEN, K. Y., CHIU, Y. L. (2019). Churn prediction and social neighbour influences for different types of user groups in virtual worlds [online]. *Expert Systems*, 36(3): 1–20. <http://doi.org/10.1111/exsy.12384>.

MIGUEIS, V. L., POEL, D. V., CAMANHO, A. S., CUNHA, J. F. (2012). Predicting partial customer churn using markov for discrimination for modeling first purchase sequences [online]. *Advances in Data Analysis and Classification*, 6(4): 337–353. <https://doi.org/10.1007/s11634-012-0121-3>.

RAHIM, M. A., MUSHAFIQ, M., KHAN, S., ARAIN, Z. A. (2021). RFM-based repurchase behavior for customer classification and segmentation [online]. *Journal of Retailing and Consumer Services*, 61(C). <http://doi.org/10.1016/j.jretconser.2021.102566>.

ROTHMEIER, K., PFLANZL, N., HÜLLMANN, J. A., PREUSS, M. (2020). Prediction of player churn and disengagement based on user activity data of a freemium online strategy game [online]. *IEEE Transactions on Games*, 13(1): 78–88. <http://doi.org/10.1109/TG.2020.2992282>.

RUNGE, J., GAO, P., GARCIN, F., FALTINGS, B. (2014). Churn prediction for high-value players in casual social games [online]. *IEEE Conference on Computatonal Intelligence and Games, CIG.* <http://doi.org/10.1109/CIG.2014.6932875>.

SEIFFERT, C., KHOSHGOFTAAR, T. M., VAN HULSE, J., FOLLECO, A. (2014). An empirical study of the classification performance of learners on imbalanced and noisy software quality data [online]. *Information Sciences*, 259: 571–595. <https://doi.org/10.1016/j.ins.2010.12.016>.

VAN HULSE, J., KHOSHGOFTAAR, T. M., NAPOLITANO, A. (2007). Experimental perspectives on learning from imbalanced data [online]. *Proceedings of the 24th international conference on Machine learning*, 935–942. <https://doi.org/10.1145/1273496.1273614>.

XIAO, J., WANG, Y., CHEN, J., XIE, L., HUANG, J. (2021). Impact of resampling methods and classification models on the imbalanced credit scoring problems [online]. *Information Sciences*, 569: 508–526. <https://doi.org/10.1016/j.ins.2021.05.029>.

XIE, H., DEVLIN, S., KUDENKO, D., COWLING, P. (2015). Predicting player disengagement and first purchase with event-frequency based data representation [online]. *2015 IEEE Conference on Computational Intelligence and Games (CIG)*, 230–237 <http://doi.org/10.1016/10.1109/CIG.2015.7317919>.

YEH, I. C., YANG, K. J., TING, T. M. (2009). Knowledge discovery on RFM model using Bernoulli sequence [online]. *Expert Systems with Applications*, 36(3 PART 2): 5866–5871. <http://doi.org/10.1016/j.eswa.2008.07.018>.