Journal of Advanced Research in Computer Technology and Software Applications
Volume 8, Print Issue 1 - 2024, Pg. No. 1-5
Peer Reviewed Journal

Research Article

Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model

Abhineet Singh

Research Scholar, MCA Thakur Institute of Management Studies, Career Development & Research (TIMSCDR), Mumbai, India.

I N F O

E-mail Id:
abhineetsingh8898@gmail.com
Orcid Id:
https://orcid.org/0009-0006-7717-4350
How to cite this article:
Singh A. Games and Big Data: A Scalable Multi-Dimensional Churn Prediction Model. *J Adv Res Comp Tech Soft Appl* 2024; 8(1): 1-5.

Date of Submission: 2024-03-03
Date of Acceptance: 2024-04-07

A B S T R A C T

The introduction of mobile games has resulted in a paradigm shift in the video game business. Game developers now have a wealth of information on their players at their disposal, allowing them to use trustworthy models that can reliably predict player behavior and scale to massive datasets. Churn prediction, a difficulty shared by many industries, is especially important in the mobile gaming business, where user retention is critical for effective game monetization. We offer a method for predicting game churn based on survival ensembles in this article. Our algorithm accurately predicts both the level at which each player will abandon the game and their total playtime up to that point. It is also resistant to varied data distributions and adaptable to a wide range of response variables, while allowing for effective parallelization of the algorithm. As a result, our methodology is well adapted to doing real-time churner assessments, even for games with millions of daily active users.

Keywords: Social Games, Churn Prediction,. Ensemble Methods,' Survival Analysis, Online Games, User Behavior, Big Data

Introduction

The appearance of mobile games has disrupted the video game business in recent years. Currently, both traditional console games and mobile games are always online, allowing game creators to record each player's action. This one-of-a-kind source of data opens the door to a thorough quantitative study of player behavior and a complete comprehension of player demands.

Preventing user abandonment is a challenge encountered by many sectors, particularly the video game industry. Acquisition initiatives to acquire new players are indeed costly; however, maintaining existing users is more cost-effective. Identifying churners ahead of time allows game owners to conduct targeted promotion campaigns to retain the most valued players and improve monetization more efficiently. Although there have been some works on modeling churn in the field of mobile games,[4,8,9] they generally use techniques that make binary predictions, rely on models that are not easily applicable to different data distributions, or are unable to capture the temporal dynamics inherent in churn. They also have certain downsides in terms of scalability.

In this study, we go beyond the traditional binary strategy to forecast turnover. Previous work[7] demonstrated how to estimate player exit in terms of days, i.e., time-to-event, by including survival analysis into ensemble modeling.

For the first time in the mobile game industry, the current study provides a model that precisely forecasts the level at which a player is predicted to abandon gaming and their hours of playtime until that point. Our methodology enables a comprehensive solution to the churn prediction dilemma from multiple angles and dimensions, assisting in fully understanding and anticipating player attrition.

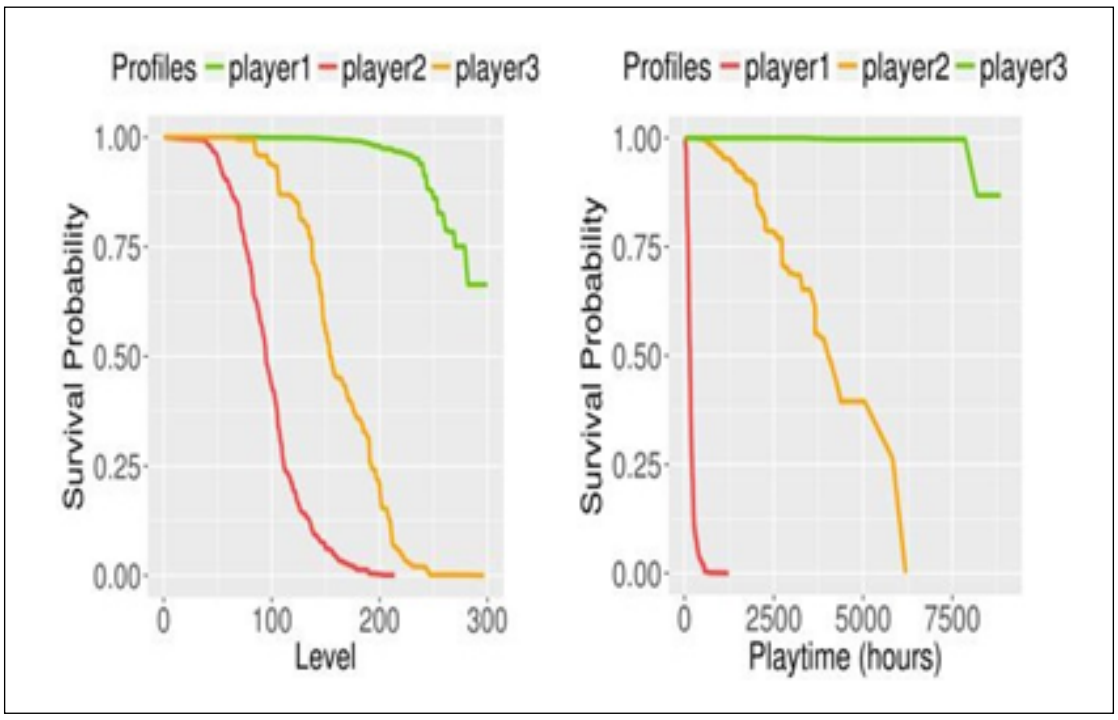

Figure 1.Survival probability predicted for three players based on level and playtime. The first is projected to churn around level 100 and play 500 hours (red), the second around level 200 and will play 5000 hours (yellow), and a loyal player who is not expected to quit (green)

Method

Model

The strategy proposed here is an extension of prior work utilizing conditional inference survival ensembles[5] to forecast churn in mobile social games.[7] According to survival analysis,[2] the model can make reliable predictions even when the response variable is suppressed.

An ensemble of survival trees is what a survival ensemble is. To separate the different survival features of each sample in the tree nodes, each tree computes weighted Kaplan-Meier estimates. Linear rank statistics are employed as the node splitting criterion in order to maximize the survival difference between the daughter nodes. Because each tree's divisions are generated in two phases, the conditional inference survival ensembles[5] are not biased towards predictors with many splits and are more resistant to overfitting. The ideal split variable is chosen first based on the relationship between the covariates and the response, and the optimal split point is determined by comparing two sample linear statistics for each potential partition of the split covariate.

We developed a parallelized version that is more feasible in a production situation and can anticipate other response variables such as level and playtime. The strategy employed here is to parallelize computations on a single machine by employing several cores.

Each core trains a subset of the overall ensemble of trees, and at the end of training, all trees are merged to produce the final model. This method is easily modified to run on several machines, with each machine taking a subset of the entire ensemble and training partial models on a shared disk before combining them back into a single model.

A similar parallelization can be utilized to obtain precise predictions on individual players: each core concentrates on only a subset of players, and full-survival probability curves for each user are computed quickly across many cores.

Dataset

The information came from player action records collected between 2014 and 2017 from Age of Ishtaria, a popular mobile social game developed by Silicon Studio. The forecast was based on a subgroup of the most valuable players, those who generate at least 50% of the revenue (in this example, 6.136 players).

We compute a set of input features that can easily be adapted to different games and accurately capture the dynamics of the data because the model should be adaptable to multiple types of games. The feature calculation may be done in parallel across all participants, and the resulting dataset is small enough to scale to millions of players.

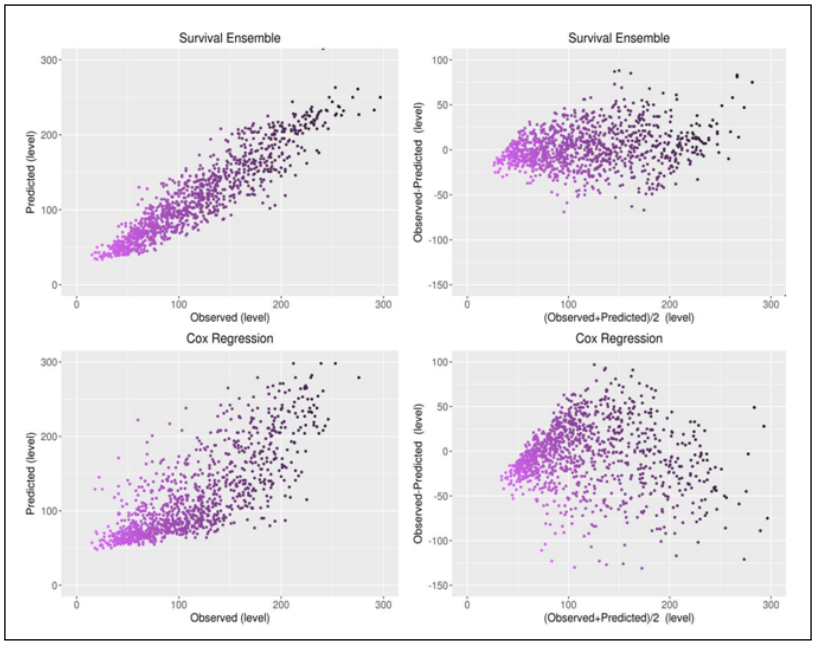

Figure 2.Using the survival ensemble and Coxregression model, predicted median survival levelvs.observed level(left) and relative deviation(right) for churning players.

Daily logins, purchases, gameplay, and level-ups are retrieved from a player's action log in particular. These are ubiquitous in most games and provide important information about player behavior. The mean is calculated for each of these data sources throughout many time periods, notably the player's first nine days, latest nine days, and lifetime. Furthermore, the time between the first and last daily transaction is tallied, as is the amount spent on that day. Finally, the total purchases, total playtime, total logins, and current level are summed to determine the player's current condition. This way of constructing the feature set enables easy extension to include other data sources (e.g., click counts, experience gained, distance traveled, and so on), and it is robust enough to explain the various variability of the data across games.

Outcome

Two further models based on[7] are implemented to forecast how many hours a user will play and at what level they will quit. Each of the following response variables is used to train the models:

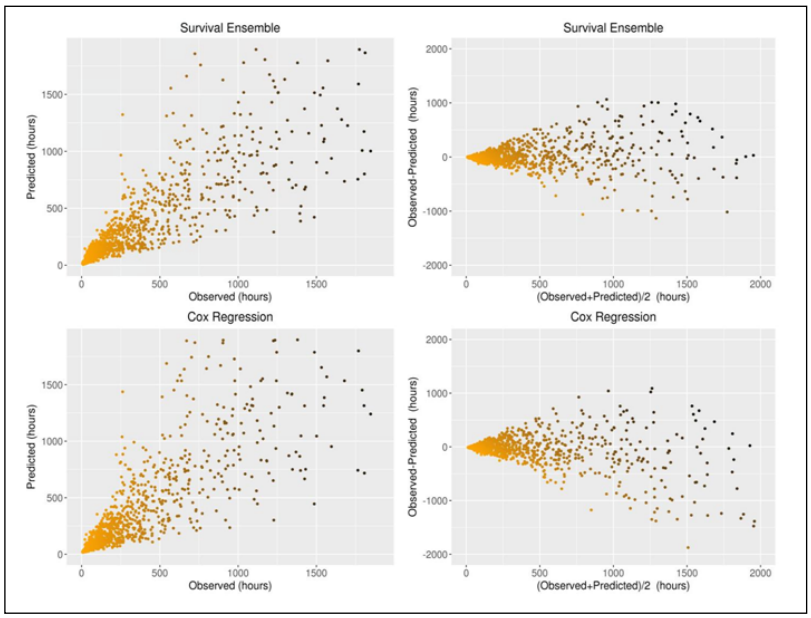

Figure 3. Survival ensemble and Cox regression models predicted median survival playtime vs. observed playtime (left) and relative deviation (right) for churning participants.

- **Playtime:** The number of seconds the user spent playing the game.
- **Level:** The player's most recent gaming level.

The censored variable in both cases is whether or not they churned (churn is defined as not logging in for 9 days).

Features

For the predictors, the most relevant variables from the dataset are chosen for each of the models.

- Playtime model: Level, Days since last purchase, First purchase amount, Last purchase amount, Purchases in the first 9 days, Loyalty index (number of days connected divided by lifetime), Days since last level up.

Level model: Lifetime, days since last purchase, first purchase amount, last purchase amount, purchases in the first 9 days, loyalty index, days since last level up.

Results

The approach presented above generates a unique survival probability curve for each player. Figure 1 depicts the survival probability for three different users, each with a unique survival expectation based on their features. The first participant is likely to churn at a relatively early stage, while the third player will reach a considerably higher level. Similarly, the survival forecast for three players is represented in terms of playtime in the right figure, identifying different degrees of playtime expectancy.

Figures 2 and 3 show the level and playtime forecast findings, respectively. We compare the standard Cox regression model[3] to the survival ensemble model (an ensemble of 1.500 trees). When forecasting at what level a player will abandon the game, the Cox regression model has a harder time capturing the problem's underlying temporal nonlinearity (i.e., the time between levels is not uniform), as seen by a significantly larger spread in Figure 2.

Table I.Integrated Brier Score (Ibs) For Level And Playtime Prediction

Model	IBS level	IBS playtime
Survival Ensemble	0.025	0.026
Cox regression	0.054	0.044
Kaplan-Meier	0.127	0.134

The survival ensemble's accuracy stays higher across the whole level range, with points lying tightly along the identity line, i.e., with minor disparities between predicted and observed values. Higher levels have lower accuracy; however, this is justified by the fact that there is less data and censorship increases (because there are fewer participants at those levels). Figure 3 shows the same result for playtime: forecasts are most accurate for players with very little experience; however, the spread gets more substantial as playtime increases, which can be explained by the fact that relatively few users have played so much. Table I shows the Integrated Brier Scores (IBS)[6] obtained using bootstrap cross-validation with a replacement of 1000 samples.[1] The survival ensemble clearly outperforms the Cox regression model for both level and playtime prediction. Figure 4 also demonstrates that the survival ensemble error is lower than the Cox regression error across the whole range of playtime and levels. Figure 4 illustrates the non-linearity of the time per level dimension (i.e., the time between levels is not evenly distributed), which will vary depending on the game.

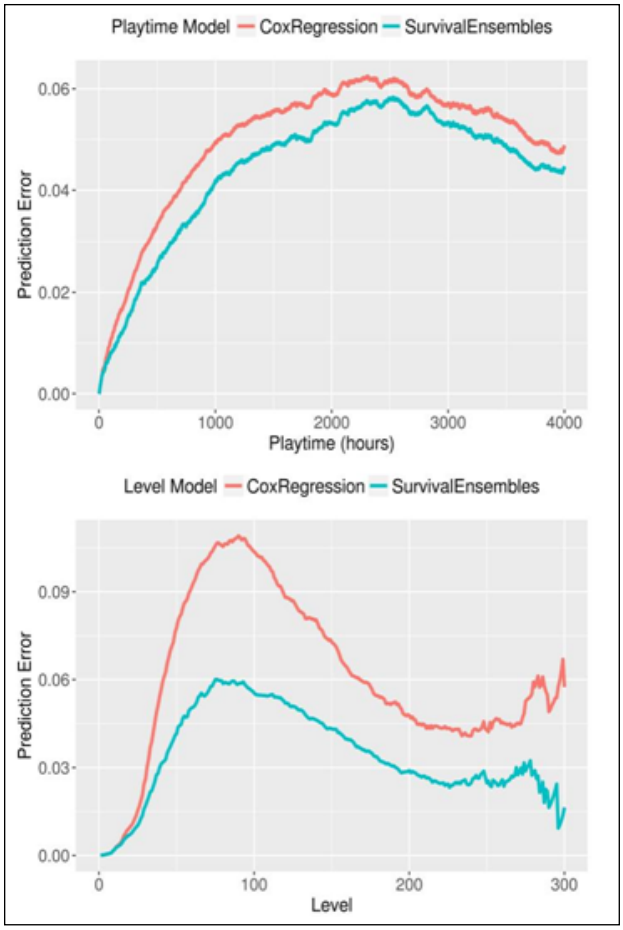

Figure 4. IBS error curves for the play time model (top) and the level model (bottom)

Summary and Conclusions

The results show that the conditional inference survival ensembles method can represent churn in terms of both playtime and level, accurately forecasting which level a player will depart from and how long they will play. This suggests that the model is resistant to various data distributions and suitable for many types of response variables. While Cox regression performed rather well, it involves a significant amount of human labor and has scalability concerns, making it inappropriate for use in a

production context. The suggested survival ensembles, on the other hand, are easily adaptable to other types of games and feature a parallelized approach that can run not only on multiple cores but also on several machines. This allows game producers to efficiently obtain comprehensive survival probability curves for each player and predict in real time not only when a player will leave the game but also at what level and for how many hours they will play before quitting.

References

1. Hicham N, Nassera H, Karim S. Strategic framework for leveraging artificial intelligence in future marketing decision-making. Journal of Intelligent and Management Decision. 2023;2(3):139-50.
2. Martínez-Carrascal JA, Sancho-Vinuesa T. Impact of Assessment Characteristics on Course Withdrawal: a Survival Analysis Approach.
3. Prentice RL, Zhao S. Regression models and multivariate life tables. Journal of the American Statistical Association. 2021 Jul 3;116(535):1330-45.
4. Weiss I, Vilenchik D. Predicting Churn in Online Games by Quantifying Diversity of Engagement. Big Data. 2023 Aug 1;11(4):282-95.
5. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics. 2006 Sep 1;15(3):651-74.
6. Mogensen UB, Ishwaran H, Gerds TA. Evaluating random forests for survival analysis using prediction error curves. Journal of statistical software. 2012 Sep 18;50(11):1.
7. Periáñez Á, Saas A, Guitart A, Magne C. Churn prediction in mobile social games: Towards a complete assessment using survival ensembles. In2016 IEEE international conference on data science and advanced analytics (DSAA) 2016 Oct 17 (pp. 564-573). IEEE.
8. Rothenbuehler P, Runge J, Garcin F, Faltings B. Hidden markov models for churn prediction. In2015 sai intelligent systems conference (intellisys) 2015 Nov 10 (pp. 723-730). IEEE.
9. Runge J, Gao P, Garcin F, Faltings B. Churn prediction for high-value players in casual social games. In2014 IEEE conference on Computational Intelligence and Games 2014 Aug 26 (pp. 1-8). IEEE.