**PROJECT TITLE**

# PREDICTING PLAYER BEHAVIOUR FOR A MOBILE GAME


**99Games Online Private Limited**

Canara Towers, Door Number 6-2-27D, 2nd Floor, Lombard Memorial Hospital Rd, Kadekoppala, Chitpady, Udupi, Karnataka 576101

**SUBMITTED BY**

**Ms. Amirta V**

**M.Sc. (Applied Statistics)**

**PRN: 21060641004**


**ACADEMIC YEAR 2022 - 23**


**Under the guidance of**

**Name of Project Guide / Mentor**

**Mr. Sourabh Jain**

**Designation: Associate Manager - Game Analytics & Performance Marketing**

**Email Id: sourabh@99games.in**

# Contents

## 1. Executive Summary

Gaming industry economically one of the fastest growing sectors and a major contributor to the economy of a country. So, studying the movement of its growth and giving a momentum to the growth is vital. In this study various models have been proposed to capture and predict certain aspects of the gaming. Logistic regression model is proposed for predicting whether a user will pay or not in the game and whether a user will churn or not with the accuracies of **0.9965** and **0.9795** respectively. Multiple linear regression is proposed to model how much money will a user spend in the game.

## 2. Introduction

In today's world gaming is one of the most exciting and engaging form of entertainment for the audience. After facing day to day hustle people throw themselves to this virtual world in order to relieve their stress and escape from reality at least for few hours. Games not only provide entertainment rather it helps people to do social networking. Since the start the industry has enormous growth economically and is expected to boom in the future. It is expected by PWC that the gaming industry around the world would be of $321 Billion by 2026.

There are various types of games under the online games such as Simulation, Action and Adventure, Massively Multiplayer online game (MMO) etc. Among these MMO and simulation is gaining popularity among the audience. MMO is an online video game in which many players come together to play simultaneously in one server. The players interact and compete with each other. It is an open world with different games in it. This type of games is played in network-capable platforms.

These online games are played in various mediums such PC, console and mobile. Among these most of the users use mobile to play these online. Everyone cannot afford PC or console but now a days phone has become a mandatory equipment and by making online games available for mobile medium the reach can be huge. In 2022, smartphone games accounted for 45 percent of video gaming revenue worldwide. The next preference is given to PC followed by console.

Asking 'how do these games makes money?' is a vital question. Generally, there are various strategies followed by the companies like **Digital download** in which the users have to pay money for downloading the game, **Subscription** where the users pay money continuously in the game once they are into it and want to continue it, **in app purchases** where if the user wants to buy any aesthetics or fancy luxuries in the game has to do a purchase using real money in order to acquire it. Plugging in **advertisements** in the game is the main source of income for the game developers. Finding different ways to monetize from game becomes useless if there are no much users. In recent times **Free to Play** strategy is used increase the number of users. According to this strategy the user can download the game for free and play it for free. In this the major monetization technique used is plugging in advertisements.

Prediction on gaming industry very significant as it is one of the fastest growing industries. Especially the online gaming industry. Studying these predictions can lead the respective company to take decision on whether to increase the investment on a game, continue or to reduce. This also helps to know about the overall growth of their company and the areas where they have to improve. Growth of online gaming industry is also beneficial for the country's economy and also helps in boosting the employment rate.

## 3. Objectives

The overall objective of this project is to build models for various prediction in the gaming industry.

The sub objectives as per the various requirements are given below

I. Will pay or not

The aim here was to predict whether the payer who enters the game will spend money in the game or not.

II. Payer's Prediction

Even if the gamer spends money in the game, how much will they spend in the game.

III. Churn Prediction

Predicting whether a player will churn out of the game or not.

## 4. Methodology

The sample player's data provided by the company to achieve the above objectives were live data that were extracted by the company from the game using various software's.

The dataset contained 40660 rows which denoted the different sample players and the 166 columns denoted the attributes corresponding to each player. The variables are the metrics present in the game that are monitored in order to understand the player's movement in the game. The variables of interest were not same for each model, it differed because the influence of variables changed as per the objective function. Different feature selection methods used helped in identifying their potential and using them to model in order to build a better model.

The steps followed in the process are as follows:


**Fig 1: Methodology of the project**

i. Understanding the Data

It is the crucial part of the process. In order to understand data from gaming domain, it is preferred to go through articles, research paper on gaming industry and to get familiarized with the terms present in the game it will be better to play the game. The above mentioned will help in understanding the data which will pave way to the next step.

ii. Cleaning Data

It is one of the most important parts, if this process is not done with at most care the rest of the process will be affected.

There were three steps involved in cleaning the data.

a) Removing the unwanted columns- In this the variables which are irrelevant to the study are removed.
b) Removing the unwanted rows- The users who cannot be included for the further study and based on whom the decision cannot be made are removed from the study.
c) Imputation of missing values- The same imputation techniques were not used for imputing the values. It differed based on variables and models. Below mentioned are the various types of imputation techniques that has been used in the study:

   Stochastic Imputation
   Random Forest Imputation
   KNN Imputation

d) There was no presence of duplicate data.

iii. Feature Selection

Two methods were used selecting the variables needed for a particular model, which are as follows:

a) Correlation: The strength of the relation between the response variable and the predicted variable was used as a main criteria for feature selection. The variables that were strongly correlated were mandatorily included as predictors in the model and the weakly correlated variables were included in order to improve the model.
b) Principal Component Analysis: It is not advisable to always choose variable based on correlation. All the variables could be useful for the study and at the same time there is a necessity to reduce the dimension. In such situation the PCA technique was used which preserved the information from all the variables in terms of linear combination as well as reduced the dimension of the data.

iv. Analysis

One the data is cleaned it is ready to be analyzed. The analysis that was carried out are mentioned below:

a) Descriptive Analysis: In this analysis the visualizations are made using various charts to understand the variables.
b) Diagnostics Analysis: With the interpretations from the visualization, the reasons behind the visualizations are understood.
c) Predictive Analysis: In this part the predictive models were built with respective to the objective.
   In order to build the model, the data was split into train and test data. The train data was used to build thee model were as the test data was used to test the model.
d) Prescriptive Analysis: In this analysis through comparing the various built model we understand the best and recommend it.

## 5. Results

Monetization and retention are vital in a gaming world. Studying and maintaining the both helps any developers to sustain in the market for a long run. Mobile gaming is even more a dynamic arena, here the players have a lot of game choices to switch. In order to sustain here using a free to play game, a deep analysis of the data has to be made with great accuracy.

The proficiency of the sample players can be understood by visualizing the level that are completed played by the users which is shown in figure 2

Table 1 shows that the minimum level completed by the user is zero that is the player has started playing the game and has not completed the level 1 and the highest level completed is 121. From the median value it can be concluded that although the highest level completed is 121, most of the players in the sample data are at beginning level.

**Table 1: Summary of levels completed by the players**

| Parameters and their Values | | | | | |
|---|---|---|---|---|---|
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0 | 1 | 4 | 8 | 10 | 121 |

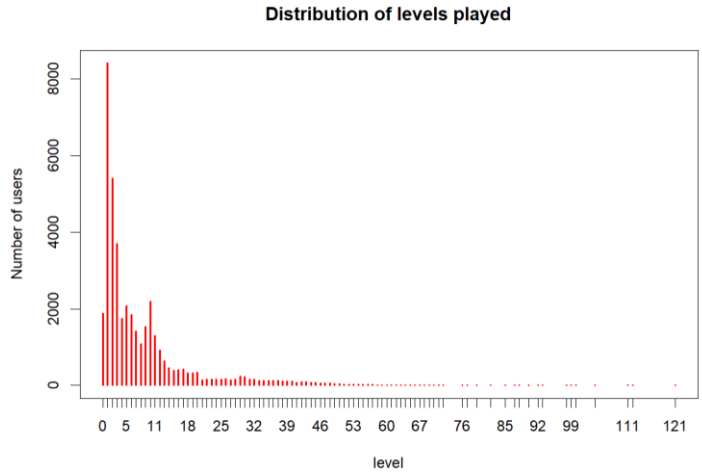


**Fig 2: Distribution of Level played by users**

Figure 1 shows that the number of sample players who playing the level 1 is 8420, that is 20.79% from 40491play level 1. From this it can be said that the most of the sample players are amateurs in the game.

Joining teams will help the sample players to build a social network and motivate them to continue in the game which will be beneficial for the game developer as the retention rate won't increase. Table 2 shows the summary of the teams joined by each user and is visualized in figure 3.

From table 2 it is seen that the minimum team joined by the player is zero that is the sample players are in their initial phase, they haven't explored the game or they do not want to build a social network in the game and retention rate of such players will be low. The highest number of teams

joined by a sample player is 12, which shows that those sample players are very social and retention rate of such sample players will be high. By observing these behaviors, the sample players can be treated accordingly to involve them more into the game.

**Table 2: Summary of Team joined by the players**

| Parameters and their Values | | | | | |
|---|---|---|---|---|---|
| Minimum | 1st Quartile | Median | Mean | 3rd Quartile | Maximum |
| 0 | 0 | 0 | 0.07923 | 0 | 12 |

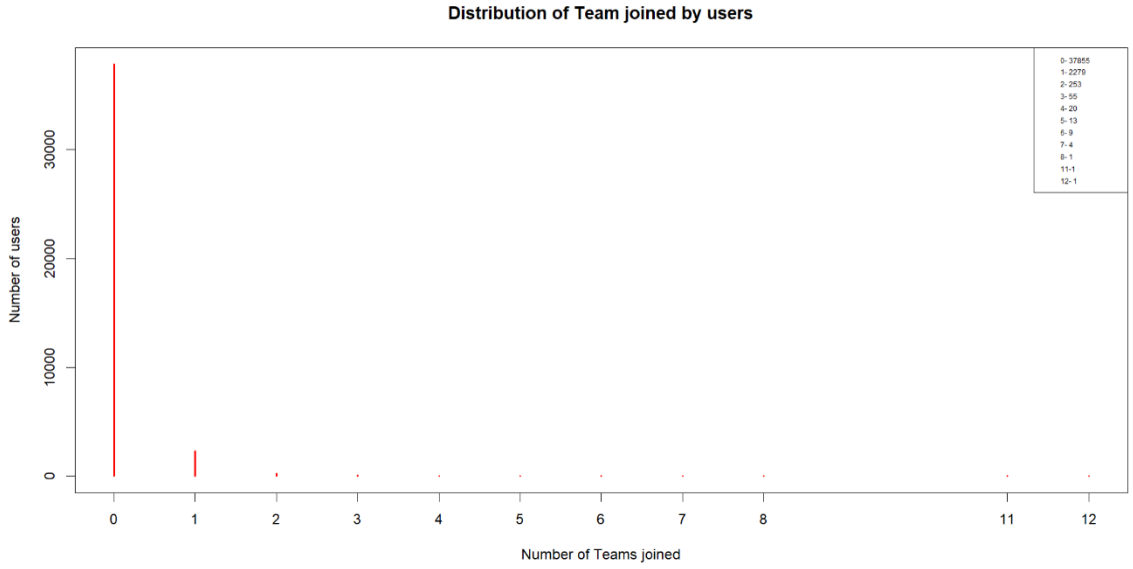


**FIG 3: Distribution of Team Joined by users**

Figure 3 shows that the 93.49% haven't completed even level 1 and 5.62% have completed level 1. The drastic change in percentage can be reasoned with the following 3 major reasons:

- Sample players are new to the game.
- Sample players do not spend enough time in playing the game.
- Sample players do not understand the game.

In order to understand the money spent by the users in the game, a bar graph of user monetization was plotted which is shown in figure 4

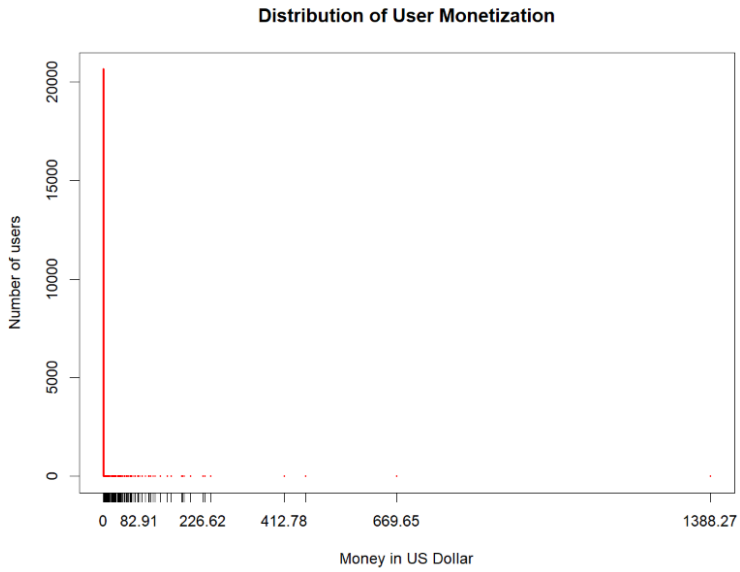


**Fig 4: US dollar spent by users in the game**

From the graph it is observed that more than 20000 sample players don't spend money the game. The dot in 1388.27 shows that only one sample player had spent the mentioned amount in the game. There are decent number of sample players who spend money between 0-200 US dollars.

## Predictive Modelling

The main focus of the project was to build predictive models in order to achieve the objectives mentioned. Three models were built for three different sub objectives.

For predicting 'will the player become payer or not', 'will the player churn from the game or not' the logistic model was built and for 'Predicting how much the player will pay' the multiple regression model was built. Although the same concept was used, still the response variable and the independent variables that were used to build the model were completely different. The accuracy measures that were used to evaluate the models were $R^2$, AIC and Accuracy.

### Model 1: Will the player become payer or not

Here the output was in a binary form that is whether the player will spend money in game or not. As mentioned before the logistic regression was used to predict the outcomes. The confusion matrix for the same is given in table 3.

The accuracy of the model built was **0.9965.**

**Model 2: Will the player churn or not**

Here too the output was in a binary form that is whether the player will churn or not. The same approach was used and logistic regression model was built but the variables that were included in the modelling were different. The accuracy of the model built was **0.9795.**

**Model 3: Predicting how much the player will pay**

The output for this model was a quantitative variable. So, accordingly the multiple regression model was used for the prediction. The first model included 11 variables which were highly correlated with the response variable. After that few more variables were added in order to improve the model.

## 6. Discussions and conclusions

The accuracy of the logistic regression models to predict 'Will the player become payer or not' and 'Will the player churn or not' is given above. When the models were applied to a sample player dataset provided by the company, there was a reduction in accuracy value of the model but the reduction rate was low. The model works good with less loss of accuracy. The fitted logistic model is a good fit for the provided data. The imputation techniques used can also affect the accuracy of the model. For model 3 the Multiple linear regression modeled with variables that had more than 0.1 correlation with the response variable was the best fit.

Gaming industry has abundance of live data. The data gets generated each and every time. The models that have been fitted for the given data has to be revised continuously. The movement of the variables that has been used for modeling should be monitored. Overall, the models that are proposed satisfies the objectives that have been proposed.

## 7. Limitation and Recommendation

As mentioned above the gaming industry deals with live data. On a long run the accuracy of models proposed will reduce. The models have to be revised. For predicting the money spent by the players in the game, regression models other than multiple linear regression can be fitted for the data. The methods like Support Vector Machine can be applied to the data.

## 8. Acknowledgement

I express my sincere gratitude to **99Games** for providing me this chance to explore the gaming industry, my mentor **Mr. Sourabh Jain** for his time and kind support throughout the project. I would also like to express my gratitude to **Ms. Shruthi Rachel D'souza** for guiding me throughout the hiring process.

I express my sincere thanks to my institute **Symbiosis Statistical Institute, Pune** for their encouragement and support to explore summer internship opportunities.

# 9. References


1) Mahlmann, T., Drachen, A., Togelius, J., Canossa, A., & Yannakakis, G. N. (2010, August). Predicting player behavior in tomb raider: Underworld. In *Proceedings of the 2010 IEEE Conference on Computational Intelligence and Games* (pp. 178-185). IEEE.
2) Rodrigues, C., & Ansari, N. (2012). Better Game Design using Association Analysis. *International Journal of Advanced Research in Computer Science*, *3*(3).
3) Fang, Z., Zhou, X., Tang, J., Shao, W., Fong, A. C. M., Sun, L., ... & Luo, J. (2014, November). Modeling paying behavior in game social networks. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (pp. 411-420).
4) Andrat, H., & Ansari, N. (2016, April). Integrating data mining with computer games. In *2016 International Conference on Computing, Communication and Automation (ICCCA)* (pp. 197-201). IEEE.
5) Siqueira, E. S., Castanho, C. D., Rodrigues, G. N., & Jacobi, R. P. (2017, November). A data analysis of player in world of warcraft using game data mining. In *2017 16th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)* (pp. 1-9). IEEE.
6) Boric, S., & Strauss, C. (2022). WHAT TURNS A FREEMIUM GAME PLAYER INTO A PAYING PLAYER. *Journal of Data Intelligence*, *3*(2), 201-217.
7) "Gaming Analytics: How to Leverage Your Customer Data for Sustained Business Growth – Indicative." *Indicative*, 16 Sept. 2021, www.indicative.com/resource/gaming-analytics.
8) Memo, Mobile Dev. "How Much Data Is Needed to Predict LTV? | Mobile Dev Memo by Eric Seufert." *Mobile Dev Memo*, mobiledevmemo.com/much-data-needed-predict-ltv. Accessed 25 Sept. 2022.