A Data Analysis of Player in World of Warcraft using Game Data Mining

Elton S. Siqueira[1]* Genaina N. Rodrigues[1] Carla D. Castanho[1] Ricardo P. Jacobi[1]

[1]Universidade de Brasília, Departamento de Ciência da Computação, Brazil

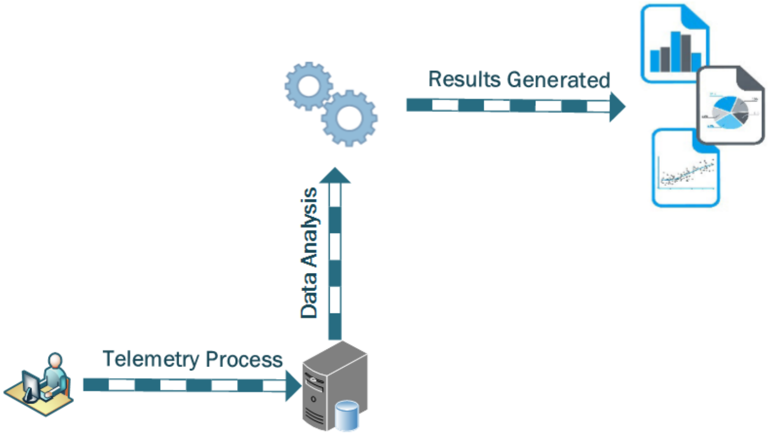

Figure 1: An overview of Game Analytics.

Abstract

Online Games platforms have become a very prospective field of investigation to the game industry. Many business models are being created to understand the dynamics of player behavior to promote improvements in game design and to keep players in the virtual world so that they maintain and renew their respective subscription. From the business perspective it is crucial to predict how many players will join the game and how many will stay in the game, important factors for the company's revenue. This work aims to identify significant data from a popular Massively Multiplayer Online Game (MMOG) - World of Warcraft - using appropriate data mining methods for deduction of players profiles. Multiple linear regression was applied to check whether a player will leave or not the game in the near future. Also, a clustering technique, i.e. K-means, was used to extract clusters of players and identify their common characteristics. The regression model showed that ”Level” and ”Playing Density” contribute significantly to predict whether a player will or not renew the subscription, while the clustering revealed four forms of player profile: Beginner, Intermediate I, Intermediate II and Professional.

Keywords: World of Warcraft, game analytics, MMOG, data Mining, clustering, k-means

1 Introduction

Massively Multiplayer Online Games (MMOGs) are a popular type of entertainment on the Internet. Data released by ESA[13] in 2015 indicate that there are 155 million Americans who play video games and 29% of them currently pay to play (P2P) online. The report released [14] in 2016 also underscores the video game industry's impact on the US economy, the industry contributed $11.7 billion in value to US GDP (Gross Domestic Product). Also, the report revealed that frequent gamers who play multiplayer and online games spend an average of 6.5 hours per week playing online. Currently, the most common business model for online gaming is based on monthly subscription fees that gamers pay to obtain credits, which allow them to start or continue a journey in the game's virtual world. From the perspective of games industry, to understand players behavior in the virtual world and predict ”how long they will stay in the game” are crucial factors to control their revenue [15]. Identifying players profiles and predicting players behavior is based on analyzing their data (for example, Playtime per month). Therefore, data collection (a process known as **Data Telemetry**) and data mining techniques are important for data analysis, as depicted in Figure 1.

There is a wealth of information hidden in process of telemetry data. However, not all of it is readily available, and some are very hard to discover without the proper expert knowledge. In addition, the challenge faced by the game industry to take advantage of telemetry data mirrors the challenge of working with big data. Simply retrieving information from databases is not enough to guide analysts. Instead, new procedures have appeared to help analysts to obtain the information they need to make better decisions. These include: automatic data summarization, the extraction of the essence of the stored information, and the discovery of patterns in raw data [12]. When datasets become very large and complex, many traditional methodologies and algorithms used on smaller datasets fail. Instead, methods designed for large datasets must be used. These methods are called data mining, and they perform a quick and effective analysis, making the results intuitively accessible to non-experts.

The aim of this work is to explore the relationships among the attributes of the World of Warcraft (WoW) players in order to extract profiles and to present a regression model that shows the probability of a player renewing the subscription. Our study contributes to a better understanding of the flow of players in massively multiplayer online games and supports improvements in the business

*e-mail: eltonsarmanho@gmail.com

model through: i) the manipulation of gameplay data and appropriate employment of data mining methods like regression model and data characterization techniques (clustering); ii) the identification of players profiles (user characterization) and their behavior in the virtual world; iii) the extraction of relevant information through data mining, which can be used by the game industry in the creation of a contingency plan to prevent players to give up the game or to motivate them to renew subscriptions.

The rest of this paper is organized as follows. Second section presents fundamental concepts we deal with in this work. In Section 3 we describe some related works. The details about the data considered in this study are given in Section 4. The application of the analysis techniques as well as the main results we obtained are shown in Section 5. At last, in Section 6 we give some final considerations.

2 Fundamental Concepts

2.1 Game Analytics

Game Analytics is the employment of data analysis techniques in games industry [12]. One of its focuses is to assist game development process in all of its phases: conception, design, development, testing and releasing. It is also relevant to other fields of game industry, such as programming, design, user test and business model. The analysis is performed on game metrics obtained from the game-user interactions (see Figure 1).

Each game company has its own method of evaluating game data, and each study may lead to a new discovery about the game and user's behavior. This way, researchers and companies are able to develop systems to support game designers to, eventually, adjust, adapt, balance and improve the game, based on such feedback information.

2.2 Game Telemetry

Game Telemetry is the process of collection and storage of game data used in data analysis [12]. It supports Game Analytics and includes receiving data from online servers, or from a set of sensors installed in a nearby game station, for example. A common scenario sees an installed game client sending data about user-game interaction to a server, where the data are processed and stored in an accessible format, supporting the data analysis process. There are many applications of game telemetry, such as analysis of game servers performance and the collection of user data from Biofeedback sensors.

The game data is employed to assist game user research, which is a field devoted to study user behavior. For instance, it can be used to study methods to analyze player's affective state after certain game phases or to induce specific reactions over them during a game session, with the purpose of supply a great gaming experience [37]. In Health Sciences, game data can be used to aid in the monitoring and treatment of patients, using games as a tool to provide a playful experience to them, using telemetry to collect data about their progression [6]. In Psychology, it is a resource to analyze the user behavior during a user-machine interaction, considering game events as a cognitive stimuli [28].

2.3 Game Data Mining

When using data mining methods in the context of games there is a term called **game data mining** [12] that covers, among others, the following aspects:

- Finding weak spots in *game design*, [22] [16],
- Figuring out how players spend their time when playing [9],
- Exploring how people play a game [1],
- Identifying how much time they spend playing [33],
- Predicting when they will stop playing [2].

In the next subsections, we will show some relevant techniques used in the context of game data mining.

2.3.1 Clustering

In the context of user behavior analysis in computer game development, cluster analysis provides a way to reduce the dimensionality of a dataset in order to find the most important features, and locate patterns which are expressed in terms of user behavior as a function of these features, which can be acted upon to test and refine a game design [12]. Clustering is a highly useful data mining method, containing many algorithms, the most commonly used being k-means [10], Principal Components Analysis (PCA) [19], Non-negative Matrix Factorization (NMF) [26] and Archetype Analysis (AA) [7]. A common objective of unsupervised data analysis using clustering in games is the player categorization (or grouping), ideally resulting in representations of the telemetry data which is interpretable by non-experts (e.g. game designers).

An important aspect in clustering [17] is that the game telemetry data can be stored in a $d \times n$ matrix $\mathbf{V} = [v_1 ... v_n] \in R^{d \times n}$, where each column corresponds to a player and each line to an attribute. Essentially, when dealing in a situation where n samples of d-dimensional vectorial data are accumulate in a data matrix $\mathbf{V}^{d \times n}$, the problem of determining efficacious clusters corresponds to finding a set of $k << n$ centroid vectors $\mathbf{W}^{d \times k}$. If a membership of the data points $\mathbf{V}$ to the centroids in $\mathbf{W}$ is expressed via a coefficient matrix $\mathbf{H}^{k \times n}$, clustering can be cast as a matrix factorization problem; where the goal is to reduce the expected Euclidean norm $\|\mathbf{V} \text{ - } \mathbf{WH}\|$. While methods such as PCA, NMF and k-means all try to reduce the same criterion, they impose distinct constraints and thus yield different matrix factors [19]. For instance, NMF assumes **V**, **W** and **H** to be non-negative matrices and often leads to sparse representation of the data. PCA constrains **W** to be composed of orthonormal vectors and produces a dense **H**, where k-means clustering constrains **H** to unary vectors. The k-means is maybe the most used clustering algorithm, and is theoretically suited for the telemetry procedure in context of games, however, it is focused on retrieving compact cluster regions, and can therefore in practice be hard to interpret.

Archetype Analysis extended to large-scale datasets by via Simplex Volume Maximization (SIVM), applies an alternating least squares procedure where each iteration requires the solution of several constrained quadratic optimization problems [7, 20]. It solves the case where **G** is restricted to convexity instead of to unarity. SIVM appears to be attractive to game telemetry analysis because it allows for the detection of a particular player behavior, as it is focused on finding extremes (e.g. outliers) in the dataset. In principle, what SIVM does is the automatic detection of a combination of characteristic that leads, when being locked in pairs, to a similar but more complex segmentation as K-means, without any user intervention (like in determining the value of k). Where the K-means algorithm produces cluster centroids, SIVM is different as it does not search for commonalities between players, but rather extreme profiles (called **archetypical**) that do not stay in dense cluster regions, but at the edges of the multidimensional space. This excellent feature of SIVM is also its central weakness in the current situation, as it is very sensitive to outliers. If the purpose of the analysis is to find outliers, like detection of bots, cheating or other peculiar player behavior, then it is advisable the use SIVM [30]. This problem does not occur for pure AA, but SIVM is an approximation of AA for large-scale data, which unlike AA fundamentally neglects the distribution of the data, thus adding a weakness to outliers.

Figure 2 illustrates the difference between these clustering methods which were applied to the same dataset. The Figure 2(a) shows the results obtained by k-means, indicating that there are three groups of behaviors in the dataset. The Figures 2(b), 2(c) and 2(d) show the results obtained by Principal Component Analysis (PCA), Non-negative Matrix Factorization (NMF) and Archetypal Analysis (AA), respectively. These three methods supply distinct cluster centroid locations and are less restrictive with respect to cluster membership of individual data samples.

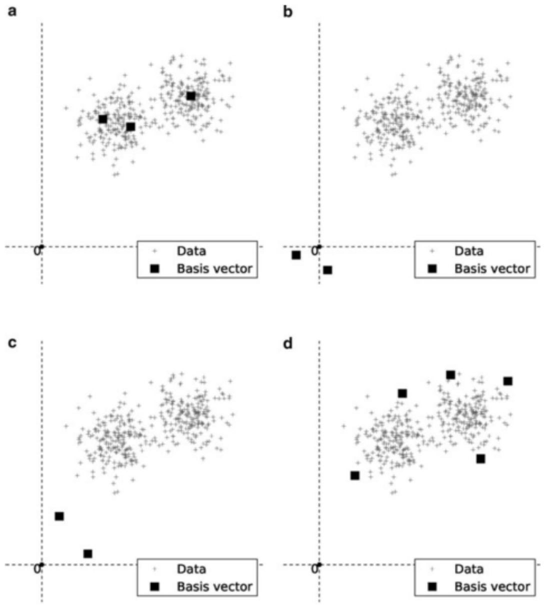

Figure 2: Distinct cluster methods can supply totally different views on the same game data. (a) k-means clustering, (b) PCA, (c) NMF, and (d) Archetypal analysis. [12]

In our study, we used k-means algorithm which is a grouping method based on partitioning and it consists of four steps:

1. Choosing K individuals as initial centers of the groups.
2. Calculating distances between each individual and each center, and assigning the individual to the nearest center (k groups are formed).
3. Replacing k centers by the centroids of k groups identified in Step 2.
4. Stop treatment if the centroids are sufficiently stable or fixed numbers of iterations is reached. Otherwise, repeat the process from Step 2.

Silhouette (S$_i$) measures the similarity of the individual with his own group compared with individuals from other groups. His value varies between [-1,1]. If most individuals have high values of similarity, partitioning result is appropriate. But, if most individuals have low values of similarity, partitioning result is considered inappropriate because it derived many (or little) groups [12].

2.3.2 Regression Models

Among the statistical models used by analysts, regression model is the most common. A regression model allows one to estimate or predict a random variable as a function of several other variables [18]. The estimated variable is called the **response variable**, and the variables used to predict the response are called **predictor variables**. Regression analysis assumes that all predictor variables are quantitative so that arithmetic operations like addition and multiplication are meaningful. Nevertheless, regression techniques can be used to develop a variety of models:

- **Linear regression**: only one predictor variable is allowed.
- **Multiple linear regression**: more than one predictor variable is used.
- **Curvilinear regression**: the relationship between the response and predictors is nonlinear.

A general regression model has a response variable Y (dependent variable) to several predictor variables X_j (independent variables) [18]. It is defined by the following equation (1):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_n + \varepsilon_m \quad \text{For } n \geq 1 \text{ and } m \geq 0 \quad (1)$$

Where

- Y: is the variable response
- β_m: are the parameters
- X_n: is the value of the predictor variable
- ε_m: is a random error term, such that expectation of the error (E) is : $E(\varepsilon_i) = 0$ with $\sigma^2(\varepsilon_i) = \sigma^2$ and $\sigma^2(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i, j \in i \neq j$

Solving regression model consists of determining the estimator b $\begin{pmatrix} b_0 \\ \vdots \\ b_m \end{pmatrix}$ of vector β. Where B_j is a constant that represents the rate of change of Y as a function of changes in X_j.

Some researches [31, 32, 3] use quality indicators of statistical models (including the regression model) to choose the most effective ones for the prediction. The main indicators are:

- **Confidence interval**: provides an estimated range of values which is likely to include an unknown regression parameter (true value), the estimated range being computed from a given set of sample data [11].
- **p-value**: is the probability of finding a test statistic as extreme as the value measured on the sample if the null hypothesis ($H_0 : \beta_1 = \beta_2 = ... = \beta_m = 0$) is true [18]. To reject the null hypothesis and take into consideration the test results significant, $p - value$ must be less than the threshold α ($p - value < \alpha$, where α is normally equal 0.05).
- **F-statistic**: this term provides of the overall significance of the regression model. The null hypothesis is rejected if its p-value $< \alpha$ [18].
- **Coefficient of determination**: is a measure that indicates how well data fit a statistical model. Model fit is better when the coefficient of determination is close to 1, on the other hand, if the regression model is not appropriate then the coefficient of determination is close to zero [18]. The coefficient of determination is denoted by R^2 because it is also square of the sample correlation between the two variables.

3 Related Work

Game analytics is a very active research subject. In [25], the authors showed a general technique to extract and cluster data from a virtual world, but used different attributes from non-game data sets. Games such as World of Warcraft (WoW) provide a rich set of user data, which has been the focus of many studies in game context.

Based on a set of World of Warcraft traces, Pittman et. al [27] proposed a realistic, empirical model to simulate users' gameplay behavior and the fluctuations in game servers' popularity over time.

The authors conjectured that at least four types of information are required to establish a prediction model: i) the rate of server's population change; ii) session length of the players; iii) the spatial distribution of avatars in the virtual world; vi) the movements of avatars over time. They observed that the number of players fluctuated in a diurnal pattern and there may be an increase in the number of players between 4 am and 6 pm. Also, they found that session duration appeared to follow a distribution where approximately 50% of the gamers remain online for 10 minutes or less.

In [4], the authors conducted a user behavior study of Counter-Strike, a popular FPS (First-person Shooter) game. Their work focused on two issues: users' satisfaction with the game, and the predictability of the game server's workload. They analyzed the number of connection attempts and sessions durations, and found that it is extremely difficult to satisfy the users. If a game server is not stable, gamers tend to go elsewhere without considering fidelity. The authors also found that users have short attention spans, and users' sessions duration are usually shorter than one hour. They also analyzed the popularity of game servers and found that the number of users on different servers follows a power-law distribution. Moreover, the server workload exhibits predictable patterns in terms of day and week scale, but the predictability diminishes with larger time scales.

Lewis and Wardrip-Fruin [24] wrote one of the first papers that attempted a large scale survey of WoW using publicly available data. They used a web crawler and screen scraper to collect information on 136,047 characters. Through the analysis of the collected data players were classified based on the items they were holding, the time they took to reach a certain level based on player class, and the number of deaths based on player class. The authors showed that game data that was previously only available to internal developers at the game companies was now available publicly to the world. Moreover, they presented a tool to easily collect the data, allowing researchers to gain insight into these games and leading to interesting qualitative studies.

The researches mentioned so far has relied upon a quantitative, data-driven approach. On the other hand, there have been many papers [8, 36, 34] that have primarily used user studies and online surveys to explore different aspects of WoW such as player motivations, personality, and demographics. These surveys typically involve several thousand respondents, usually found through. In Debeauvais *et al.* [8], the authors used online user surveys to analyze the players' commitment and retention in WoW. A number of 2,865 players completed the survey and authors used the answers to analyze topics such as number of hours played per week, numbers of years the respondents had been playing WoW, and the ratio of respondents who stopped playing the game and returned to it at a later moment. In addition, this data was used to address such game metrics as playing time by character level, in-game demographics (such character races, classes and genders), and character abandonment rate by class.

4 Data Collection

In this section, we introduce the game World of Warcraft (WoW), which is the object of this study, and give some details about the data collection phase.

World of Warcraft is a MMOG (Massively Multiplay Online Game) developed by the Blizzard Entertainment Incorporation. It is a very popular MMOG and according to the annual report of the Entertainment Software Association (ESA) [13] on the Video Game Industry, WoW is among the top ten selling computer games of 2015. Because of its popularity, it has become a field for researchers to study psychology [35], social behavior [5], and gameplay behavior [4, 21].

In this study, we used the data in the repository called **World of Warcraft Avatar History (WoWAH) dataset** [23] for the year of 2008. This year was selected to be analyzed because in that year was released an expansion (called **Wrath of the Lich King**) that had great success. To protect players' privacy, the repository mapped the avatars' names and guild names randomly as positive integers with a consistent mapping (i.e., the same names were always mapped to the same integers). Each element of a sample is a string that contains 7 fields: **Time**, **avatar ID**, **guild**, **level**, **race**, **class** and **zone**. The meanings and valid values of the fields are detailed in Table 1.

Table 1: Field Description

Field	Valid Values
ID	Integer >1
Guild	Integer within [1,513]
Level	Integer within [1,80]
Race	Blood Elf, Orc, Tauren, Troll, Undead
Class	Death, Knight, Druid, Hunter, Mage,
	Paladin, Priest, Rogue, Shaman, Warlock, Warrior
Zone	One of 229 Zones in WoW world
Time	Between Jan.2008 and Dec.2008

We also provide three sample records in Table 2. The first record indicates an avatar with ID 59425 at 00:02:04 on 01/01/08, and the avatar was a level 1, guild 165 Orc Rogue in Orgrimmar.

Table 2: Sample data

Time	ID	Guild	Level	Race	Classe	Zone
01/01/08 00:02:04	59425	165	1	Orc	Rogue	Orgrimmar
01/01/08 00:02:05	65494	-1	9	Orc	Durotar	Durotar
01/02/08 00:03:31	1367	19	60	Undead	Warrior	ArashiMountain

We based our classification method on gamers' Playtime[1] (monthly hours) and Playing Density[2]. First, we randomly chose 800 gamers from the data repository. Second, we performed some data manipulation (data cleaning, data grouping, and averages of grouped data) using R Language[3] to extract game metrics (Playtime and Playing Density).

5 Data Analisys

In this section, we present some basic statistics derived from the collected dataset. Assuming that each avatar is associated to one player, we analyzed the behavior of World of Warcraft players in terms of game metrics - playtime and density - as these are relevant indicators of online gaming [29, 12]. We can examine players' profiles according to the game metrics, thus helping to improve business model of game. In addition, we created a regression model that can predict whether a player will leave this game in the near future.

5.1 Player Categorization

El-Nasr et. al [12] showed that the two game metrics - **average monthly playtime and average playing density** - depict important aspects of player behavior over the time. We performed four experiments using **k-means clustering algorithm**, and met the following criteria in this experiment:

[1] The Playtime is the time that each player spent playing the game

[2] The Playing Density is the occurrence of a gamer's playing days within all available days. For example, if a gamer has logged in the game at least once a day for 15 days out of 30 days, his playing density in that month will be 0.5.

[3] https://www.r-project.org/

- 50 players between levels [0,20];
- 50 players between levels [21,40];
- 50 players between levels [41,60];
- 50 players between levels [61,80];
- The players were randomly selected;
- No repetition of players in the experiments.

The experiment has been executed four times and similar results were obtained concerning to the player behavior (always using k-means clustering algorithm). A high silhouette value when K = 4 indicated better clustering during the experiment, this way was defined that optimal number of clusters is four. We defined four categories of players (Figure 3): **Beginners**, **Intermediate I**, **Intermediate II** and **Professional**. Then, we infer the following information based on the results of the experiments:

- **Professional (Black Circles)**: these are the players that play several days and hours. They are between levels $]70,80]$ of the game. In this group, there is a fast progress in the player's level and a strong engagement of the players in game because they are accustomed to the game environment and always visit it.
- **Intermediate II (Red Circles)**: these are the players that play many days and reasonably many hours. They are between levels $]50,70]$ of the game. In this group, there is a great engagement of players.
- **Intermediate I (Green Circles)**: these are the players that play reasonably many days and few hours. They are between levels $]30,50]$ of the game. In this group, there is a slight progress in the player's level because they play a few hours per day.
- **Beginners (Blue Circles)**: these are the players that play few days and few hours. They are between levels $[0,30]$ of the game. In this group, there is a slow progress in the player's level because they are not familiarized with the game and they rarely visit it.

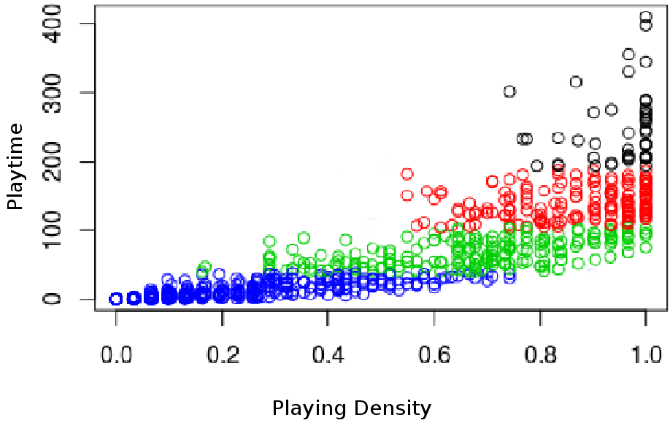

Figure 3: Clustering Overview using K-means

5.2 Monthly Player Behavior

Here we examine how long players play in terms of the overall subscription time (playtime) and monthly gameplay activity (playing density). In addition, in this experiment we had the players' records for the 12 months of the year 2008.

The *cumulative distribution function* (CDF) of average monthly hours playtime and average monthly playing density are shown in Figures 4 and 5, respectively.

In Figure 4, we find out that 50% of the players play longer than 48 hours per month, which fits the following groups of players: Intermediate (I and II) and Professional. Moreover, we find out that only the Professionals and Intermediate II play longer than 84 hours per month (25% of the players). Therefore, it is clear that the players spend a long time immersed in the game.

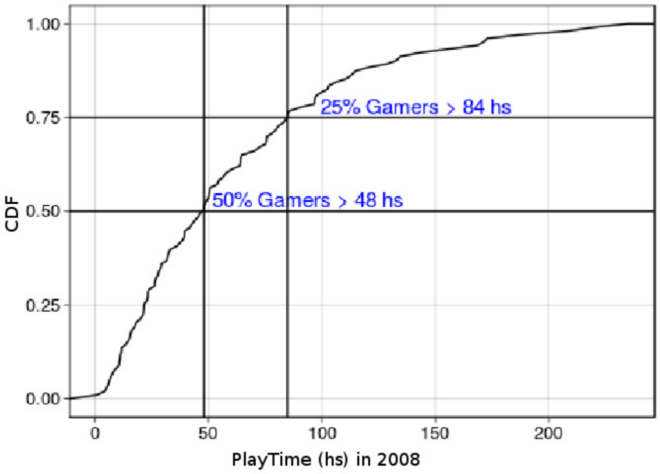

Figure 4: Cumulative distribution function of the Playtime

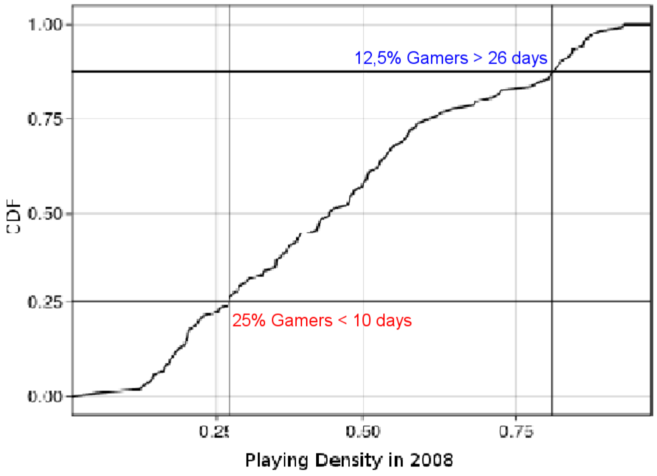

Figure 5: Cumulative distribution function of the Playing Density

Figure 5 shows that 25% of the players play less than 10 days per month, which fits the group of Beginner players, and 12.5% play more than 26 days per month, which indicates the group of Professional players.

Lastly, we summarize the quartiles, averages of the monthly playtime and monthly playing density in Table 3. It shows some important information, such as that about 25% (first quartile, denoted by Q1) of the players playtime in the dataset lies below 22 hours and that about 75% (third quartile, denoted by Q3) lies below 84 hours of playtime. The minimum value (min) of playtime hours is equal to 0.56 and the maximum value (max) is equal to 234.80 hours. Also, we found that about 25% (Q1) of the players played below 28% of all available days and that about 75% (Q3) of them played below 61%. The maximum value is 93%, indicating that some players often interact with this game. Based on these analysis, we have substantial indicators showing that the game is very attractive for its gamers.

Table 3: Summarization of Playing Density and Playtime data

	Min	Q1	Median	Mean	Q3	Max
Playtime	0.56	22.41	48.19	61.28	84.78	234.80
Density	0.057	0.28	0.44	0.47	0.61	0.938

5.3 Analysis of the Results of the Regression Model

In our study on the regression analysis, we selected the player characteristics as the predictor variables (X_i), while **Score** (the player's probability to renew his subscription in the next year) as the response variable (Y) (Table 4).

Table 4: Predictor Variables and Response Variable

	Variables	Domain
Predictor Variables (X_i)	Level of Player Playing density	[1 .. 80] [0.0 .. 1.0]
Response Variable (Y)	Score	[0.0 .. 1.0]

Due the complexity of our dataset, we selected the elements as follows: i) random extraction of 15% of the population (a subpopulation of records); ii) removal of noisy data; iii) the study of the entire subpopulation when it is small. We used "random selection" to prevent bias in the results and to check the stability of the model obtained with several samples. In addition, the **PlayTime** variable is not in the model because it presents a strong correlation coefficient with **Playing Density**, causing a decrease in the coefficient of determination (R^2).

Using the **Playing Density** of the players and their last **level** detected by the system, we came up with Equation 2, where the coefficient of determination equals 90%.

Due to the considerable data range of the **Level** factor, we performed the **logarithmic transformation** on Level (**LevelLog** variable) to normalize the data.

$$\begin{aligned} score = -1.5094 + 0.6623 \times LevelLog \\ -3.9101 \times Density \\ +0.7687 \times LevelLog \times Density \end{aligned} \tag{2}$$

We show the significance of the factors in Equation 2 throught the values presented in Table 5. Since none of the factors include zero, all factors are significant at the 90% confidence level.

Finally, applying the F-Test, we have **F-statistic** value equal to 179.0 and **F-Table** is 2.76 (using significance level $(\alpha) = 0.05\%$). So we have the computed **F-statistic** way greater than the provided **F-table** value. Additionally, since the $p-value < 2.2e^{-16}$ is much lower than the significance level, we can reject the **null hypothesis** that the factors have no effect on the response variable, i.e., the score.

Table 5: Confidence Interval

	X_{min}	X_{max}
Intercept	-2.57	-1.16
LevelLog	0.48	0.84
Density	-5.63	-2.18
LevelLog×Density	0.36	1.17

From our regression model, we can thus provide useful information based on the computed score for a player. For example, this could represent the fact that the higher the score of the player, the higher the probability that the player will renew his subscription. In fact, we have actually confirmed such information in our records as presented in Table 6. There we have the following information that Professional and Intermediate II players, which get significantly higher scores (greater than 80%), renewed their subscriptions, note their high score due to their high level and large playing density. On the other hand, Intermediate I and Beginners players usually don't renew their subscriptions and their scores are less than fifty percent (score < 50%).

Table 6: Results of regression model

Category	Density	Level	LevelLog	Score(%)	Renewed Subscription
Professional	0.89	80	4.38	90%	Yes
Intermediate II	0.72	70	4.24	83%	Yes
Intermediate II	0.50	60	4.09	81%	Yes
Intermediate I	0.40	40	3.68	49%	No
Beginners	0.30	30	3.40	34%	No
Beginners	0.28	28	3.33	31%	No

6 Conclusion

The improvements in playability (or adoption of new strategies in the business model) are important factors to raise the game revenue. However, to accomplish so it is crucial to have knowledge of the players's profile, which can be determined through the analysis of their data recorded during play sessions.

The present study intended to identify important data as well as appropriate game data mining methods to discover players profiles and predict whether a player will renew his subscription. We firstly studied the database of World of Warcraft (WoW) to prepare the data (removing noisy data). Then, we studied the correlation between variables in this database to generate the game metrics that are important for research. In the analysis stage, were applied the following techniques: clustering and CDF (for players characterization) and multiple linear regression (for overall player score).

Our analysis revealed relevant player metrics (**PlayTime**, **Playing Density** and **Level**) and derived some motivating results: i) we could identify four categories of players, namely: Beginner, Intermediate I, Intermediate II and Professional, which interacted with the game according to their experience; ii) Using the cumulative distribution function we showed that the game is highly attractive because several players seem to become dedicated a lot time to virtual world; iii) the regression model showed a measure called **score** that estimated the probability of a player renewing their subscription in the next year. This model aids the game company as it predicts when a player will stop (or continue) playing the game. Also, it may be useful in user-oriented testing, where it is possible to use this information to locate the types of behaviors that lead players to quit playing. Finally, our results may indicate a direct relationship with player retention (the ability of the game to keep people playing it), which is central to revenue.

For future works, we suggest the analysis and exploration of several years' worth of data that can extend our results and enrich information about the players profile. Also, we envision the execution of other experiments using supervised (or unsupervised) classification methods to construct a predictive model based on other variables.

References

[1] C. Alessandro and D. Anders. Patterns of play: Play-personas in user-centred game development. In *DiGRA &#3909 - Proceedings of the 2009 DiGRA International Conference: Breaking New Ground: Innovation in Games, Play, Practice and Theory*. Brunel University, September 2009.

[2] C. Bauckhage, K. Kersting, R. Sifa, C. Thurau, A. Drachen, and A. Canossa. How players lose interest in playing a game: An empirical study based on distributions of total playing times. In *2012 IEEE Conference on Computational Intelligence and Games (CIG)*, pages 139–146, Sept 2012.

[3] S. Benmakrelouf, N. Mezghani, and N. Kara. Towards the Identification of Players' Profiles Using Game's Data Analysis Based on Regression Model and Clustering. *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 - ASONAM '15*, pages 1403–1410, 2015.

[4] C. Chambers, W.-c. Feng, S. Sahu, and D. Saha. Measurement-based characterization of a collection of on-line games. In *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*, IMC '05, pages 1–1, Berkeley, CA, USA, 2005. USENIX Association.

[5] V. H.-h. Chen and H. B.-L. Duh. Understanding social interaction in world of warcraft. In *Proceedings of the International Conference on Advances in Computer Entertainment Technology*, ACE '07, pages 21–24, New York, NY, USA, 2007. ACM.

[6] H. Converse, T. Ferraro, D. Jean, L. Jones, V. Mendhiratta, E. Naviasky, M. Par, T. Rimlinger, S. Southall, J. Sprenkle, and P. Abshire. An emg biofeedback device for video game use in forearm physiotherapy. In *SENSORS, 2013 IEEE*, pages 1–4, Nov 2013.

[7] A. Cutler and L. Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[8] T. Debeauvais, B. Nardi, D. J. Schiano, N. Ducheneaut, and N. Yee. If you build it they might stay: Retention mechanisms in world of warcraft. In *Proceedings of the 6th International Conference on Foundations of Digital Games*, FDG '11, pages 180–187, New York, NY, USA, 2011. ACM.

[9] A. Drachen, R. Sifa, C. Bauckhage, and C. Thurau. Guns, swords and data: Clustering of player behavior in computer games in the wild. In *Computational Intelligence and Games (CIG), 2012 IEEE Conference on*, pages 163–170, Sept 2012.

[10] A. Drachen, C. Thurau, R. Sifa, and C. Bauckhage. A comparison of methods for player clustering via behavioral telemetry. *CoRR*, abs/1407.3950, 2014.

[11] V. J. Easton and J. H. McColl. Statistics glossary v1.1, 1997.

[12] M. S. El-Nasr, A. Drachen, and A. Canossa, editors. *Game Analytics, Maximizing the Value of Player Data*. Springer, 2013.

[13] E. S. A. (ESA). 2015 essential facts about the computer and video game industry. `http://www.theesa.com/wp-content/uploads/2015/04/ESA-Essential-Facts-2015.pdf`, 2016. (Accessed on 06/14/2017).

[14] E. S. A. (ESA). 2016 essential facts about the computer and video game industry. `http://essentialfacts.theesa.com/Essential-Facts-2016.pdf`, 2017. (Accessed on 06/14/2017).

[15] W.-c. Feng, D. Brandt, and D. Saha. A long-term study of a popular mmorpg. In *Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games*, NetGames '07, pages 19–24, New York, NY, USA, 2007. ACM.

[16] A. R. Gagné, M. Seif El-Nasr, and C. D. Shaw. Analysis of telemetry data from a real-time strategy game: A case study. *Comput. Entertain.*, 10(1):2:1–2:25, Dec. 2012.

[17] J. Han, M. Kamber, and J. Pei. Data mining concepts and techniques, third edition, 2012.

[18] R. Jain. *The art of computer systems performance analysis: techniques for experimental design, measurement, simulation, and modeling*. Wiley New York, 1991.

[19] I. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[20] K. Kersting, M. Wahabzada, C. Thurau, and C. Bauckhage. Hierarchical convex nmf for clustering massive data. In M. Sugiyama and Q. Yang, editors, *Proceedings of 2nd Asian Conference on Machine Learning*, volume 13 of *Proceedings of Machine Learning Research*, pages 253–268, Tokyo, Japan, 08–10 Nov 2010. PMLR.

[21] J. Kim, J. Choi, D. Chang, T. Kwon, Y. Choi, and E. Yuk. Traffic characteristics of a massively multi-player online role playing game. In *Proceedings of 4th ACM SIGCOMM Workshop on Network and System Support for Games*, NetGames '05, pages 1–8, New York, NY, USA, 2005. ACM.

[22] D. King and S. Chen. *Metrics for social games*. Presentation at the social games summit 2009, game developers conference, 2009.

[23] Y.-T. Lee, K.-T. Chen, Y.-M. Cheng, and C.-L. Lei. World of warcraft avatar history dataset. In *Proceedings of the Second Annual ACM Conference on Multimedia Systems*, MMSys '11, pages 123–128, New York, NY, USA, 2011. ACM.

[24] C. Lewis and N. Wardrip-Fruin. Mining game statistics from web services: a world of warcraft armory case study. In *FDG*, pages 100–107. ACM, 2010.

[25] G. B. Orgaz, M. D. R-Moreno, D. Camacho, and D. F. Barrero. Clustering avatars behaviours from virtual worlds interactions. In *Proceedings of the 4th International Workshop on Web Intelligence &#38; Communities*, WI&#38;C '12, pages 4:1–4:7, New York, NY, USA, 2012. ACM.

[26] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[27] D. Pittman and C. GauthierDickey. A measurement study of virtual populations in massively multiplayer online games. In *Proceedings of the 6th ACM SIGCOMM Workshop on Network and System Support for Games*, NetGames '07, pages 25–30, New York, NY, USA, 2007. ACM.

[28] A. R. Subahni, L. Xia, and A. S. Malik. Association of mental stress with video games. In *Intelligent and Advanced Systems (ICIAS), 2012 4th International Conference on*, volume 1, pages 82–85, June 2012.

[29] P.-Y. Tarng, K.-T. Chen, and P. Huang. On prophesying online gamer departure. In *Proceedings of the 8th Annual Workshop on Network and Systems Support for Games*, NetGames '09, pages 16:1–16:2, Piscataway, NJ, USA, 2009. IEEE Press.

[30] R. Thawonmas and K. Iizuka. Visualization of online-game players based on their action behaviors. *Int. J. Comput. Games Technol.*, 2008:5:1–5:9, Jan. 2008.

[31] G. Wallner. Sequential Analysis of Player Behavior. *CHI PLAY '15 Proceedings of the 2015 Annual Symposium on Computer-Human Interaction in Play*, pages 349–358, 2015.

[32] G. Wallner and S. Kriglstein. Technical section: Plato: A visual analytics system for gameplay data. *Comput. Graph.*, 38:341–356, Feb. 2014.

[33] D. Williams, M. Consalvo, S. Caplan, and N. Yee. Looking for gender (LFG): Gender roles and behaviors among online gamers. *Journal of Communication*, 59, 2009.

[34] D. Williams, N. Yee, and S. E. Caplan. Who plays, how much, and why? debunking the stereotypical gamer profile. *Journal of Computer-Mediated Communication*, 13(4):993–1018, 2008.

[35] R. Wright. Expert: 40 percent of world of warcraft players addicted. *In tom's GAMES*, 2006.

[36] N. Yee. The demographics, motivations, and derived experiences of users of massively multi-user online graphical environments. *Presence: Teleoper. Virtual Environ.*, 15(3):309–329, June 2006.

[37] A. Çakır İlhan, Y. Ng, C. Khong, and H. Thwaites. The world conference on design, arts and education (dae-2012), may 1-3 2012, antalya, turkey a review of affective design towards video games. *Procedia - Social and Behavioral Sciences*, 51:687 – 691, 2012.