Data Mining League of Legends Logs to Compare New Players to Experienced Ones and to Understand What Causes Game Continuation or End

A study submitted in partial fulfilment of the requirements for the degree of Information School

at

THE UNIVERSITY OF SHEFFIELD

by

David Otto Istvanovszki

September 2016
Abstract

**Background:** Since the Big Data concept reached the computer game industry, telemetry data mining got bigger focus. However, the lack of the former studies and previous models make it hard for the companies to make connection between these datasets and player engagement.

**Aims:** The aim of this research is mining League of Legends logs to understand game skill development over time, to compare new players versus experienced ones and to understand what causes game continuation or end.

**Methods:** The methodology followed by the CRISP-DM processes. Firstly, the data collection and preparation methods will be introduced. The used analyses are focus on the comparing of the new players and experienced ones, and try to find an explanation of the game skill development over time. The data mining model is focus on the measurements of the attributes.

**Results:** The results show that there are no significant differences between the players who already stopped and who’s still playing. However, in between the tier leagues there are huge differences in every game aspect. The data mining model shows a strong correlation between the tier, level and total number of games and the player engagement.

**Conclusions:** The conclusion focus on the already known results, and make a comparison of the literature review and the results. Moreover, it tries to make it clear, that because of the lack of former studies and models, the telemetry data analysis is still very new.
## CONTENTS

1 Introduction ..................................................................................................................4
  1.1 Research aims and objectives ..............................................................................5

2 Literature Review .......................................................................................................7
  2.1 Overview ..................................................................................................................7
  2.2 Psychological effects of the computer games ......................................................7
  2.3 Data analyses in computer games ..........................................................................9
  2.4 Game telemetry data researches ............................................................................9
  2.5 Association mining rules .......................................................................................10
  2.6 League of Legends researches .............................................................................13
  2.7 Measuring player experience ...............................................................................14
  2.8 Data analytics for League of Legends ...................................................................15
  2.9 Data Mining introduction .....................................................................................17
  2.10 Limitations of predictive analytics .......................................................................17
  2.11 Data mining processes and models .....................................................................18
    2.11.1 CRISP-DM Processes .....................................................................................18
    2.11.2 SEMMA ..........................................................................................................20

3 Methodology ..............................................................................................................22
  3.1 Overview ................................................................................................................22
  3.2 Practicalities ...........................................................................................................22
  3.3 Data collection and preparation ............................................................................23
  3.4 Analyses ..................................................................................................................26
  3.5 Model Evaluation and errors .................................................................................27
  3.6 Overfitting .............................................................................................................30
    3.6.1 Pruning .............................................................................................................31
  3.7 Final data mining model .........................................................................................31
  3.8 Ethical aspects .......................................................................................................32

4 Results and Interpretation .........................................................................................33
  4.1 Analyses results .....................................................................................................33
  4.2 Model evaluation ...................................................................................................35
  4.3 Limitations of the analyses ...................................................................................38

5 Conclusion ..................................................................................................................40
  5.1 Contribution to knowledge ...................................................................................40
  5.2 Further research ....................................................................................................41

6 References ..................................................................................................................42

7 Appendices ..................................................................................................................46
Acknowledgements

I would like to thank Dr Gianluca Demartini for his support and guidance as well as all the staff at the Information School for the excellent services and help they provided throughout the course.
1 INTRODUCTION

The history of computer games started with an arcade game called Pong which released by the Atari (Kent, 2002). Since this time the video game industry changed a lot. In the beginning, the computer games were not played on personal computers, more like on consoles. Since the ‘90s when the Windows operating system released, playing with video games started to be part millions of people’s life globally (Kent, 2002).

In this time the most popular games were “single player” or “cooperative” based. These definitions refer to computer games which can be played on one computer (Oosterhuis&Feireiss, 2006). In the middle of 2000 this trend started to be change. The most important reason for that is the Internet penetration. Since this time a new game type started to be more popular called “multiplayer”. Multiplayer means, more than one person can play at the same time, through Internet channels in the same game environment (Kent, 2002). This trend is shown by a statistic, from the 18 most popular computer game 14 is only support online gaming and the other four are essentially “single player” games but they have multiplayer game content as well (Statista, 2016).

Nowadays, the most popular video game in the earth is the League of Legends, around 23% of the players play with it at least once a week (Statista, 2016). The history of League of Legends (LoL) started in 2009. League of Legends is a multiplayer strategy video game developed and published by Riot Games for personal computers. “Lol is a matched based team competition game. Teams are most often composed of five players who are randomly matched together” (Blackburn&Kawk, 2014, p:878). The idea and the gameplay of LoL are not unique at all. There are games which released earlier and based on the same idea, so what makes the League of Legends the most popular computer game in the world? The Riot Games latest statistics show that more than 67 million players play with LoL in every month and more than 27 million plays at least one game in every day (Riot Games, 2014b). It means the users do not just become regular players; there is a huge amount of number who becomes hard-core players. The aim of my dissertation is to find the reason which players become pro and which players will stop playing using data mining and data visualisation techniques.
Using data mining techniques in the computer games have multiple roles. Designers can find out the weak spots in the game design, make a marketing model to convert the non-paying to paying users, discover geographical patterns in the player community, find the most motivating achievements or develop better AI-controlled opponents (Drachen, 2012a). Moreover, specialist can explore how people play the game, how much time they spend playing and predict when they will stop playing or predict what they will do while playing (Drachen, 2012a).

1.1 RESEARCH AIMS AND OBJECTIVES

The aim of my research is mining League of Legends logs to understand game skill development over time, to compare new players versus experienced ones and to understand what causes game continuation or end.

The objectives which are crucial for this part of this research are multiple. The first and most important part is to understand and seek different former researches about League of Legends and other computer games related to this subject. There are two different types of researches in this area. The first analyses are focus on the psychological part of the gaming world. Why gamers become addictive of a game and what are the important elements of the game to players who are playing these games for a long period of time. The second part of my literature review contains former works about data analyses in video games most of them focusing on game design improvements.

The next objective is the data processing. Before data processing, however, data collection about the different usernames in the League of Legends is crucial. There are two different types of methods introduced in this research methodology section. This step does not just include the understanding of the data; it contains data preparing jobs as well, which will make easier the data analyses.

Because I am looking for players who become pro or stopped playing in this research required to specify exactly what the pro player mean is. There are two options for this. League of Legends are ranking players into six different tiers. Each tier is a different group of players with similar skill level. Every tier, except the highest “Challenger tier” is broken into five different divisions. According to the Gameguyz (2016) statistic, more than 71% of the players play in the three lowest tiers. Another option
to collect pro player’s data is not tier, but also the players who made their account before 2013 and since than still play with League of Legends. This indicates that using the LoL built in ranking system and the date of account creation will be able to identify the pro players. On the other hand, another data point are the players who do not become top tier players or stopped playing with League of Legends. It means, the players who were in the three lowest rank (bronze, silver, gold) and their account is inactive in the last six months or longer period of time.
2 LITERATURE REVIEW

2.1 OVERVIEW

In this section different theoretical backgrounds and tools will be introduced which is connected to the computer game user experiences and data mining. This section contains three different types of researches. In this first section the psychological effects of the computer games will be introduced, how the players react and what can be the reason that the player stops playing with the game. The second section focuses on the different game analytics methods from in-game data using with League of Legends. The last part introduces the data mining as a business intelligence method. It contains the limitations and opportunities of the predictive analyses and the different data mining concepts which are available.

2.2 PSYCHOLOGICAL EFFECTS OF THE COMPUTER GAMES

Bartle (2009) introduced in his job four different types of players. Each player types have a different interest in the game. He used four main type of interests which explain the relationship between these players; these are the acting, players, world and interacting (Bartle, 2009). The achievers “like acting on the virtual world” (Bartle, 2009, p: 5). Their aim is to reach the top content in the game, reach the maximum level and get the best gear or skins in the virtual world (Bartle, 2009). The explorers “like interacting with the virtual world” (Bartle, 2009, p: 6). Their aim is to understand the virtual world and try to explore it. The socialisers “like to interact with other players” (Bartle, 2009, pp.6). They look at the game like a social media; they like to talk to other players and helping the other gamers (Bartle, 2009). The killers “like acting on other players” (Bartle, 2009, p:6). The player theory of Bartle (2009) explains why different people play with virtual words and give a basic guideline to the designers which part they should focus on. In my research, it is a crucial part to understand why players start to play with this game and afterward while they stop playing or become professional gamers.

One of the crucial parts of people who become addictive in a computer game are the psychological effects. Chappel et.al (2006) made a research about the Everquest
computer game. They aim was to realise what are the important factors of a computer game where players stop playing with the game or become hard-core players (Chappel et.al, 2006). They used twelve individuals who were in different stages of their Everquest careers (Chappel et.al, 2006). They made three different player stages for the analysis (Chappel et.al, 2006). The first was the players who stopped playing but they start to play again (Chappel et.al, 2006). The second was the players who are playing with Everquest for a long period (Chappel et.al, 2006). The third type was the players whose stopped playing and do not want to play again. “The accounts analysed for this study show that the positive feelings and the fun of playing the game in the early stages of being an Everquest player fade as the game is experienced as making more and more demands on the individual” (Chappel et.al, 2006, p.209). They realised that the hard-core gamers ingame friendships are as meaningful as their real life friends (Chappel et.al, 2006). Their conclusion was that online games are more addictive than single player games because of the competitive and cooperative aspects of the game.

Bowman et.al (2012) research aim was the same like Chappel’s research. However, meanwhile Chappel’s research was more like to realise the anti-social behaviour at the computer game players and what are the important factors for them, Browman’s (2012) research was more like to realise how hard-core players feel about their characters, is there any strong attachment between the characters and players. To get these answers they made an analysis from three perspectives. The first one was the pro-social and anti-social gaming, so each player how much they like to help other players, how much they are social in the virtual world (Bowman et.al, 2012). The second one was the video game skill. Chappel used a seven scale Likert scale to assess the self-reported video game skill. The last one was the analytical strategy. They used this measure “to examine the influence of character attachment and our control variables on pro- and anti-gaming motivations, each motivation factor was regressed on all study variables.” (Bowman et.al, 2012, p.171) One of the main conclusion of their research was “there are strong and interpretable patterns that suggest antisocial gaming to be associated with winning and inciting others (Bowman et.al, 2012, p.172).” “Anti-social gaming motivations attached with skilled, young male gamers and who have a weakened sense of responsibility for their game character’s well-being (Bowman et.al, 2012, p.171). “
2.3 DATA ANALYSES IN COMPUTER GAMES

Previous researches what was introduced in this research are relating to the personality traits of the video game, and they are all based on questionnaires about the game, statistics about the game environment or statistics about the amounts or types of game played. The benefit of the questionnaires’ is that they can provide “verified indicators about stable, real-life personal traits” (Kokkinakis et al., 2014, p.605). However, respondents may respond untruthfully even with the anonymity across the internet and these questionnaires is time consuming, thereby limited the number of participants who can include in each study (Kokkinakis et al., 2014). Kokkinakis et al. (2014) however used an alternative approach to the psychological analysis of gaming data. They used very large datasets for scientifically relevant relationships (Kokkinakis et al., 2014). The aim of their research was to find correlations between the anti-social tendencies and estimated age of the player and in-game interactions. Kokkinakis et al. (2014) research was the first who examined player-player interactions in a MOBA game (Multi-player online battle arena) using gaming data.

These researches are all based on the psychological effects of the computer games. In this research these addiction effects play an important role. One of my research questions is to realise who will be pro gamer and who will stop playing LoL using the pattern of the gaming style. Meanwhile these researches tried to find an answer for this question, however, instead of using the raw data of the game style they research based on the human effects.

2.4 GAME TELEMETRY DATA RESEARCHES

According to Drachen (2012a), without game data mining, using only the telemetry data the results will be limited only to simple aggregates. To know in-depth knowledge about more complex issues, such as player behaviour, not enough to analyse the raw data sets, data mining is required to reach the goal (Drachen, 2012a). The game telemetry is the data logged from the clients or servers, saved the information how each individual plays the game, and how the servers respond for this behaviour (Drachen, 2012a). Moreover, telemetry data are the server logs as well. In total, telemetry is the raw data itself. Another important definition in computer game analyses what Drachen introduced (2012a), is the game metrics. “Game metrics are
interpretable measures, of something, whereas telemetry is the raw data that specialist work with (Drachen, 2012a)."

Bauckhage et al. (2012) made a research about the connection of the game telemetry data and player experiences. In their study they used the total playing times to get the answer how players lose interest in playing a computer game (Bauckhage et al., 2012). They analysed ingame data from five different type of shooter game, two single-player games, Just Cause 2 and Tomb Raider: Underworld; and three multi-player games, Battlefield Bad Company 2, Medal of Honor and Crysis 2 (Bauckhage et al., 2012). In their research they used first passage time distributions and underlying processes (Bauckhage et al., 2012). The academical paper shows that the best results were given by the Brownian motion or Wiener processes. The Wiener process assume that time is $t$, the quantity of interest $I_t$ depends on its history, in a Wiener process, the current value of $I_t$ would result from random Gaussian perturbations $dW_t$ of its previous value. In a Wiener process with drift $\nu$, individual updates are still random, yet, over time, the quantity of interest will tend towards decreasing values” (Bauckhage et al., 2012, p. 142). The results of their research is in the Appendix 1.

The equation of Wiener Process (Brownian motion) with drift: $dI_t = \nu t + \beta dW_t$

Bauckhage et al. discussion based on the fact that there are no significant differences in the total playing hours in the five games and for each game there is a significant number of players who played the game for only a couple of hours (Bauckhage et al., 2012). In their research they Weibull distributions which „provide convincing abstractions of empirically determined distributions of total playing times per player” (Bauckhage et al., 2012, p. 144).

### 2.5 ASSOCIATION MINING RULES

Frequent item set mining rules are interesting area of the data mining, which focuses on “looking at sequences of events or actions” (Drachen, 2012b). Sequence analyses used in different areas, however, it has a crucial role in game data analytics as well, which helps to build patterns to understand particular player’s behaviour, such as a player quitting the game (Drachen, 2012b).
The original theory at data mining frequent itemsets introduced by Agrawal and Srikant in 1994. The discovery of large itemsets follows algorithms making several passes over the data (Agrawal & Srikant, 1994). Firstly, the support of individual items is counted and then the large ones are specified (Agrawal & Srikant, 1994). The following subsequent passes begin with a seed set of the itemsets which were determined as large in the prior pass (Agrawal & Srikant, 1994). These seed sets used in the passes are supposed to discover candidate itemsets – new potentially large itemsets – for which the actual support is counted during the pass over the data (Agrawal & Srikant, 1994). The pass ending is defined by the determination of the truly large candidate itemsets (Agrawal & Srikant, 1994). These itemsets again represent the seed set for the following pass (Agrawal & Srikant, 1994). The procedure ends when no new large itemsets are discovered anymore (Agrawal & Srikant, 1994).

The frequent itemset generation approach based on the “finding of frequent itemsets is to determine the support count for every candidate itemset in the lattice structure (Tan & Steinbach & Kumar, 2006, p.332) “. “To do this we need to compare each candidate against every transaction (Tan & Steinbach & Kumar, 2006, p.332) “. “If the candidate is contained in the transaction its support count will be incremented (Tan & Steinbach & Kumar, 2006, p.332)”.

There are two ways to reduce the complexity of frequent itemset generation:

1) Reduce the number of candidate itemsets. The best solution is the Apriori Principle. The theory of the Apriori Principle is, “if an itemset is frequent, then all of its subsets must also be frequent” (Tan & Steinbach & Kumar, 2006, p.333).

The first part of the algorithm is to count the incidents of the large itemsets (Agrawal & Srikant, 1994). The next part consists two different phrases, where the k is a subsequent pass (Agrawal & Srikant, 1994). First, the large itemset, describes as \(L_{k-1}\), which are found in the \(K-1\)th pass, used to generate the appropriate candidate itemset, which is described as \(C_k\) (Agrawal & Srikant, 1994). Secondly, these \(C_k\) candidates are scanned in the database and counted (Agrawal & Srikant, 1994). The last part of the algorithm contains buffer management techniques which helps to efficiently determines the candidates in \(C_k\) (Agrawal & Srikant, 1994). The table below shows the notations of the Apriori algorithm and the Figure 1. shows the algorithm itself.
### Notation for Apriori Algorithm

<table>
<thead>
<tr>
<th>k - itemset</th>
<th>An itemset having k items</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L_k$</td>
<td>Set of large k-itemsets</td>
</tr>
<tr>
<td></td>
<td>Each member of this set has two fields:</td>
</tr>
<tr>
<td></td>
<td>1) Itemset</td>
</tr>
<tr>
<td></td>
<td>2) Support count (Agrawal &amp; Srikant, 1994)</td>
</tr>
<tr>
<td>$C_k$</td>
<td>Set of candidate k-itemsets, so the potentially large itemsets. Each member of this set has two fields:</td>
</tr>
<tr>
<td></td>
<td>1) Itemset</td>
</tr>
<tr>
<td></td>
<td>2) Support count (Agrawal &amp; Srikant, 1994)</td>
</tr>
<tr>
<td>$\tilde{C}_k$</td>
<td>Set of candidate k-itemsets when transactions are kept associated with the candidates (Agrawal &amp; Srikant, 1994).</td>
</tr>
</tbody>
</table>
1) \( L_1 = \{ \text{large 1-itemsets} \}; \)
2) \( \text{for } (k = 2; \ L_{k-1} \neq \emptyset; \ k++) \ \text{do begin} \)
3) \( C_k = \text{apriori-gen}(L_{k-1}); \ // \ \text{New candidates} \)
4) \( \text{forall transactions } t \in D \ \text{do begin} \)
5) \( C_t = \text{subset}(C_k, t); \ // \ \text{Candidates contained in } t \)
6) \( \text{forall candidates } c \in C_t \ \text{do} \)
7) \( c.\text{count}++; \)
8) \( \text{end} \)
9) \( L_k = \{ c \in C_k \ | \ c.\text{count} \geq \text{minsup} \} \)
10) \( \text{end} \)
11) \( \text{Answer} = \bigcup_k L_k; \)

**Figure 1. Apriori Algorithm (Agrawal & Srikant, 1994)**

2) Reduce the number of comparisons. Rather than matching each candidate itemset with every transaction, we can reduce the number of comparisons by using more advanced data structures (Tan & Steinbach & Kumar, 2006).

Frequent itemset mining can be used in several ways to find different understanding in game data. One of the most important way is to find patterns among players (Drachen, 2012b). If the database is organised in a clear way than each instance describes player and the attributes describes the players’ playing style, so itemset mining is useful to find playing style characteristics that frequently co-occur (Drachen, 2012b).

### 2.6 League of Legends Researches

Blackburn and Kwak (2014) made a research about the LoL report and banning system. The Riot Games use a unique system called Tribunal. This system uses the wisdom of the crowd to make judgement of the toxic players’ behaviour. Instead of the players automatically get punished the other players can vote which players were guilty and which are not (Blackburn, Kwak, 2014). They defined two questions in their research: “Can we predict the crowdsourced decisions?” (Blackburn, Kwak, 2014, p.879) and “What do the important features imply? (Blackburn, Kwak, 2014, pp.880).” They made a prediction model to understand the crowdsourced decisions using data mining tools (Blackburn, Kwak, 2014). On the other hand, they explained “Toxic behaviour has been typically considered hard to define. If we obtain a good
quality supervised-learned classifier, it indicates the important building blocks in defining and understanding toxic behaviour.” (Blackburn, Kwak, 2014, p.880) In the research they realised that the model that they made had more accuracy with low agreement decisions (Blackburn, Kwak, 2014). Moreover, through their research they realised another important question for further studies, this model that they make can be used other computer games as well or not.

The Big Data concept finally reached the computer game industry as well. There are more than thirty-two million active user of the League of Legends. Altogether, the average hours that Riot Game keep is more than billion cumulative hours in every month and the Riot Games keeps track on everything: “Because it’s all happening inside a computer, [they] can attach a telemetry sensor to every player, every joint, every part of the field, and gather all that data” (Johnson, 2013). However, we have to know what exactly the main aspects of the game are where specialist can use these raw data. Johnson (2013) and Kennerly (2003) introduced five different types of usage of these data. Kennerly (2003) introduced four other types of usage of data mining: to balance the economy (in the meaning of the virtual world economy), to catch cheaters, to cut production costs and to increase customer renewal. He introduced a cycle about recycling old data into a new design. The steps are the next: live data collection, pre-process data, statistics, analysis, hypothesis and test (Kennerly, 2003). In contrast, Johnson’s article (2013) is about the pro players who can use these data to understand the gaming style of the enemy teams. He argued that these data are useful for some point, however, it’s not enough to understand the whole gaming style, because it contains a lot of strategical and more complex steps, which are not able to get from these datasets (Johnson, 2013). In conclusion, he believes that these data are useful for the beginners, who don’t have such a big experience with the game.

2.7 MEASURING PLAYER EXPERIENCE

One of the most crucial topic in player experience researches is how to measure the players’ in-game behaviour which is relates the subjective experience of playing a game (Drachen, 2012c). A computer game producer has a big interest to analyse a player’s behaviour (Drachen, 2012c). As analysing a player’s behaviour provides useful information about a player’s habit and reasons why or why not he/she is enjoying the game (Drachen, 2012c). Since the producer wants to keep the player they will invest to make unsatisfied players satisfied and therefore prevent them from
quitting the game (Drachen, 2012c). There are a couple evidences to make conclusions about this topic, such as game user research reports at the Game Developers Conference over the past few years. Big computer game companies such as EA and Microsoft have released a couple of case studies about analysing behavioural telemetry data (Drachen, 2012c).

In 2005 Larry Mellon and Nicole Lazzaro introduced the use of “fun metrics”, which is essentially the collection of metrics of user behaviour indicating the fact that the users enjoy or not the gaming experience. “Inferring PX from behavioural telemetry can however be prone to errors, as it is not possible to verify whether conclusions drawn from behavioural analysis are correct unless some form of measure of the playing experience is used” (Drachen, 2012c). To solve this issue, the solution is the triangulation, which means obtaining both, the experience measure and behavioural telemetry (Drachen, 2012c).

Another way to measure user experience is using qualitative or semi-qualitative methods, such as user feedback, like surveys or interviews, while combining usability testing (Drachen, 2012c). Meanwhile, the usability testing focuses on the operation testing of the game, the playability testing measures the player’s experience in-game (Drachen, 2012c). However, the biggest issue with this type of measurement is that the interaction flow between the players and the game is interrupted (Drachen, 2012c). There are alternative approaches, such as combining metrics with psychophysiological measures (e.g. heart rate) (Drachen, 2012c). Nowadays, a new type of player’s experience collection method became more popular, where pop-up surveys combined with behavioural telemetry to detect experience related problems (Drachen, 2012c).

2.8 DATA ANALYTICS FOR LEAGUE OF LEGENDS

A data analytics blog, called Manaless Blog, introduced an analyses about League of Legends champion benefits and most popular champions, using statistical methods and data mining methodologies. The data and their codes what they used for analyses are publicly available and in this paper them researches used for deeper research.

The first starter point in the research was to draw a scatterplot which introduced the kill and death ratios in each champion (Manaless Blog, 2015). The plot is crowded in
the middle, where most of the champions are take place (Manaless Blog, 2015). However, as a first point the graph shows that the supporter champions such as Janna or Nami has the lowest kills in the game and the lowest death rate, meanwhile the assassin champions Akali, Talon and Katarina tends to have higher kills in the game and highest death rate as well (Manaless Blog, 2015). On the other hand, this chart only uses two variables, while the Riot Games API contain hundreds of them (Manaless Blog, 2015). Due to this reason, data analytics blog made a deeper research to get the answer which hero is the most useable in the game, using 16 different variables (Manaless Blog, 2015). The first scatterplot result is in appendix 2 and the used dataset is shown in the appendix 5.

In his research he used a new type of method called Principal Component Analyses (PCA). "It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences (Smith, 2002, p.13). Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analysing data (Smith, 2002, p.13)". Another fundamental benefit of using PCA, if the pattern was already found than the help of the PCA is possible to reduce the number of dimensions without losing any information (Smith, 2002, p.13). “Essentially, the method assumes a relationship between data similarity and concentrations of energy in the eigenspace that spans the data (Sahouria & Zakhor, 1999, p.1290).” In the solution of the game analytics blog they used R methods for PCA analyses.

1) data <- read.table("***.csv", header = T, row.names = 1, sep = ",", as.is = T)
2) pca <- prcomp(data, scale = T)
3) x <- summary(pca)$x[,1:2]
4) plot(x, pch = 16)
   text(x, labels=row.names(data), cex = 0.8, pos=1)

   Figure 2. R code for the LoL scatterplot (Manaless Blog, 2015)

R is a language and environment for statistical computing and graphics (R Project, n.d.). R provide a widely range of statistical methods such as linear and nonlinear
modelling, time-series analyses and clustering, and provide graphical techniques and is highly extensible (R Project, n.d.). The data analytics blog R code is divided into four different parts (Manaless Blog, 2015). The first part loads the data into R (Manaless Blog, 2015). The second part performs the PCA itself, however specialist choose to normalise the data before, due to the variables such as number of kills and gold earned are operating in completely different scales (Manaless Blog, 2015). The third line contains the first two principals for plotting (Manaless Blog, 2015). The last part draws the scatterplot and adds the champion name labels to the chart (Manaless Blog, 2015).

The benefit of using the PCA is that the champions roles are automatically grouped together, we can see the support champions on the lower left corner, meanwhile the mid laners are in the lower right corner, and the support / mid laners are in the middle (Manaless Blog, 2015). “What we see here is that a support champion generally has high ward killed, ward placed, and assists (arrows to the left); mid laners tend to have high magic damage dealt to champions (arrow to the bottom); junglers and other non-jungle tanks tend to have an assortment of high damage taken and neutral monsters killed (arrows to the top); assassins and ADCs tend to have high gold, high champion kills, high minion kills, and high gold earned (arrows to the right) (Manaless Blog, 2015)”.

The final scatterplot of the model is in appendix 3.

### 2.9 DATA MINING INTRODUCTION

There are many definitions for data mining. Data Mining is a multi-disciplinary area, which used in science, engineering and business. The best known definition of the data mining is “The nontrivial extraction of implicit, previously unknown, and potentially useful information of data” (Piatetsky Shapiro). There are two types of data mining, called predictive analytics and descriptive analytics. These two types of analytics often called “knowledge discovery in data” or KDD. According to Elkan (2013), descriptive analyses is often fascinating and useful, however harder to get direct benefit from descriptive analyses than predictive analytics.

### 2.10 LIMITATIONS OF PREDICTIVE ANALYTICS

It is important to understand the limitation of the predictive analytics, and be aware the potential problems to face with. First of all, we cannot make progress without a
good training dataset which has an adequate size and quality (Elkan, 2013). On the other hand, it is really important to have a clear definition of the data mining concept, and to know what is the aim of the prediction (Elkan, 2013). Moreover, it is good to have historical examples of the concept (Elkan, 2013). According to Elkan (2013), for a successful data mining application “the actions to be taken based on predictions need to be defined clearly and to have reliable profit consequences. (p.9)” On the other hand, a successful data mining application, the new system can not cause any unintended consequences (Elkan, 2013). It means, a new system cannot really change or modify the behaviour dramatically (Elkan, 2013).

Another crucial point is the training data must be a representative of the test data (Elkan, 2013). Most of the time the training data is coming from the past, meanwhile the test data predict the future (Elkan, 2013). If what we want to predict is not stable over time, then the prediction has a chance that it will be less useful (Elkan, 2013). Last but not least, Elkan (2013) said that “for a successful application it helps if the consequences of actions are essentially independent for different examples” (p.9).

2.11 DATA MINING PROCESSES AND MODELS

One of the most popular data mining model is the CRISP-DM. In this section the different data mining models will be introduced. There are different models in data mining, the most popular are the CRISP-DM (Cross Industry Standard Process for Data Mining) (Chapman et.al, 2000), the SAS SEMMA (Santos & Filipe, 2008) model (Sample, Explore, Modify, Modell, Assess) and the IBM SPSS 5A, which based on the assess, access, analyse, act and automate steps (Santos & Filipe, 2008). However, lot of companies introduce and implement their own system.

2.11.1 CRISP-DM Processes

The CRISP-DM contains the following steps:

1) Business understanding: All the data mining work start with an initial phrase which focuses on the project objectives and requirements from the business perspective (Chapman et.al, 2000). Than this knowledge is converted into a data mining problem definition, and design initial plans for the data mining project. (Chapman et.al, 2000)
2) **Data understanding:** This phase starts with an initial data collection. Data miners in this step get familiar with the data, identify the quality problems “to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information. (Chapman et.al, 2000, p.10)”

3) **Data preparation:** This step covers all the required activities to construct the final data set from the raw initial data (Chapman et.al, 2000).

4) **Modelling:** The best model is selected and applied (Chapman et.al, 2000).

5) **Evaluation:** At this stage the models are evaluated and reviewed, that the model is fit for the business requirements. If it does not happen, then they have to start the modelling stage again (Chapman et.al, 2000).

6) **Deployment:** The model is implemented into the business processes (Chapman et.al, 2000).

The figure below shows the stages of the CRISP-DM processes.

![Figure 3. CRISP-DM life cycle (Chapman et.al, 2000, p.10)](image)

The sequence of the six stages is not rigid, as is schematize in Figure 3. CRISP-DM is extremely complete and documented (Santos & Filipe, 2008). All his stages are duly organized, structured and defined, allowing that a project could be easily understood or revised (Santos & Filipe, 2008).
2.11.2 SEMMA

The SEMMA system was developed and introduced by the SAS Institute (Santos & Filipe, 2008). The SAS Institute considered five different stages in the data mining processes:

1) *Sample:* In this stage from a huge dataset extract for a smaller dataset. To make the model, the dataset has to be big enough to contain the relevant information, yet small enough to be able to manipulate (Santos & Filipe, 2008).

2) *Explore:* This stage consists the data exploration process, where searching for unanticipated trends and anomalies to get a better understand on data and gain new ideas (Santos & Filipe, 2008).

3) *Modify:* This stage focuses on modifying the data by creating, selecting, modifying and transforming the variables (Santos & Filipe, 2008).

4) *Model:* This stage consists the data modelling part itself (Santos & Filipe, 2008).

5) *Assess:* “This stage consists on assessing the data by evaluating the usefulness and reliability of the findings from the data mining process and estimate how well it performs (Santos & Filipe, 2008, p.3). “

Although, the SEMMA model is independent from the chosen DM tool, the model was introduced as a guideline for the SAS Enterprise Miner software (Santos & Filipe, 2008). Moreover, this model is not really a data mining model, because it focuses on the analyses and model building phrases (Santos & Filipe, 2008).
According to Piatetsky (2014) research, currently the CRISP-DM is the most popular methodology. However, big proportion of the companies use their own solution, and the popularity of SEMMA is decreasing. Since 2014, the own solutions got bigger role, because the CRISP-DM is not prepared for the Big Data challenges, so until there will not be a new data mining methodology, the biggest companies have to find their own solutions (Piatetsky, 2014).
3 METHODOLOGY

3.1 OVERVIEW

In this research the CRISP-DM data mining processes will be followed. This process has six different steps; however, not all the steps are important in this paper, but additional steps are required. The steps are these: problem definition, which means what is exactly the aim of the research and what I will look for using the data mining tools (Chapman et al., 2000). The second step is the data exploration (Chapman et al., 2000).

In this thesis, the data collection part is divided into four different areas. The first area was the data collection part. For the data collection the research needed to find users who’s currently playing in different tiers and players whose already inactive. Active players have been defined the inactive players to the players who have not played in the 2016 season. All the data was collected from the western European servers. The number of players whose data have been used in this research more than one hundred. Approximately half of these players are inactive and the other half is still active player. The next dividing point in the research was into pro players and less experienced ones. There were introduced different classifications techniques for this. First, the players who never played ranked games before, and maybe even they have not reached the maximum level in the game. Secondly, the players who has ranked experience, however they are in the low ranked tiers, such as bronze and silver. Thirdly, the players who have significant amount of games in ranked mode, and they are or were in the top tiers, such as gold, diamond or platinum. The data was collected primarily from two different sources; forums and actual game participants.

3.2 PRACTICALITIES

There are no special costs of my dissertation that I am aware of. The dataset is available at the Riot Games main page for every registered user. The data will be stored at my personal computer in a password protected folder and the cloud drive which offered by the university for every student. The software (RapidMiner and Tableau) that I will use for analyses and data visualisation are free licensed software
or contains a free version with limited restrictions. The RapidMiner is free licensed software which supports not just data mining processes, text analysing and machine learning; on the other hand, it supports data visualisation techniques as well (RapidMiner, 2014a). However, the data visualisation and result analyses are a more convenient process with the Tableau which supports research and educational purposes.

One of the crucial parts of my dissertation is the achievability which is strongly connected to the time factor. To achieve my research aim I need to collect hundreds of username in the game which is fit to the description what I described in the objectives part. Before I start to use the dataset of the individual user I have to be aware of that the user will be relevant to my research. Moreover, I have to build a proper database for the analyses.

3.3 DATA COLLECTION AND PREPARATION

In this sector the data collection and preparation will be introduced. The original Riot Games data is written in JSON codes. “JavaScript Object Notation (JSON) is a lightweight, text-based, language-independent data interchange format. It was derived from the ECMAScript Programming Language Standard. JSON defines a small set of formatting rules for the portable representation of structured data” (Crockford, 2006). JSON code is represent two structured types, such as objects and arrays, and four primitive types, such as strings, numbers, booleans and null (Crockford, 2006). JSON is directly supported by the internet browsers, which leads to a significant performance (Nurseitov et. al, 2009). However, Crockford (2006) addresses such arguments, like “every object is a namespace. Its set of keys is independent of all other objects, even exclusive of nesting. Also, JSON uses context to avoid ambiguity, just as programming languages do,”. However, to get the usable data from this JSON formats the code had to be transferred into an excel readable format. It was necessary to use a pre-created JavaScript which was able to “read” this JSON format, and save it directly into CSV format. The required JavaScript code is shown in the Appendices (Appendix 6.).

The Riot Games collects four types of data points, game performance, system configuration, game configuration and game event information (Riot Games, 2015a). They collect information about when a player cast a certain spell, identify build orders,
death locations and win rates. However, in this research it is not relevant to use all the data which the Riot Games made available. Another complication, that the Riot Games made available all the data which is connected to the ranked games. It contains all the available information about the ranked matches since 2012. However, the players whose only have unranked games, the dataset is smaller, and needed to use other APIs or statistical methods to collect the required data.

The Summoner Details table contains the collected summoner names and summoner ID-s. It shows the highest tier what the player reached, the current level of the account and the fact that the current player is active or inactive. The table contains 123 data points in total.

The champions table contains the currently available and retired champions. It shows the information about the champion spells, and benefits and drawbacks and the main and secondary stats of the champion. Moreover, it contains information about the bot mechanism is enabled or disabled in that current champion. The table contains 132 champions in total.

The Recent Games table shows the information about each players recent ten games. This table describes the champions which used by the game, the stats what the players achieved and the items what they bought in the game. The table contains around 1200 data in total.

The Match Basic table contains the information about the player games in their most popular season. It shows which was the champion what they used, what kind of queue they joined what was the role and lane in the game. This dataset contains around 51,000 data.

The Ranked Statistics table shows the information about the players who played ranked games in their most popular seasons. It shows the summarized information about each type of ranked queues, shows the win ratio, and different aggregated statistics such as minion kills or champion kills. This table contains around 1000 data.

The Leagues Map table shows the relevant information about the current or previous ranked players according to their achievements. Its shows the tiers, the divisions, the number of wins and losses and the different league points changes over time. This dataset contains around 66 data.
The map data and item list is statistical tables, which helps to understand and explain
the item sets and map data in the game. The item list contains the information about
all the item in the game and the effects of the champion statistics. There are 213
currently available item in the game.

The data preparation is one of the most crucial and most time consuming parts of the
CRISP-DM processes (Chapman et al., 2000). In this part the clearing of the data and
fill the missing values are important. However, the collected dataset about the ranked
players does not contain any missing values. Meanwhile, the data about the unranked
players are not so clear. It contains information about the recent games, where some
of the results were made, but the number of the total matches, and statistics such as
number of aggregated kills, gold earnings or losses are missing. It made it complicated
to prepare the data for the analyses.

According to Witten et.al (2004), there are two different types of methods to replace
the missing values. Firstly, we can ignore these values or use a statistical method to
replace them. There is no best method to replace the missing values. In this research
there were four statistics approach which were used on the dataset. To replace the
missing values, the already known attributes average value was implemented, if the
cases required rather than using average, median and modus were used (Witten et.al,
2004). The benefit of this the easy implementation and application, however if the
already known attributes are random, then the results of the functions will end up as a
fake result (Witten et.al, 2004). The other method used in this paper for data collection
is the regression method. The aim of the regression method is that the we split the
dataset into two different parts. The first will be the dependent variables and the second
one is the independent variables (Witten et.al, 2004). In this case if the dependent
variables are missing we can estimate the values with the independent variables. This
method is more useful with the modelling phrase, rather than the analyses part.

The next step is the modelling. This part is involved making the different algorithms
and make the proper hypothesis (Chapman et al., 2000). The last two steps of the
CRISP-DM are the evaluation and deployment (Chapman et al., 2000). However,
these two steps do not really connect to my research. Instead of these steps I will make
a data visualisation about my results to make it clearer and easier to understand my
results.
3.4 ANALYSES

In this section the required analyses and tools will be introduced to find an explanation why players are stop playing a computer game or become expert in a game. On the other hand, in this section the data will be prepared for the further data mining model. Different data visualisation and data manipulation techniques will be introduced. Data visualisation is a way to present the data in a graph format (Witten et.al, 2004). It helps, to get an easier understanding of the results and identify the new patterns between the data (Witten et.al, 2004).

In the first analyses, the most popular champions and the champion choices between the pro and least pro players will be introduced. After collecting the data from the Riot Games API, there were missing attributes which was required for my first analyses. However, the League of Legends offers a feature in-game, where players are available to retrieve other players in-game performance. This feature shows the summoner top three favourite champions and the total number of win games in the different game modes (it does not show the total number of games). Between the favourite champions we do not make differences, so we are not ranking them as first favourite of the player, the second or third favourite of the player. The reason for this is two-sided. Firstly, the number of played games and earned experience with each top heroes do not show significant differences. Secondly, there are League of Legends rules and methods which require to do not devotion only to one hero. Each player can only use one hero in each game and another player cannot use that hero in that game, in ranked game mode the enemy team can choose three heroes which the is banned from the game, as well as the team setup require different hero choice depending on the line what the player choose to play.

For the analyses the excel built-in pivot table tools was used. Two different types of table were made. The first table shows the count of the heroes to know which one is the most popular and which one is the least popular heroes. The second table shows the count of the champions regarding to the summoner rank. For the easier understanding the tiers were splitted into three groups. The first group the “unranked” players have not changed. The second group is the “beginner” which includes the bronze and silver tier players. The last group is the “pro” which are the players in gold or higher tiers (diamond, master, challenger). The aim of this analyses was to get the answer is there any correlation between the champion choice and the different players
types, and there is a main aspect or least important question that players stop playing the game or keep playing it for a longer period of time.

A strongly connected analyses to the previous analyses is the in-game lane analyses. There are four different types of lanes in the game, bottom, jungle, middle and top. Each lane requires a different playing style. Champions have different capabilities regarding what lane they play within. These data were available in the match_basic table.

The next analyses was a ratio for the win. The match_basic table in my dataset contains each played ranked games. After summerise these matches we get the total number of the match. Meanwhile, the ranked_statistics table contains information about wins and losses. After this with an easy division we can get the win ratio. There were missing values in this section. Most of them with the players who has or had unranked status and the least of them the players who are or were pro players. To replace the missing values, average values were used. The average values were calculated using the same tier players win rates.

3.5 **Model Evaluation and errors**

According to the CRISP-DM processes, the data preparation is followed by the data modelling (Chapman et al, 2000). In this chapter, an initial data mining model will be built, followed by an introduction about the model evaluation. First to get the prediction model a decision tree was built. To build the tree, Hunt’s algorithm was followed. “Hunt's algorithm grows a decision tree recursively by partitioning a training data set into smaller, purer subsets” (Kesavaraj & Sukumaran, 2013, p.2). To be easier to understand the algorithm, let Dt is the set of training records which are associated with node t and \( y = \{ y_1, y_2, \ldots, y_c \} \) are the class labels (Tan & Steinbach & Kumar, 2006). Hunt’s algorithm contains two steps.

**Step 1:** “If all the records in Dt belongs to the same class in \( y_t \) then t is a leaf node labelled as \( y_t \)” (Tan & Steinbach & Kumar, 2006, p.152)
**Step 2:** If $D_t$ belongs to more than one class then the selected partition the records into smaller subsets (Tan & Steinbach & Kumar, 2006). From here a child node is created for each subset (Kesavaraj & Sukumaran, 2013).

For the decision tree model, RapidMiner was used. The model contains four different operators. The first operator is the X-Validation. “This operator performs a cross-validation in order to estimate the statistical performance of a learning operator (Rapidminer, 2016b)”. The main use of this operator to estimate the accuracy of the model. The second operator was the decision tree operator. It “generates a Decision Tree for classification of both nominal and numerical data (Rapidminer, 2016b)”. The apply model operator “applies an already learnt and trained model on the example set”, which is in our case is the decision tree model (Rapidminer, 2016b)”. The last operator which used in the research is the performance operator. “This operator is used for performance evaluation. It delivers a list of performance criteria values. These performance criteria are automatically determined in order to fit the learning task type” (Rapidminer, 2016b).

The used dataset is shown below.

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Description</th>
<th>Data Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tier</td>
<td>Contains the information about each player tier</td>
<td>Polynominal</td>
</tr>
<tr>
<td>Level</td>
<td>Current summoner level in the game</td>
<td>Integer</td>
</tr>
<tr>
<td>Inactive</td>
<td>Shows that the player was inactive or not in the current, 2016 season</td>
<td>Polynominal - Label</td>
</tr>
<tr>
<td>Total Ranked Game</td>
<td>Shows the total number of the ranked game</td>
<td>Integer</td>
</tr>
</tbody>
</table>
Fav_Line  | Shows the most popular line for each player | Integer  
--- | --- | ---  
Win Ratio | Shows the win ratio of the players, calculated from the total games divide with wins | Integer  

These were the main measures which used in the model. Although the confusion matrix provides the information how good is the performance of the classification model, to summarize this information into a single number would make it more convenient (Tan & Steinbach & Kumar, 2006). This can be done using performance metric such as accuracy, which is defined like this:

\[
\text{Accuracy} = \frac{\text{Number of correct predictors}}{\text{Total number of predictors}} = \frac{F_{11} + F_{00}}{F_{11} + F_{10} + F_{01} + F_{00}}
\]
In the build model the value of the accuracy was surprisingly good, 86%. Meanwhile the precision is 97.5%. The figure below (Figure 5) shows the confusion matrix of the model results.

<table>
<thead>
<tr>
<th></th>
<th>true no</th>
<th>true yes</th>
<th>class precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>pred. no</td>
<td>45</td>
<td>10</td>
<td>81.82%</td>
</tr>
<tr>
<td>pred. yes</td>
<td>1</td>
<td>24</td>
<td>96.00%</td>
</tr>
<tr>
<td>class recall</td>
<td>97.83%</td>
<td>70.59%</td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Confusion matrix of the results.

### 3.6 OVERFITTING

Even though the result accuracy and precision was convincing, the model is not ready for implementation. According to Tan et.al (2006), a classification model is generally divided into two types: training errors and generalization errors. “Training error, is the number of misclassification errors committed on training records, whereas generalization error is the expected error of the model on previously unseen records” (p. 172). This situation is called as overfitting. When the test error rates of the model are large, meanwhile the size of the tree is small, the situation called underfitting (Tan & Steinbach & Kumar, 2006). Meanwhile, the overfitting is when the “model has yet to learn the true structure of the data” (Tan & Steinbach & Kumar, 2006, p.174). The overfitting model based on the 14th century scientist work, called William of Occam, and his thesis called Occam’s Razor (Hawkins, 2003). According to the Occam, a model should only contain the procedures which is relevant and necessary and nothing more (Hawkins, 2003). So if two or three predictors is enough to explain y, then no more than these predictors should be used (Hawkins, 2003). This phenomenon is when the training error rate begins to increase, meanwhile the training error rate is decreasing (Tan & Steinbach & Kumar, 2006).

In the model that was introduced previously, the overfitting effect reason is the lack of representative samples. “Models that make their classification decisions based on a small number of training records are also susceptible to overfitting (Tan & Steinbach
These kind of models are mostly generated due to the lack of training data and learning algorithms which continue to process to generate newer leafs on the decision tree even when lack of the proper amount of training records (Tan & Steinbach & Kumar, 2006).

### 3.6.1 Pruning

In the previous section the overfitting effect was described. In this section, there will be an introduction about how the overfitting was handled, introducing two different methods. By building a complete tree and pruning afterward I adapted a method called postpruning rather than prepruning (Witten & Frank & Hall, 2011). The aim of the prepruning is to stop the tree to make more subtrees. The benefit of this method is to avoid generating overly complex subtrees to overfit the training data (Tan & Steinbach & Kumar, 2006). However, it is really difficult to choose the right threshold. If the threshold is too high it will result underfitted models, while if it is too low, it “may not be sufficient to overcome the model overfitting problem” (Tan & Steinbach & Kumar, 2006, pp. 184).

There are two different operators for postpruning, subtree replacement and subtree raising (Witten & Frank & Hall, 2011). At each node, the pruning model decide that it should use subtree replacement, subtree raising or leave the subtree as it is, unpruned (Witten & Frank & Hall, 2011). The idea behind the subtree replacement is to select some subtrees and replace them with a single leaf (Witten & Frank & Hall, 2011). The subtree raising is more complex and not always worth it to implement (Witten & Frank & Hall, 2011). The idea behind it is to change the original subtree into a new subtree (Witten & Frank & Hall, 2011). This new, changed subtree can contain other subtrees as well (Witten & Frank & Hall, 2011). “Postpruning tends to give better results than prepruning because it makes pruning decisions based on a fully grown tree, unlike prepruning, which can suffer from premature termination of the tree-growing process (Tan & Steinbach & Kumar, 2006, pp. 185).

### 3.7 Final Data Mining Model

In the final data mining model, the tier, level and total number of games were the best measurements. After applying prepruning and postpruning, the RapidMiner automatically did not use the win ratio, favourite line and favourite heroes’
measurements. The reasons for this, because there is no strong correlation between these attributes and labels. In this case the label shows the player is inactive or not.

To avoid overfitting effect, the postpruning confidence values was set to 0.25. The confidence value shows the confidence level used for the pessimistic error calculation. In prepruning four different types of parameters were set up. First, the minimal gain was set up to 0.1. This value shows the gain of node is calculated before splitting it. Secondly, the size of a leaf node is the number of examples in its subset. The minimal leaf size value was 2. Next, the minimal size for split value is equal to the total number of examples in the dataset, so the tree is generated in the way that every leaf node subset has at least the minimal leaf size number. Only those nodes were split which greater than or equal the minimal size parameter. The value for it was 4. Finally, number of prepruning alternatives parameter adjusts the number of alternative nodes tried for splitting when split is prevented by prepruning, the value of it was set up to 3. The used operators for the final model were the same as it was introduced earlier, the differences in the final and previous model are based on the measurement changes, the data collection and data manipulation.

3.8 Ethical Aspects

The Riot Games made available for every registered user to download and use the different data about the gaming style. These data contain information about the spells, potions, champions and matches what you or other players played. The data attached for each players’ username. However, to access to players’ data I need the individuals’ username. These usernames are available playing in the arena and/or different forums. However, collecting nicknames about the users lead to a potential risk. The data will be used anonymously and the data will be kept in a password protected folder on my personal computer. There are no potential risks to participate in this research and the users will not get any respond that I use their nicknames. Moreover, the usernames will be kept confidentially; those will not be publicized in my dissertation. After the discussion with my supervisor I understood that my research contains “low risk” data elements, because there is a possibility that the username can lead back to the name of an actual user. The results of this study will be included in my master’s dissertation which will be publicly available.
4 RESULTS AND INTERPRETATION

4.1 ANALYSES RESULTS
The Blackhauge et.al (2012) research shows, that there are a significant number of players whose stop playing after a couple of hours spending in the game. This paper result tries to find an explanation what are the reasons for this effect, and what are the factors which cause this. The results show that there are big differences between the three groups of player that were used in the analyses.

![Figure 6. Average win ratio chart](image)

The unranked players average win ratio is about 45.68%. Meanwhile the beginners win ratio is 52.64% and the pros are 59.07%. Like the results show, the differences are significant. The explanation for this is two-sided. First of all, the League of Legends has a matchmaking system (the system which creates the matches and pull together the players), has some problems. Riot Games had to solve the issue, the players who play against each other have the same level and skills, meanwhile the queues are short. The problem with this system is, that the started players often play against players who have more experience and skill in the game. On the other hand, the beginner players with less experience want to learn the game mechanism and want to enjoy the game, and not being competitive right in the beginning. The representative difference between the win ratios show that beginner players loose more often than
the more experienced ones. This reason lead to the fact that players leave the game in the early period because of lack sense of satisfaction.

The champion use shows big differences as well. The most popular champions in the game is Vayne, Blitzcrank and Yasuo. Meanwhile Vayne and Yasuo are assassin type of champion, Blitzcrank is a tank / supporter type of champion. The least popular champions in the game are the healer / supporter champions. However, there are representative differences at the results of the champion use, divided into different tiers. Meanwhile, the unranked players prefer to play with Master Yi and Blitzcrank, however, there are no significant scatter between the champions. It means the unranked players play with each champions in the game, and the data do not show especially favourite heroes. The beginner players, in bronze and silver tiers are playing Vayne, Blitzcrank and Yasuo. Like the first analyses shows that the most popular heroes are the same like the “beginner” champion’s use. The reason for this is the biggest tier, most of the players play in this level. The pro players’ favourite champions are Vayne, Riven and Yasuo. It shows that each favourite hero are assassin heroes, with huge damage potential, good gold earnings, however big death rate as well, according to the game analytics research which was introduced earlier.

![Most popular heroes by tier](image)

**Figure 7. Most popular heroes in the game**

The analyses result is based on the lane choices for each individual player. The most popular lanes were bottom and top, meanwhile the least popular lanes are the jungle and mid. There were no significant differences between the player tier level and lane
choice, or the activity and lane choice as well. The fact that the bottom and top lane are the most popular is based on the game mechanism. There are five players in each team and four lanes in the game. Each lane requires at least one or more heroes’ presence. However, according to different League of Legends forums and former tournament analyses, the most working tactic is two players are going to the bottom or top lane, meanwhile one – one player is going to the last two lanes. The results show that only an insufficient number of unranked players tried or played at jungle lane. The explanation probably is that the jungle lane requires the least teamwork, and the chance to fight against enemy champions are smaller, than at the other lanes.

Figure 8. Lane popularity chart

4.2 MODEL EVALUATION

In this section the data mining model will be evaluated. Like the results show the accuracy of the model is 70.45%, meanwhile the precision is 43.75%. The reason for the low accuracy and precision is probably the fact, that game continuation is highly depends on the psychological factors rather than the raw data. Like the Figure 9. shows the number of “no” predictors are high. The “no” predictor shows the players who will not stop playing the game. Meanwhile, the number of “yes” predictors, so the players who has a significant chance that they will stop playing is low. In this case the confusion matrix, among 122 data points that are classified, 86 was correctly classified, meanwhile 36 are misclassified.
Rather than using two different models, it is more convenient to have a single one, due to in this research will be computer two metrics from the confusion matrix, which will be converted into one:

1. True Positive Rate (TPR), defined as: \[ \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \]

   “Intuitively this metric corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. In other words, the higher TPR, the fewer positive data points we will miss. (Dernoncourt, 2015, p.129)”

2. False Positive Rate (FPR), defined as: \[ \frac{\text{False Positive}}{\text{False Positive} + \text{True Negative}} \]

   “Intuitively this metric corresponds to the proportion of negative data points that are mistakenly considered as positive, with respect to all negative data points. In other words, the higher FPR, the more negative data points we will misclassified. (Dernoncourt, 2015, p.130)”

To get one single metric from FPR and the TPR combined, firstly the two former metrics, which have many different thresholds, must be calculated for the logistic regression (Dernoncourt, 2015). Secondly these should then be plotted onto a graph. The TPR values should be placed on the ordinate whilst the FPR values should be on the abscissa (Dernoncourt, 2015). The curve that appears is known as the ROC curve. The metric that is used is the AUC area (Dernoncourt, 2015).
Figure 10. The model AUC and ROC curve

In this figure (Figure 10.), the blue area corresponds to the Area Under the curve of the Receiver Operating Characteristic (AUROC). The dashed line in the diagonal we present the ROC curve of a random predictor: it has an AUROC of 0.6. Meanwhile the value of the AUC is 0.759.

The Figure 8. below show the decision tree results. The decision tree contains four main areas. The results will be introduced starting from the left side of the tree. The first part we can see, that the players whose are lower than level 6 have the biggest chance that they will stop playing it. This result compare to the result of Bauckhage et al. (2012) is same. Bauckhage et.al (2012) argued in his research, there is a significant number of players who stop playing after a couple of hours.

Another crucial point in the model the players who are level 14-17. The chance that they will stop playing with League of Legends is high. This is the range where the players are not starters anymore, however, they did not reach the max level. They have already invested a decent amount of hour in the game, however, they became bored of it. A quite similar result is the players who has less than 578 matches, but they did not reach the maximum level.

The highest chance that somebody will not stop play the game, the players who are in the “beginner” or “pro” tiers, they have more than 578 matches and they are in the top level. The Figure 11. below show the results of the decision tree.
4.3 LIMITATIONS OF THE ANALYSES

The limitations of the study were two sided. First of all, the lack of the former studies and modelling processes about player engagement and using telemetry data. There were two big difficulties to address the issues in modelling the player engagement. First of all, interest of playing a game is an abstract quantity. “Psychologist would call it a latent or hidden variable which influences measurable quantities such as the frequency of playing sessions or total playing times but cannot itself be observed directly (Bauckhage et.al, 2012, p. 140).

Secondly, the industry needs abstract mathematical models which can be applied to the real world and not only to a set of data. This implies one of the most important features of such a developed model is its implementation in the real world. Even though, data analysts try to justify these theoretical and abstract models it is rarely possible to interpret these assumptions arise by mathematics (Bauckhage et.al, 2012). Precondition for results of these models to be useful for the industry is their provision of real explanations giving the concrete needs and requirements of game developers (Bauckhage et.al, 2012).

Another limitation was the lack of proper data. The Riot Games made available all the ranked game details, however, the lack of the unranked matches or the number of
missing values in these dataset made it really hard to make a comparison between the
different player types. The research data are based on a significant number of
prediction and statistically used data, rather than raw telemetry data. The reason why
the Riot Games did not make these data available is coming from their company’s
privacy policy issues. They believe that the ranked games are public, meanwhile the
not ranked games are everyone’s private information. However, if the complete
dataset would be available on the Riot Games API, the results would have been more
accurate, and would have been able to make more remarkable studies about the
connection between game telemetry data and player engagement at the computer
game.
5 CONCLUSION

5.1 CONTRIBUTION TO KNOWLEDGE

Data mining starts to get more and more important role in the computer game industry. Since the Big Data millions of data collected and the companies just start to learn how to use these data for their benefits. Specialist can explore how people play the game, how much time they spend playing and predict when they will stop playing or predict what they will do while playing (Drachen, 2012a). This research aim was to data mine League of Legends telemetry data to compare new players versus experienced ones and to understand what causes game continuation or end.

In the literature review the former researches were introduced about the computer game player’s engagement, League of Legends former analyses and the data mining itself. In the methodology section the CRISP-DM processes were followed (Chapman et.al, 2000). This method is a data mining methodology which based on six steps, data understanding, data collection, data preparation, model build, model evaluation and implementation. There were not introduced all the sections in this paper, the aim was to introduce in basics a data mining process, using League of Legends telemetry data. The methodology focused on the data collection and preparation, as well as the data mining model build and evaluation.

The results section introduced the analyses findings and the data mining model evaluation. The analyses results show significant differences between the tier groups, however, it does not show such a significant difference between the players who already stopped playing and players who are still playing. However, the players who are already stopped playing the game are coming each tier, meanwhile the proportion of the players whose are not playing is bigger in the unranked and low level tiers than in the pro tiers. The data mining model showed that the best measures to predict that a player will stop playing the game or not are the current level, tier and total ranked games, rather than the playing mechanism, such as win ratio, favourite line or champion choice.
5.2 Further Research

To understand the gamer behaviour, the game data is not enough. There are researches which maps the telemetry data, and there are researches which explore pattern at the psychological effects of the computer games, however the connection between these analyses are missing. The factors such as friendships in the computer games and fact that the players start playing with a computer game and the possibility for that if the players friend stop playing in the computer game will have an effect or not to the other users or not are still missing. There are methods which used by the most significant computer game companies, to strengthen the relationship between players. However, these solutions do not really work at the moment.

One of the big issue with the CRISP-DM method that it is hard to implement for Big Data analytics (Chapman et.al, 2000). The CRISP-DM was conceived in 1996, and since than the fundamentals of the method have not changed (Chapman et.al, 2000). However, a new open source standard, the PMML will replace the old methods, according to the research of Guazzelli et.al (2009). “The PMML package exports a variety of predictive and descriptive models from R to the Predictive Model Mark-up Language (Guazzelli et. al, 2009, p.1).” PMML is based on XML language, and nowadays it started to be de facto standard not just for predictive and descriptive models, but also data pre-processing and post-processing (Guazzelli et. al, 2009). One of the biggest advantage of this model is that it interchanges between different tools and environment, avoiding proprietary issues and incompabilities (Guazzelli et. al, 2009).

A strongly connected development for the PMML system is the ADAPA scoring engine, which was introduced by Zementis (Guazzelli et. al, 2009). The ADAPA engine enable to the users to upload their models over the Internet, and other users can “execute or score the models with any size of dataset in batch-mode or real time” (Guazzelli et. al, 2009, p.5).

(Word count: 11590)
6 REFERENCES


Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. (2000). *CRISP-DM 1.0.* SPSS Inc., (pp. 6-64)


Oosterhuis, K., Feireiss, L. (2006). The architecture Co-Laboratory: Game Set and Match II. Delft University of Technology. (pp. 172-181)


R Project (n.d.). *What is R*? Retrieved from: https://www.r-project.org/about.html


Appendix 1. “Observed empirical distributions of playing times per player (in hours) and fitted first passage time distributions for the two single-player games Just Cause 2 (JC2) and Tomb Raider: Underworld (TRU). Shape and scale parameters of the fits were obtained from maximum likelihood estimation (Blauckhage et.al, 2012, p.143).
Appendix 2. Shows the champions ordered by average death per game and average champion kills per game (Manaless Blog, 2015).

**Appendix 4.** Introduction the dataset which used be the Manaless Blog (2015)

<table>
<thead>
<tr>
<th>Column Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>champion</td>
<td>Contain all the champions in the game</td>
</tr>
<tr>
<td>matchduration</td>
<td>The average duration of the matches of each champion</td>
</tr>
<tr>
<td>kills</td>
<td>The average number of kills of each champion</td>
</tr>
<tr>
<td>deaths</td>
<td>The average number of deaths of each champion</td>
</tr>
<tr>
<td>goldearned</td>
<td>The average rate of the earned gold of each champion</td>
</tr>
<tr>
<td>killingspeers</td>
<td>The average number of killing sprees (killing spree is killing more than two champions in a short period of time)</td>
</tr>
<tr>
<td>physicaldamagedealltochampions</td>
<td>The physical damage dealt to other champions in the match (average value)</td>
</tr>
<tr>
<td>magicdamagedealltochampions</td>
<td>The magic damage dealt to other champions in the match (average value)</td>
</tr>
<tr>
<td>physicaldamagetaken</td>
<td>The physical damage taken by other champions (average value)</td>
</tr>
<tr>
<td>magicdamagetaken</td>
<td>The magic damage taken by other champions (average value)</td>
</tr>
<tr>
<td>truedamagedealltochampions</td>
<td>The true damage dealt to other champions (true damage is the damage which is actually not reduced by armour and magic resistance)</td>
</tr>
<tr>
<td>truedamagetaken</td>
<td>The true damage taken by the champion</td>
</tr>
<tr>
<td>wardsplaced</td>
<td>The number of placed wards (ward gives extra help in the fight, such as visibility, cooldown reduction, etc.)</td>
</tr>
<tr>
<td>wardskilled</td>
<td>The number of killed wards.</td>
</tr>
<tr>
<td>minionskilled</td>
<td>The average number of killed minions (minions are respawning on the top, bottom and mid lane in every 30 seconds)</td>
</tr>
<tr>
<td>neutralminionskilled</td>
<td>The number of natural minions killed in the game (natural minions are the minions which are in the jungle lane)</td>
</tr>
</tbody>
</table>
Appendix 5. The Rapid Miner model

The second picture shows the training and testing procedures after the validation process.
Appendix 6. The JavaScript code to convert JSON into CSV files.

```html
<html>
<head>
  <title>Demo - Convert JSON to CSV</title>
  <script type="text/javascript" src="http://code.jquery.com/jquery-latest.js"></script>
  <script type="text/javascript" src="https://github.com/douglascrockford/JSON-js/raw/master/json2.js"></script>
  <script type="text/javascript"> // JSON to CSV Converter
    function ConvertToCSV(objArray) {
      var array = typeof objArray != 'object' ? JSON.parse(objArray) : objArray;
      var str = ''; 
      for (var i = 0; i < array.length; i++) {
        var line = ''; 
        for (var index in array[i]) {
          if (line != '') line += ',';
          line += array[i][index];
        }
        str += line + '\n';
      }
      return str;
    }
    // Example
    $(document).ready(function () {
      // Create Object
      var items = [
        { name: "Item 1", color: "Green", size: "X-Large" },
        { name: "Item 2", color: "Green", size: "X-Large" },
        { name: "Item 3", color: "Green", size: "X-Large" }];
      // Convert Object to JSON
      var jsonObject = JSON.stringify(items);
      // Display JSON
      $('#json').text(jsonObject);
      // Convert JSON to CSV & Display CSV
      $('#csv').text(ConvertToCSV(jsonObject));
    });
  </script>
</head>
<body>
  <h1>JSON</h1>
  <pre id="json"></pre>
  <h1>CSV</h1>
  <pre id="csv"></pre>
</body>
</html>
```
Access to Dissertation

A Dissertation submitted to the University may be held by the Department (or School) within which the Dissertation was undertaken and made available for borrowing or consultation in accordance with University Regulations.

Requests for the loan of dissertations may be received from libraries in the UK and overseas. The Department may also receive requests from other organisations, as well as individuals. The conservation of the original dissertation is better assured if the Department and/or Library can fulfill such requests by sending a copy. The Department may also make your dissertation available via its web pages.

In certain cases where confidentiality of information is concerned, if either the author or the supervisor so requests, the Department will withhold the dissertation from loan or consultation for the period specified below. Where no such restriction is in force, the Department may also deposit the Dissertation in the University of Sheffield Library.

To be completed by the Author – Select (a) or (b) by placing a tick in the appropriate box

If you are willing to give permission for the Information School to make your dissertation available in these ways, please complete the following:

☐ (a) Subject to the General Regulation on Intellectual Property, I, the author, agree to this dissertation being made immediately available through the Department and/or University Library for consultation, and for the Department and/or Library to reproduce this dissertation in whole or part in order to supply single copies for the purpose of research or private study

☐ (b) Subject to the General Regulation on Intellectual Property, I, the author, request that this dissertation be withheld from loan, consultation or reproduction for a period of [ ] years from the date of its submission. Subsequent to this period, I agree to this dissertation being made available through the Department and/or University Library for consultation, and for the Department and/or Library to reproduce this dissertation in whole or part in order to supply single copies for the purpose of research or private study

Name: David Otto Istvanovszki
Department: Information School
Signed: [Signature] Date: 30/08/2016

To be completed by the Supervisor – Select (a) or (b) by placing a tick in the appropriate box
(a) I, the supervisor, agree to this dissertation being made immediately available through the Department and/or University Library for loan or consultation, subject to any special restrictions (*) agreed with external organisations as part of a collaborative project.

*Special restrictions

(b) I, the supervisor, request that this dissertation be withheld from loan, consultation or reproduction for a period of [ ] years from the date of its submission. Subsequent to this period, I, agree to this dissertation being made available through the Department and/or University Library for loan or consultation, subject to any special restrictions (*) agreed with external organisations as part of a collaborative project.

Name

Department

Signed Date

THIS SHEET MUST BE SUBMITTED WITH DISSERTATIONS BY DEPARTMENTAL REQUIREMENTS.