Exploring the most important variables for winning a professional League of Legends match

C.J. Slotboom
STUDENT NUMBER: 2029764

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE OR DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee:
Dr. Maryam Alimardani
Dr. Silvy Collin

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
January 2021

Preface

Before you lies the Thesis “exploring the most important variables for winning a professional League of Legends match". This study is based on the dataset openly provide by Oracle elixir. This Thesis was written for the master program Data Science & Society at Tilburg University. I would like to thank my supervisor Maryam Alimardani for the continues support during my thesis.

Exploring the most important variables for winning a professional League of Legends match

C.J. Slotboom

E-sport is a growing business. One of the most popular e-sport games is League of Legends. League of Legends has a very popular competitive scene in which players from all around the world play against each other to obtain prizes which go into the millions of dollars. In this study the most important in-game statistics are analyzed and used to predict matches of professional League of legends teams. In order to achieve this, a dataset which contains professional League of Legends matches is used. By using the data, a team score is created. This team score is intended to be used as a predictor for future matches. By making use of various machine learning techniques such as Support Vector Machine, k-Nearest Neighbor and Random Forest, the most important in-game statistics are analyzed. The statistics provide excellent conditions for these classification models, some predicting up to 99% accuracy.

1. Introduction

This thesis explores the use of feature engineering and various machine learning algorithms to find the most valuables in-game League of Legends (LoL) statistics. These will be applied to find the most important variables for winning a professional LoL match.

E-sport is, compared to traditional sport, a fairly new branch with a lot of potential. Since e-sport in a fairly new branch, the amount of viewers are much lower compared to traditional sports. On the other hand, e-sport is growing each year and might reach equal viewers compared to traditional sports in the future (Jones, 2019).

While traditional sports have many studies conducted around win prediction, e-sport still has opportunities in this area. LoL has a daily player base of approximately eight million players. Therefore, making LoL the game with the second highest player base in the world (Gibson, 2020).

Many LoL players are playing to grow in their rank and improve their playstyle. Not only does LoL have a balanced ranking system but also a competitive side. Each year there are multiple tournaments which are located all around the world. The LoL season ends with the world championship[1]. At the world championship the best teams play against each other to win the world title and take a piece of the prize pool. This year the price pool was $2,225,000 in total[2]. The pro teams train all year to find their strengths and weaknesses. This is mostly done

[1] Riot Games. (n.d.). Lol e-sport. Retrieved December 3, 2020, from https://lolesports.com/news
[2] Liquipedia. (n.d.). 2020 World Championship. Retrieved December 3, 2020, from https://liquipedia.net/leagueoflegends/World_Championship/2020

by live coaching, reviewing matches and overall performances. Currently, there are no studies that analyze the game with various machine learning techniques to find the most important statistics for a win. Therefore, this study will focus on these statistics by using feature engineering. These statistics will be analyzed to create a team score. This team score is an indication of how well the team performed. The field of e-sports, specifically LoL, has a research gap surrounding this problem. Within this study the most important game statistics will be identified and multiple machine learning approaches are used to find the best way to improve e-sport performances. This will provide team coaching, players, fans and other stakeholders an easy way to analyze and understand LoL matches without having to understand all the elements of the game. This study also focuses on exploring the most important variables for pro teams to win an e-sport match. In order to contribute to diminish this research gap, this study will answer the following research questions:

- What in-game statistics are the most valuable to win?
- In what way can a scoring system, based on the in-game statistics of a team, predict the chance of winning a professional League of Legends match?

The structure of the thesis is as followed. Chapter 2 is about the related work of prediction techniques in sports, e-sports and League of Legends. Chapter 3 contains the models used in this study and the overall general approach. Chapter 4 contains the details of the dataset and the experimental procedure. Chapter 5 presents the results of the study, followed by chapter 6 where the results are evaluated. Finally, chapter 7 elaborates on the conclusion of this thesis.

2. Related work

This chapter provides the theoretical framework of the thesis.

2.1 Win prediction in traditional sports

Traditional sport and e-sport share a lot of similarity in the way games are analyzed. For this reason, many studies around traditional sports are evaluated based on the various machine learning techniques they use to try and predict the winner of a match. Another reason to compare traditional sport with e-sport is that traditional sports are enormously popular and well researched, thus having a broad selection of studies to explore. One of the most popular traditional sport is football. Tax and Joustra (2015) applied nine classification algorithms on Dutch football matches. Their study resulted in a classification accuracy of 54% while using various machine learning techniques. The best performing classifiers were naïve Bayes and Artificial Neural Network.

While Razali, Mustapha, Yatim, and Ab Aziz (2017, p. 99) succeeded in predicting the results of English Premier League football matches with an accuracy of 75% using Bayesian

Networks. Another example is the study conducted by Hsu (2020, p. 84). Hsu applies classification models such as Neural Network, Support Vector Machine and ensemble learning to predict the outcome of a professional American Football league match. Hsu applied feature engineering to select the most influential factors. As a result of selecting the most influential factors the classification models reached a prediction accuracy close to 70%.

Passi and Pandey (2018, p.32) made predictions based on the selected players in a game of cricket. By using naïve bayes, decision trees, random forest and SVM the study got a prediction accuracy of 43%, 77%, 89% and 50%.

Given the above it can be concluded that various machine learning techniques can be used to achieve high prediction accuracy. However, according to Bunker and Thabtah (2019, p30), there is still a need for more accurate models. Bunker and Thabtah did a summarizing study on the use of multiple machine learning techniques. These studies vary from the National Football League, rugby, English Premier League Football, Dutch football and horse races. Most of those studies provide promising results. Even though these studies show promising results, there is still a high demand for better models. The reason for this is that there are a ton of betting websites and managers seeking for new information about either their own team or the team they have to play against (Bunker & Thabtah, 2019, p. 30).

2.2 Prediction in e-sport

Now there is knowledge on prediction in traditional sports, it is relevant to look into e-sport. E-sport is becoming a booming business. The worldwide revenue of the e-sport market in 2020 is 950.3 million US dollars[3]. Expectations are that this will grow to about 1.6 billion US dollar in 2023. At the moment the revenue and viewers of traditional sports are still much higher compared to e-sport. However, this can change in the future (Jones, 2019b).

Whereas traditional sports mostly stay the same, as in, there are rarely any big changes in the way the sports are, the area of e-sports is a continuously changing. Certain updates can change a game completely, such as champion reworks or adding new items.

DOTA2, a Multiplayer Online Battle Arena (MOBA) which is very similar to LoL, updates the game almost monthly[4]. Even though most e-sport games change continuously, various machine learning techniques applied on traditional sports are also applicable on e-sport.

In the study of Yang, Qin, and Lei (2016) classification methods such as logistic regression are used to predict a win based on the selected champion. Whereas Anshori, Marri, Alauddin, and Abdurrahman Bachtiar (2019, p. 9) use SVM to predict the winner with an accuracy of 60%. Both of these focus on the selected champions or a combination of champions. Another

[3] Statista. (2020, October 13). Revenue of the global eSports market 2018-2023. Retrieved December 3, 2020, from https://www.statista.com/statistics/490522/global-esports-market-revenue/

[4] Valve. (n.d.). Game Updates. Retrieved December 3, 2020, from http://www.dota2.com/news/updates/

form of win prediction is done by Hodge et al (2020b, p.1). Hodge et al.(2020b, p.1) applied feature engineering and machine learning techniques to do live game prediction on a game of DOTA 2. By selecting the right features some models got a prediction accuracy of 85% after five minutes of gameplay. Most of these studies focus on real-time predictions, selected champions, combination of champion etcetera. What can be seen as a limitation in these studies is the availability of professional games. Most of the studies are based on amateur game datasets. Hodge et al. (2020b, p. 1) concluded in his study that he had insufficient professional data to work with. This is also the problem in some of the League of Legend predictions.

2.3 prediction in League of Legends

Just like DOTA2, as mentioned in the previous paragraph, League of Legends is an example of a MOBA game. The game consist out of two teams, each containing five players. These two teams fight against each other in an arena. This arena is divided into four parts; bottom lane, mid lane, top lane and jungle as seen in figure 1. Each lane has certain roles, champions and playstyles. The main goal of the game is to destroy the nexus of the enemy. The nexus is the last objective in the enemy base. To reach the main goal, the teams have to fight against each other, minions, monsters and try to obtain various objectives. When killing players, minions, monsters or objectives the players are rewarded with gold. The gold is used to buy various items which also gives a champion advantages such as increased in-game statistics or empowered minions. Some objectives also give a champion advantages against other champions[5].

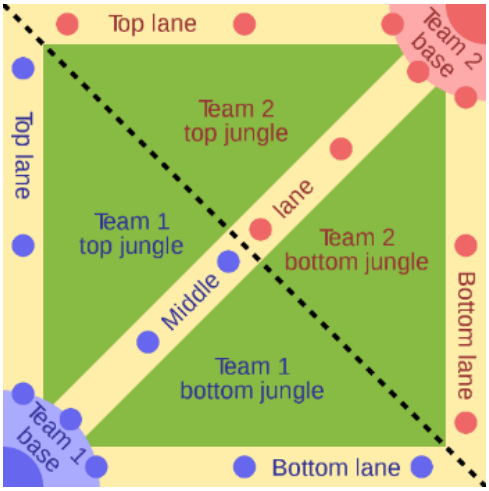

Figure 1. *General arena of a MOBA-game*

[5] League of Legends. (n.d.). Retrieved December 3, 2020, from https://na.leagueoflegends.com/en-us/

As discussed before, there is little research around feature engineering and win prediction of LoL professional e-sport data. As well as in other MOBA studies, most of the research around LoL is based on old data or data from amateur games. The study of Cong (2020) used the same machine learning techniques as the traditional sport and other MOBA win prediction techniques such as SVM, naïve Bayes and kNN. Even though the predictions work well, it still misses the element of professional data. Another study, the research of Novak, Bennett, Pluss, and Fransen (2019, p. 5) compared the win ratio of teams from North America (NA) versus Europe (EU). The study compares various statistics and win rates to see the overall difference between NA and EU. However, the data of the study is limited since the dataset only contains a total number of 30 games (15 NA games, 15 EU games). This is a relatively low amount.
The research gap in these studies is a feature engineering to obtain the most important features and using these to create a team score. Despite of the various studies conducted on e-sport games, there are no studies that create a team score. Providing professional LoL teams with a team score helps to improve the overall performances of individual team members as well as the team (Driskell & Salas, 1992, p. 285). The team score is not the only research gap. Likewise there is also little to no research of statistical analysis on professional e-sport games.

3. Methods

This chapter includes the description of the models used in this study.

3.1 Classification models

In this study, classification models are used to make predictions. A classification model observes values and tries to draw conclusions based on the observations. Depending on the input, a classification model will try to predict the value of this input. The output of a classification model is split into two classes, either positive or negative. The output is often either “0” for the negative class or “1” for the positive class (Dreiseitl & Ohno-Machado, 2002, p. 56). In this study the positive class “1” stand for a win, whereas “0” stands for a loss. This study makes use of supervised models. Supervised models are used for classification tasks (Love, 2002, p. 32). Models used for classification tasks are kNN, Trees, Logistic Regression, Naïve Bayes and SVM. These models are widely used as discussed in chapter 2. The models applied in this study are Random Forest, Support Vector Machine and k-Nearest Neighbor.

3.2 Random forest

The first applied model is Random Forest (RF). RF consist out of multi decision trees. The name decision tree comes from the tree-like structure. The tree-like structure is built from many if-then rule set. The advantages of using RF is that the model often produces a good prediction

result. The hyperparameters are easy to interpret because there are only a few. A drawback of RF models is that they easily overfit. This can, most of the time, be solved by adding more trees. RF are easy and fast to train but adding more trees can make the model slow (Oshiro, Perez, & Baranauskas, 2012, p. 165).

3.3 Support Vector Machine

The second model applied in this study is Support Vector Machine (SVM), which is a binary classifier. As discussed in chapter 3.1 it separates values into two classes. The support vectors are the data points which are the nearest to the hyperplane. The hyperplane is the line that separates the two classes. The hyperplane tries to get a best possible margin. The margin is the distance between the nearest point of each set (Noble, 2006, p. 1565). A SVM can do nonlinear classification by replacing the linear SVM-algorithm with a kernel function. This way the models can fit to the maximum-margin hyperplane (Noble, 2006, p. 1565).

3.4 k-Nearest neighbor

The third model is k-Nearest neighbor(kNN), which is called a lazy learner. A lazy learner does not learn in the training period. A kNN is an easy implementable algorithm because it only has two parameters. The parameters for a kNN are either Euclidean or Manhattan. The Euclidean distance gives the minimum distance between two points. The Manhattan distance is used to calculate for distances in a grid like path. Manhattan is a preferred parameter when there is an increasing number of dimensions (Beckmann, Ebecken, & Pires de Lima, 2015, p. 112).

3.5 Statistical test

In order to extract the most important in-game statistics, feature engineering was applied. This was done as discussed in chapter 4. All game categories, which are further discussed in chapter 4 (player vs player, player vs minions/monsters, vision, objectives, gameplay), were analyzed on correlation between the variable and result. To explore the correlation between the variables of table 2,3,4,5,6, linear models(lm) and local regression(loess) are applied. In addition to exploring the variables on correlation, the machine learning techniques as discussed in chapter 3.2, 3.3 and 3.4 are used. All three models are tested for accuracy. In addition to testing the accuracy, the significance of each model and machine learning technique is taken into account. Thus aiming to reach a significance of p value <0.05.

3.6 Class balancing

This part of the thesis outlines the classes and how they are balanced. All variables are well balanced. Figures 2,3,4,5,6 show a variety of classes and the balancing of these variables. All variables consist out of 5565 games as shown in the table. The variables are well balanced.

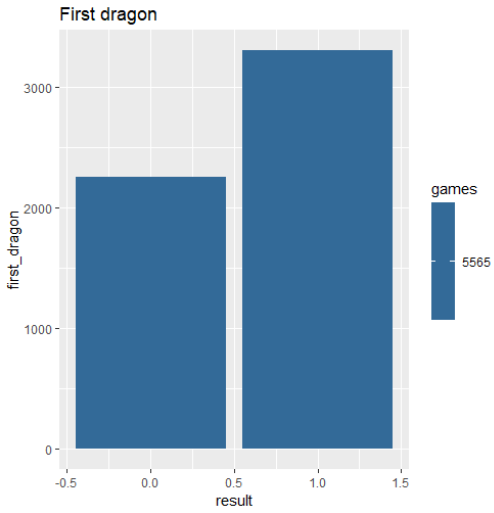

Figure 2. *Balancing of the first dragon variable. X axis shows the result (0 = loss, 1 = win), y axis is the number of times first dragon was taken.*

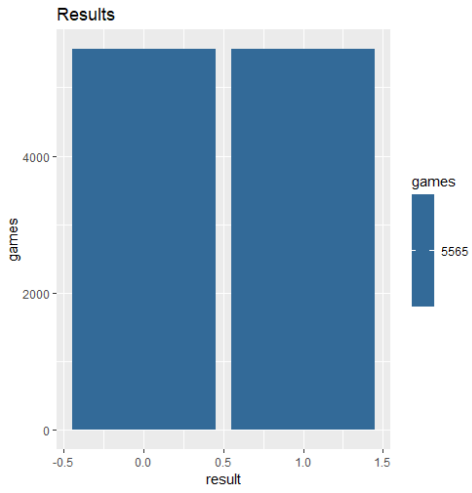

Figure 3. *Balancing of the result variable. X axis shows the result (0 = loss, 1 = win), y axis is the number of times it occurred.*

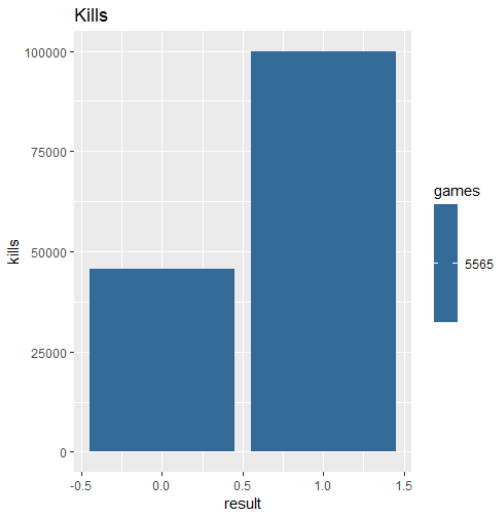

Figure 4. *Balancing of the kill variable. X axis shows the result (0 = loss, 1 = win), y axis is the number of total kills.*

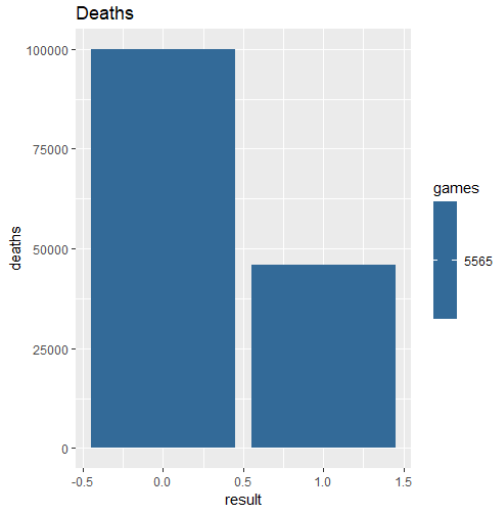

Figure 5. *Balancing of the deaths variable. X axis shows the result (0 = loss, 1 = win), y axis is the number of total deaths.*

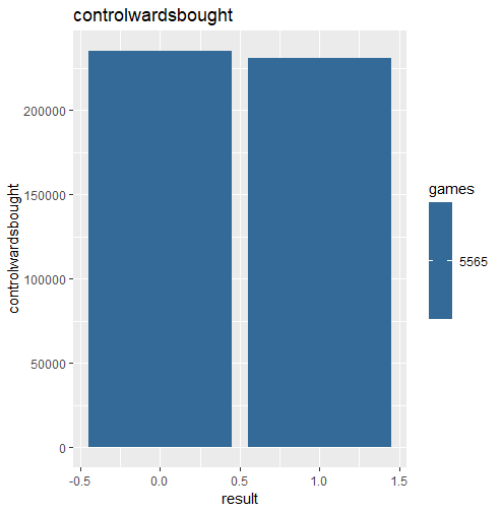

Figure 6. *Balancing of the controlwardsbought variable. X axis shows the result (0 = loss, 1 = win), y axis is the number of total control wards bought.*

4. Experimental setup

This chapter contains the experimental setup. It explains the features, sample size, how the data was collected and all other relevant information of this study.

4.1 Data

The dataset which is used for this study is downloaded from an openly available website[6]. Oracle Elixir collects data from all professional League of Legends games since 2015. The dataset of this study contains data from all professional LoL games from 01-01-2020 till 12-09-2020. The data was collected from professional LoL games out of every region in the world. The dataset has 69080 observations and 103 variables. Each variable is a different in-game statistic. Each observation contains the statistics of a player per game and date. The list of features can be found in appendix A table 14. The table contains the name of the variable and description of the variable. Within the dataset are 246 different teams. Each team falls under the term professional LoL team. A selection of teams has been made. These teams were selected because they qualified for the world championship[7]. Table 1 shows the top 16 teams.

[6] Oracle's Elixir - LoL Esports Stats. (n.d.). Retrieved December 3, 2020, from https://oracleselixir.com/
[7] Riot Games. (n.d.). Lol e-sport. Retrieved December 3, 2020, from https://lolesports.com/news

Table 1. *Top 16 teams based on the qualification for the world championship.*

Top 16 teams			
Top Esports	Fnatic	Suning	JD Gaming
Gen. G	G2 Esports	DAMWON Gaming	DRX
Team Liquid	Machi Esports	PSG Talon	Rogue
LGD Gaming	Team SoloMid	FlyQuest	Unicorns of love.CIS

4.2 Preprocessing

In order to transform the dataset into usable data, the raw dataset was preprocessed. First, the partial complete data was removed. The variable in the raw dataset “datacompleteness” consisted out of two variables: “complete” or “partial”. The “partial” variable was filtered out. The reason for this is that there were important missing statistics, thus making it unusable for this study. Second, the variable “complete” contained some NA values. This issue was solved by creating a function. The function takes the NA values and fills them with the mean of the column. The variables affected by this problem were “minionkills”, “cspm”, “firstbaron” and “firstdragon”.

4.3 Feature engineering

Some of the columns within the dataset are removed. The reason for this is that the data within the columns is not useable for this study. The data is grouped by the variables “date” and “game”. The selected variables can be found in appendix A, table 15. Feature engineering was applied to find the most important statistics of the team performances. Each of the selected variables contain the performances of LoL teams. The variables containing individual statistics were not explored. What was left out of the dataset were variables that contained either information about the enemy team or unusable data, such as the “URL” variable. The “URL” variable did not add anything to the research. The variables “bans” and “champions selected” were left out because this study focusses on player statistics and does not take into account which champions were played or banned.

The features used in this study each have specific in-game purposes. For this reason, the features are separated into the following categories: Player vs Player (table 2), Player vs minions/monsters (table 3), Vision (table 4) and objectives (table 5).

Table 2. *Selected variables which fall under the player vs player category.*

Player vs Player
Kills
Deaths
Assists
Damagetochampions
Firstblood

Table 3. *Selected variables which fall under the player vs minions/monsters category.*

Player vs minions/monsters
Minionskills
Cspm
Monsterkillsownjungle
monsterkillsenemyjungle

Table 4. *Selected variables which fall under the vision category.*

Vision
Visionscore
Wardsplaced
Controlwardsbought
vspm

Table 5. *Selected variables which fall under the gameplay category.*

Gameplay
Gamelength
Team
Teamscore
Result

Table 6. *Selected variables which fall under the objectives category.*

Objectives		
Firstbaron	Firstherald	Firstdragon
Barons	Heralds	Dragons
Infernals	Mountains	Clouds
Oceans	Elders	Firsttower
Firstmidtower	Firsttothreetowers	Towers
Inhibitors		

Each specific purpose is analyzed by checking correlation between the statistic itself and the result. This was done with the summary() and lm() function. These categories were applied to the methods which are discussed in the previous chapter. The result of these scores are discussed in chapter 5.

4.4 Team score

The team score comes forth from all variable means. This was done by summarizing the variables with the mean() function. The calculation of the team score can be divided into three groups. The first group consist of positive influences for a team score (table 7). The positive influence was tested by using the lm() function.

Table 7. *Variables which have a positive influence on team score.*

First group variables			
Kills	Assists	Gamelength	Wardsplaced
Controlwardsbought	Visionscore	Minionkills	Monsterkillsownjungle
Monsterkillsenemyjungle	Cspm	Vspm	Damagetochampions
towers			

The reason that these variables are grouped is because the means can be calculated for these statistics. These variables where divided by the overall mean of that same variable. After that, the means were summed up to the team score.

The second group consists out of the following variable:

- Deaths

This variable has a negative influence for a win. This was also calculated by using the lm() function. The deaths are divided by the overall mean of deaths. While the first group all summed up, the deaths score is deducted from the team score.

Table 8. *All variables in this table are objectives.*

Third group variables			
Inhibitors	Firstblood	Firstmidtower	Firsttower
Firsttothreetowers	Firstbaron	Firstherald	Firstdragon
Barons	Heralds	Infernals	Mountains
Clouds	Oceans	Elders	

Table 8 consist out of objectives. The reason for this is that the overall mean of most objectives is 0.5. This can be explained by the fact that these variables contain either “0” or “1”. So instead of dividing the variable by the overall mean, the scores in this group are just added to the team score. For example: if a team obtained first blood, the team score increases with 1. The objective is either obtained or not. Then, the team score is added as a new column to the dataset, which creates a team score per match and team.

4.5 Win prediction

As discussed in chapter 3, three models were used for win prediction. The models are Random Forest, Support Vector Machine and K-nearest Neighbor. To prepare the data for the classification models, the result variable is transformed from “0” and “1” to “no” and “yes”. For all models the data was divided into a training(0.6) and testing(0.4) set. The models were evaluated on accuracy.

4.6 Software used

The programming for this thesis was done in R studio. R studio is ideal for statistical analyses, data manipulation and data visualization. Table 9 includes the packages that were used for this study.

Table 9. *R studio packages used for this study.*

R packages			
Dplyr	Tidyr	Ggplot2	Caret
RandomForst	Ranger	Ggcorrplot	Tidyverse
Lubridate	E1071		

5. Results

In this chapter the results of this study are discussed. Starting with chapter 5.1 where the first research question is answered, followed by chapter 5.2 for the second research question.

5.1 The most important variables

This chapter evaluates the results of the conducted study. Starting with the first research question “What in-game statistics are the most valuable to win?”. The results of the various applied methods show that all of the variables, except for “deaths” and “controlwardsbought”, have a positive effect on obtaining a win. In addition to the positive effect, all variables except for “minionkills” and “monsterkillsownjungle” have a significant effect ($p < 0.05$). The correlation on the objectives variables revealed that by taking an objective, the chance of winning increases between 30% – 40%. On the other hand, each kill increases the chance of winning with 4%. Thus, making the objectives variables the most important factors in the game. As a result of the promising results, machine learning techniques were trained to try and predict a win based on the given variables. The results of the tests are shown in table 10.

Table 10. *Accuracy results of multiple classification models on 5 categories.*

	Random Forest	Support Vector Machine	k-Nearest Neighbor
Player vs player	0.9607	0.9636	0.7248
Player vs minions/monsters	0.7439	0.7457	0.7186
Vision	0.7287	0.7381	0.6736
Objectives	0.9751	0.9704	0.9674
All	0.9695	0.9935	0.6806

The results illustrate that “player vs player” and “objectives” are the two most important categories for a high accuracy win prediction. The highest scoring result is SVM, which has an accuracy of 99.35% ($p < 0.05$). Table 11, 12 and 13 show the confusion matrixes of SVM, kNN and RF based on all variables. The recall value achieved an even better score (99.46%) than measuring for accuracy. Followed by F1 score (99.32%) and last precision (99.23%).The kNN model scores 82% on F1 score, 71% on recall and 67% on precision. Last, the RF scores 99% on recall, 98.7% on F1 score and 94% on precision.

Table 11. *Confusion matrix of SVM.*

Prediction	No	yes
No	2209	12
yes	17	2214

Table 12. *Confusion matrix of kNN.*

Prediction	No	yes
No	1444	640
yes	782	1586

Table 13. *Confusion matrix of RF.*

Prediction	No	Yes
No	2105	17
Yes	121	2209

SVM even has an enormously high prediction rate of 0.9935 when using all variables as an input. Based on the best predicting model, in this case the SVM, the variables are ranked in the following order ranking from most valuable to least valuable:

1. Objectives
2. Player vs player
3. Player vs minions/monsters
4. Vision

Taking objectives, getting kills, assists and avoiding deaths increase the chance of winning the game. While vision and player vs minions/monsters are less important to obtain a win, they still increase the chance of winning by smalls bits.

5.2 Team score

The team score is created to see how well each team performances per game. The team score is based on various variables as discussed in chapter 4. The results show that it is very hard to create a team score to predict the chance of a team winning a future match. Figure 7 illustrates the mean team score of each team based on all played matches. According to figure 7 Unicorns Of Love should be the best performing team, but at the world championship they did not win a single game[8]. Figure 8 consist out of 4 of the 16 teams. The teams are grouped in clusters of four. The reason for pairing the teams by four is that the visualization will still be interpretable and they can easily be compared. By visualizing them all together instead of smaller clusters, the figure becomes too crowded and therefore making the results unreadable and unclear.

The results show that the team score barely changes over time. As seen in figure 8, all four teams barely increase or decrease over the course of the year. Results of the other 12 teams can be found in appendix A figure 9-11. Figure 12 illustrates the team score of the team Top Esports. The team score varies a lot over the course of the season. The team score of Top Esports is below 10 in March while peeking in June. Similarly is the team score of G2 Esports as illustrated in figure 13. The overall performances of the other teams can be found in appendix A figure 14-27.

When looking at the individual games, the score per match varies a lot. When looking at figure 12, Top Esports has a really low score in early June and a high team score in the end of June. The same counts for G2 Esports (figure 13) and the other teams (appendix A figure 14-27).

[8] Riot Games. (n.d.). Lol e-sport. Retrieved December 3, 2020, from https://lolesports.com/news

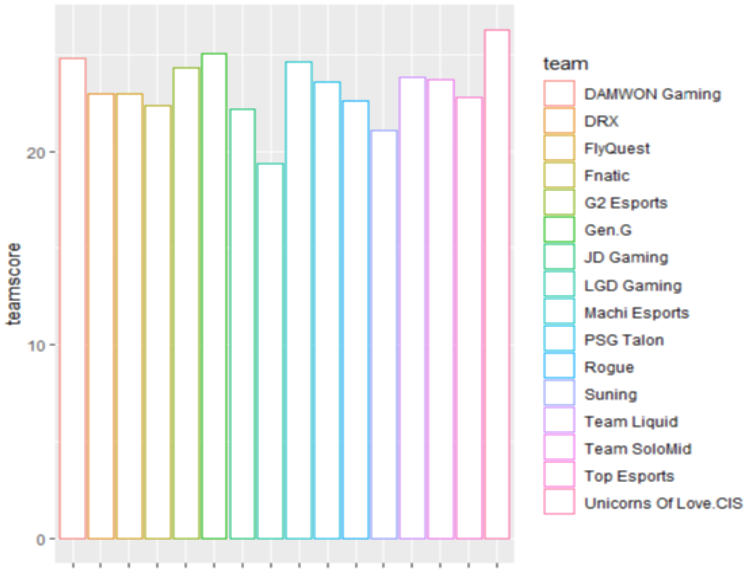

Figure 7. *Mean team score of top 16 professional League of Legends teams. Unicorns Of Love.CIS has the highest team score while LGD Gaming has the lowest.*

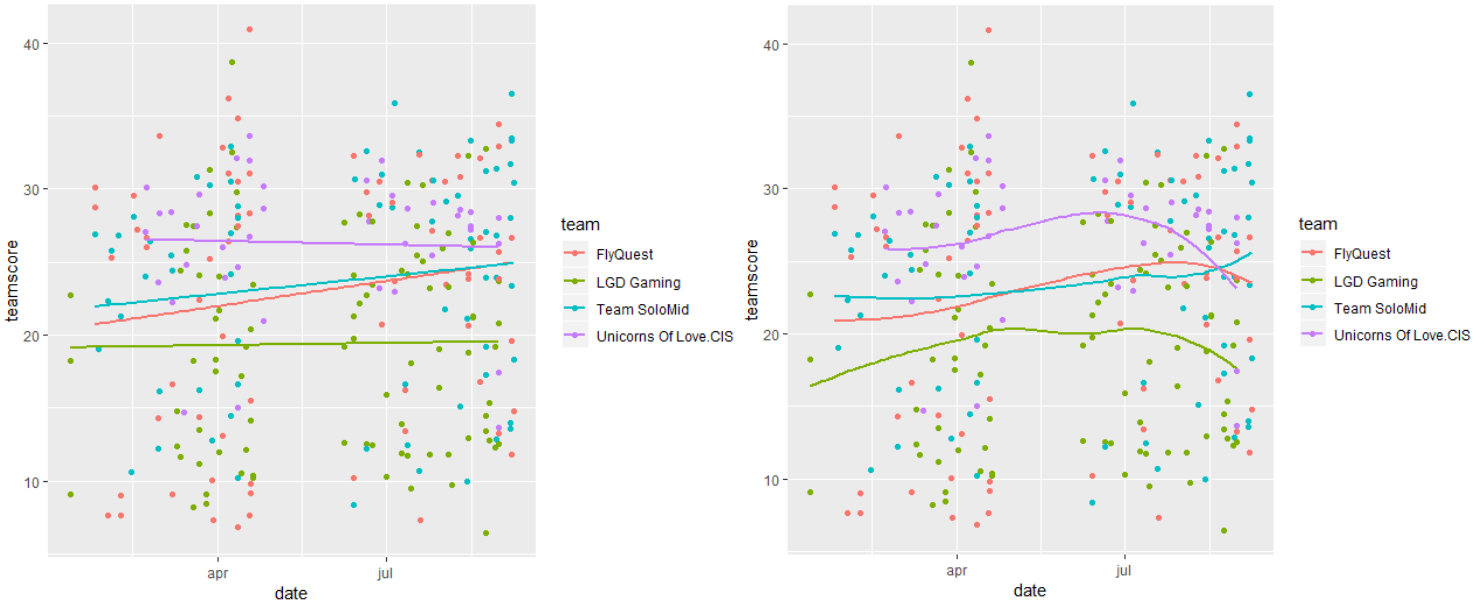

Figure 8. *Plot of lm and loess method. The left figure shows the lm method while the right shows the loess method. The plot shows the team scores of four teams (FlyQuest, LGD Gaming, Team SoloMid, Unicorns Of Love.CIS) for each day they played over the course of the season.*

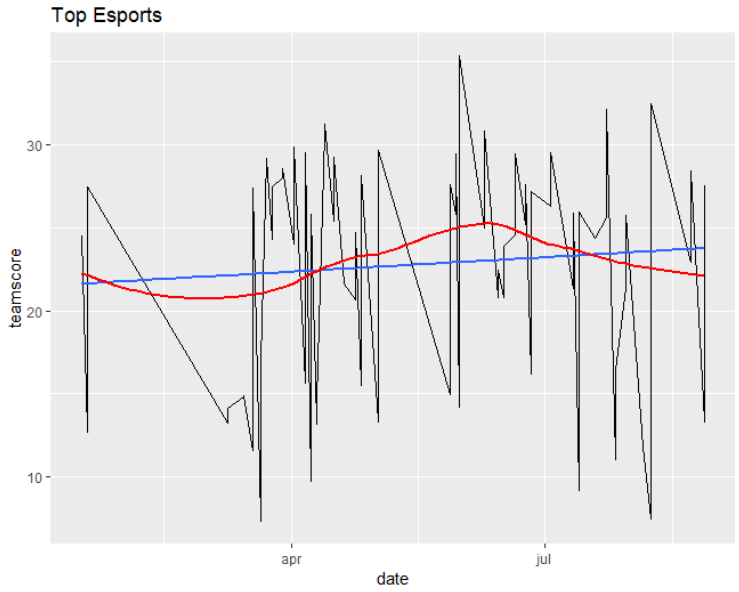

Figure 12. *Plot of all games of Top Esports played over the year with lm(blue line) and loess(red line) methods applied.*

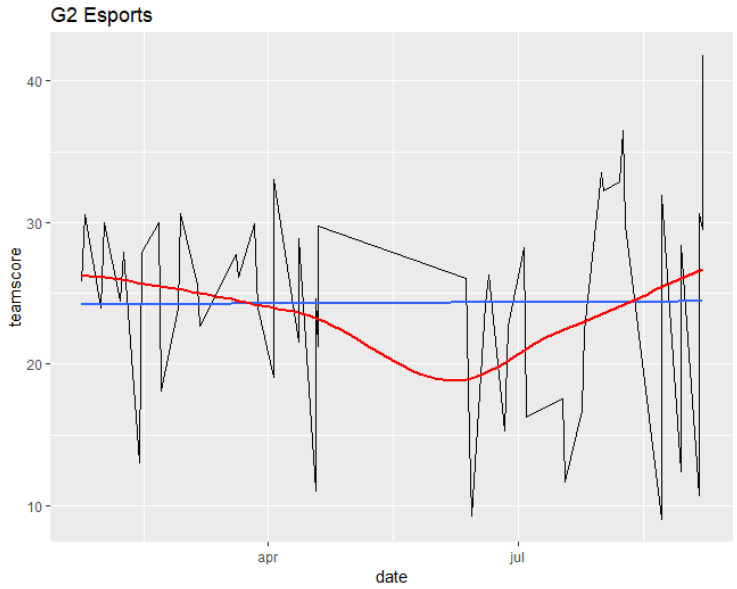

Figure 13. *Plot of all games of G2 Esports played over the year with lm(blue line) and loess(red line) methods applied.*

6. Discussion

In this chapter the results of the conducted study and the study itself are evaluated. This chapter includes the findings, limitations and the possibilities for future research.

6.1 Finding and limitations

This study explores the possibilities of feature engineering and various machine learning algorithms to find the most important in-game statistics. To achieve this goal, the study is separated into two research questions. The first research questions is: “What in-game statistics are the most valuable to win?”. All the variables were controlled on how each variable influences the chance of winning. All variables except for “deaths” and “controlwardsbought” had an positive effect on the outcome. Furthermore, all variables had a significant effect as well ($p < 0.05$), except for “minionkills” and “monsterkillsownjungle”.

While the “kills” variables increased the chance of winning with 4%, some objectives increased the chance between 30-40%. Therefore, proving that taking objectives is a crucial part of winning a match. In addition to these tests the variables were divided into separate groups. Each group indicates the role they have in-game. These categories were then applied as the input of the machine learning techniques; RF, SVM, kNN. SVM scored the highest accuracy (99%) followed by RF (97%) and kNN (68%). Based on the strong positive effects of objectives and high prediction accuracy, the in-game statistics are ordered as followed:

1. Objectives
2. Player vs player
3. Player vs minions/monsters
4. vision

This study does however have some limitations. First, the dynamic and playstyles within the game are not taken into account. Second, the study only explored the team statistics. Individual players might have an important role in overall team performances, this however is not taken into account. Third, the group dynamic is also not taken into account. In fact this study only focusses on team effort. Last, several variables were removed from the dataset. Those variables contained information about selected champions, banned champions, gold gain, gold differences and gained experience.

The second research question is “In what way can a scoring system, based on the in game statistics of team, predict the chance of winning a professional League of Legends match?”. The same in-game statistics were applied to answer this question as were used in the first research question. The reason for this is that the variables proved to be good indicators for win predictions. The team scores were created based on the average score of each variable. Therefore, giving teams a score per day. The team scores of all teams were

compared against each other. While winning teams did in fact have higher team score than losing teams, the score was not useable to predict future matches. To illustrate, the team with the highest team score is Unicorns of Love.CIS. However, their performance in the world championship did not show this at all. Within the championship they lost every match. This indicates that the team score is unreliable. The performances of all team fluctuated from high team scores (30>) to low team scores (<10) per day. In addition, the teams all have ups and down spread out throughout the season as seen in figure 12. Therefore, making the team score invalid as an indicator on how well a team is performing in the future.

A limitation of the study is a good scoring measurement. To clarify, the team score is based on the mean of each variable. While objectives just add extra points, this might not be sufficient. This study does not take into account how important each variable is. Therefore, the team score is not based on how important a variable is but just on the overall mean of each variable. The way the dataset is organized can also be seen as a limitation. The dataset does not include all data for individual players and the role they chose. When objectives are obtained the results are added as a team effort which results in objectives not being assigned to an individual player.

Another limitation is the variable which shows the patch number. As discussed in chapter 2; the rules and therefore the way a game plays, changes each patch. A patch can result in a shift in playstyles. To clarify, the result of a patch can shift an objective based playstyle into a player vs player based playstyle. The last limitation is on the various regions. Even though certain limitations have to be taken into account, this study was still able to provide new insights and relevant scientific knowledge for e-sports and hereby contributed to diminish the research gap. Finally, the results also show that machine learning techniques can be applied in a positive way.

When comparing this study to the various studies on traditional sports, it can be concluded that the variables in e-sport provide much better prediction qualities. In comparison, the highest accuracy in the analyzed sports is 89% while this study scores 99%. The reason for this is assumably on the fact that all e-sport statistics are automatically documented.

6.2 Future research

As discussed in chapter 6.1, there are some limitations in this study. These limitation provide opportunities for future research. First, the scoring system still has challenges to overcome. Currently, the scoring system is based on the mean of each variable. However, future research could explore the various variables and map them on how important they are. Thus, creating a new way to calculate the team scores. Therefore, the scoring system might be used as a future match predictor.

Second, this research could be extended by exploring the played and banned champions. The role of certain champions might provide interesting information or new insights. Future studies could even focus on providing a scoring system for each champion. This could then be compared to the team performances.

7. Conclusion

This chapter contains the conclusion of this thesis.

7.1 Conclusion

In this thesis the dataset of professional League of Legends games was analyzed and used to find the most important in-game statistic and to create a team score. By comparing the correlations with the prediction models, this study concludes that the most valuable in-game statistics are the objectives. The in-game statistics proved to be excellent variables for classification models such as SVM and RF. By using linear models, the correlations between various variables and the results were found. After that the game statistics could be categorized. The categories were then scored based on the highest prediction accuracy.

From the same statistics a team score was created. The study intended to use this team score to predict future match outcomes. This was not successful. The team score does need improvement which can be done in a extending research. The team score can be useful to get a quick overview of how a team played without having to watch a replay or analyze the statistics by hand.

References

Anshori, M., Marri, F., Alauddin, M. W., & Abdurrahman Bachtiar, F. (2018). Prediction Result of Dota 2 Games Using Improved SVM Classifier Based on Particle Swarm Optimization. *2018 International Conference on Sustainable Information Engineering and Technology (SIET)*, 1–10. https://doi.org/10.1109/siet.2018.8693204

Beckmann, M., Ebecken, N. F. F., & Pires de Lima, B. S. L. (2015). A KNN Undersampling Approach for Data Balancing. *Journal of Intelligent Learning Systems and Applications*, *07*(04), 104–116. https://doi.org/10.4236/jilsa.2015.74010

Bunker, R. P., & Thabtah, F. (2019). A machine learning framework for sport result prediction. *Applied Computing and Informatics*, *15*(1), 27–33. https://doi.org/10.1016/j.aci.2017.09.005

Cong, J. (2020, June). *Prediction of game results based on League of Legends*. Retrieved from http://courses.cecs.anu.edu.au/courses/CSPROJECTS/20S1/reports/u6502494_report.pdf

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: a methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/s1532-0464(03)00034-0

Driskell, J. E., & Salas, E. (1992). Collective Behavior and Team Performance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*(3), 277–288. https://doi.org/10.1177/001872089203400303

Gibson, A. (2020, August 17). Most Played Games in 2020, Ranked by Peak Concurrent Players. Retrieved December 3, 2020, from https://twinfinite.net/2020/08/most-played-games-in-2020-ranked-by-peak-concurrent-players/

Hodge, V., Devlin, S., Sephton, N., Block, F., Cowling, P., & Drachen, A. (2020a). Win Prediction in Multi-Player Esports: Live Professional Match Prediction. *IEEE Transactions on Games*, 1. https://doi.org/10.1109/tg.2019.2948469

Hodge, V., Devlin, S., Sephton, N., Block, F., Cowling, P., & Drachen, A. (2020b). Win Prediction in Multi-Player Esports: Live Professional Match Prediction. *IEEE Transactions on Games*, 1. https://doi.org/10.1109/tg.2019.2948469

Hsu, Y.-C. (2020). Using Machine Learning and Candlestick Patterns to Predict the Outcomes of American Football Games. *Applied Sciences*, *10*(13), 4484. https://doi.org/10.3390/app10134484

Jones, K. (2019a, September 3). How the eSports Industry Fares Against Traditional Sports. Retrieved December 3, 2020, from https://www.visualcapitalist.com/how-the-esports-industry-fares-against-traditional-sports/

Jones, K. (2019b, September 3). How the eSports Industry Fares Against Traditional Sports. Retrieved December 3, 2020, from https://www.visualcapitalist.com/how-the-esports-industry-fares-against-traditional-sports/

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review*, *9*(4), 829–835. https://doi.org/10.3758/bf03196342

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567. https://doi.org/10.1038/nbt1206-1565

Novak, A., Bennett, K., Pluss, M., & Fransen, J. (2019). Performance analysis in esports: part 1 - the validity and reliability of match statistics and notational analysis in League of Legends. *PREPRINT –NOT PEER REVIEWEDCorresponding Author:Andrew R. Novak -Andrew.Novak@uts.Edu.AuAuthor Agreement Statement:We the Authors Agreeto the Sharing of This Preprint on SportRxivTwitter Handles:@NovakSportSci / @KyleJMBennett / @PlussMatt / @JobFranPerformance Analysis in Esports: Part 1 - the Validity and Reliability of Match Statisticsand Notational Analysis in League of Legends*, 5–8. https://doi.org/10.31236/osf.io/sm3nj

Oracle's Elixir - LoL Esports Stats. (n.d.). Retrieved December 3, 2020, from https://oracleselixir.com/

Oshiro, T. M., Perez, P. S., & Baranauskas, J. A. (2012). How Many Trees in a Random Forest? *Machine Learning and Data Mining in Pattern Recognition*, 154–168. https://doi.org/10.1007/978-3-642-31537-4_13

Passi, K., & Pandey, N. (2018). INCREASED PREDICTION ACCURACY IN THE GAME OF CRICKET USING MACHINE LEARNING. *International Journal of Data Mining & Knowledge Management Process*, *8*, 32–34. Retrieved from https://arxiv.org/ftp/arxiv/papers/1804/1804.04226.pdf

Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. (2017). Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL). *IOP Conference Series: Materials Science and Engineering*, *226*, 012099. https://doi.org/10.1088/1757-899x/226/1/012099

Tax, N., & Joustra, Y. (2015, September). *Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach*. Research gate. https://doi.org/10.13140/RG.2.1.1383.4729

Wagner, M. G. (2006). On the Scientific Relevance of eSports. *Research Gate*, *1*, 1–3. Retrieved from https://www.researchgate.net/publication/220968200_On_the_Scientific_Relevance_of_eSports

Yang, Y., Qin, T., & Lei, Y. H. (2016, December). *Real-time eSports Match Result Prediction*. Retrieved from https://arxiv.org/pdf/1701.03162.pdf

Appendix A

Table 14. *Included all variables of the dataset with short explanation.*

GameID	ID of the game to track it back
Datacompleteness	Indicates if data is complete
Url	URL to find the matchhistory
League	The league of which the match was played in
Year	Year of which the match was played
split	Contains opening or closing
Playoffs	Contains a 1 or 0
Date	The data of which the game is played
Game	The game number
Patch	Patch number
playerID	The ID of the player
Side	The side the player played on
Position	The position of the player
Player	In-game name of the player
Team	Team name of the player
Champion	Selected champion
Ban1	The first ban of the team
Ban2	The second ban
Ban3	The third ban
Ban4	The fourth ban
Ban5	The fifth ban
Gamelength	The length of the game. Four digits, for example: 1535 stand for 15 minutes and 35 seconds
Result	Result of the match 1 = win, 0 = loss
Kills	The amount of kills the player achieved
Deaths	The amount of times the player died
Assists	The amount of assists the player achieved
Teamkills	Total number of kill of the team
Team deaths	Total number of death of the team
Double kills	Contains 0 or 1. 1 if player got two kills in a short period. 0 if not achieved in the match

Triple kills	Contains 0 or 1. 1 if player got three kills in a short period. 0 if not achieved in the match
Quadra kills	Contains 0 or 1. 1 if player got four kills in a short period. 0 if not achieved in the match
Penta kill	Contains 0 or 1. 1 if player got five kills in a short period. 0 if not achieved in the match
First blood	Contains 0 or 1. 1 if player helped with the first kill.
First blood kill	Contains 0 or 1. 1 if player got the first kill of the game.
First blood assist	Contains 0 or 1. 1 if player assisted with first blood
First blood victim	Contains 0 or 1. 1 if player was the first one to die
Team kpm	Kpm stand for kills per minute.
Ckpm	Combined kills per minute
First dragon	Contains 0 or 1. 1 if team killed the first dragon. 0 if not.
Dragons	Contains the amount of dragons killed
Opp_dragons	Contains the amount of dragons killed by the enemy
Elemental drakes	Contains the amount of dragons killed
Opp_elementaldrakes	Contains the amount of dragons killed by the enemy
Infernals	The amount of infernals dragons
Mountains	The amount of mountain dragons
Clouds	The amount of cloud dragons
Oceans	The amount of ocean dragons
Dragons (type unknown)	The amount of dragons where the class is not known
Elders	The amount of elder dragons
Opp_elders	The amount of elder dragons by the enemy
First herald	1 if team got the first herald. 0 if niet
Heralds	The amount of heralds killed
Opp_heralds	The amount of heralds killed by the enemy
First tower	1 if team got first tower. 0 if not

Opp_towers	1 if enemy got first tower. 0 if not
Firstmidtower	1 if team destroyed the mid tower as first. 0 if not
Firsttothreetowers	1 if team destroyed three towers as first. 0 if not
Inhibitors	Amount of inhibitors destroyed
Opp_inhibitors	Amount of inhibitors destroyed by the enemy
Damagetetochampion	The amount of damage afflicted to other champions
Dpm	Damage per minute.
Damage share	Amount of damage shared with other players
Wardsplaced	The amount of wards placed
Wpm	Wards per minute
Wardskilled	The amount of wards killed
Wcpm	Wards killed per minute
Controlwardsbought	The amount of controlwards the player bought
Visionscore	The score of which a player placed and kill wards
Vspm	Visionscore per minute
Total gold	Total gold of the player at the end of the match
Earned gold	Gold earned by the played
Earned gpm	Earned gold per minute
Earnedgoldshare	Earned gold shared with other players
Goldspent	The amount of gold spent
Gspd	Gold spent per death
Total cs	The total of creeps slain.
Minionkills	Total of minions killed
Monsterkills	Total of monsters killed
Monsterkillsownjungle	Total of monsters killed in own side of the jungle
Monsterkillsenemyjungle	Total of monsters killed in enemy side of the jungle
Cspm	Creeps slain per minute
Goldat10	The amount of gold at 10 minutes

Xpat10	Amount of experience at 10 minutes
Opp_goldat10	Amount of gold of the enemy in the same role at 10 minutes
Opp_xpat10	Amount of experience of the enemy in the same role at 10 minutes
Opp_csat10	Amount of cs of the enemy in the same role at 10 minutes
Golddiffat10	Difference in gold between the player and the enemy in the same role at 10 minutes
Xpdiffat10	Difference in experience between the player and the enemy in the same role at 10 minutes
Csdiffat10	Difference in creeps slain between the player and the enemy in the same role at 10 minutes
Goldat15	Gold at 15 minutes
Xpat15	Experience at 15 minutes
Csat15	Creep slain at 15 minutes
Opp_goldat15	Gold of enemy in same role at 15 minutes
Opp_xpat15	Experience of enemy in same role at 15 minutes
Opp_csat15	Creeps slain of enemy in same role at 15 minutes
Golddiffat15	Difference in gold between the player and the enemy in the same role at 15 minutes
Xpdiffat15	Difference in experience between the player and the enemy in the same role at 15 minutes
Csdiffat15	Difference in creeps slain between the player and the enemy in the same role at 15 minutes

Table 15. *All selected variables used for predictions*

Date
Game
Team
Result
Gamelength
Kills
Deaths
Assists
Wardsplaced
Controlwardsbought
Visionscore
Minionkills
Monsterkillsownjungle
Monsterkillsenemyjungle
Cspm
Vspm
Damagetochampions
Inhibitors
Firstblood
Firstmidtower
Firsttower
Firsttothreetowers
Firstbaron
Firstherald
Firstdragon
Towers
Barons
Heralds
Dragons
Infernals
Mountains
Clouds
Oceans
elders

Team score linear models

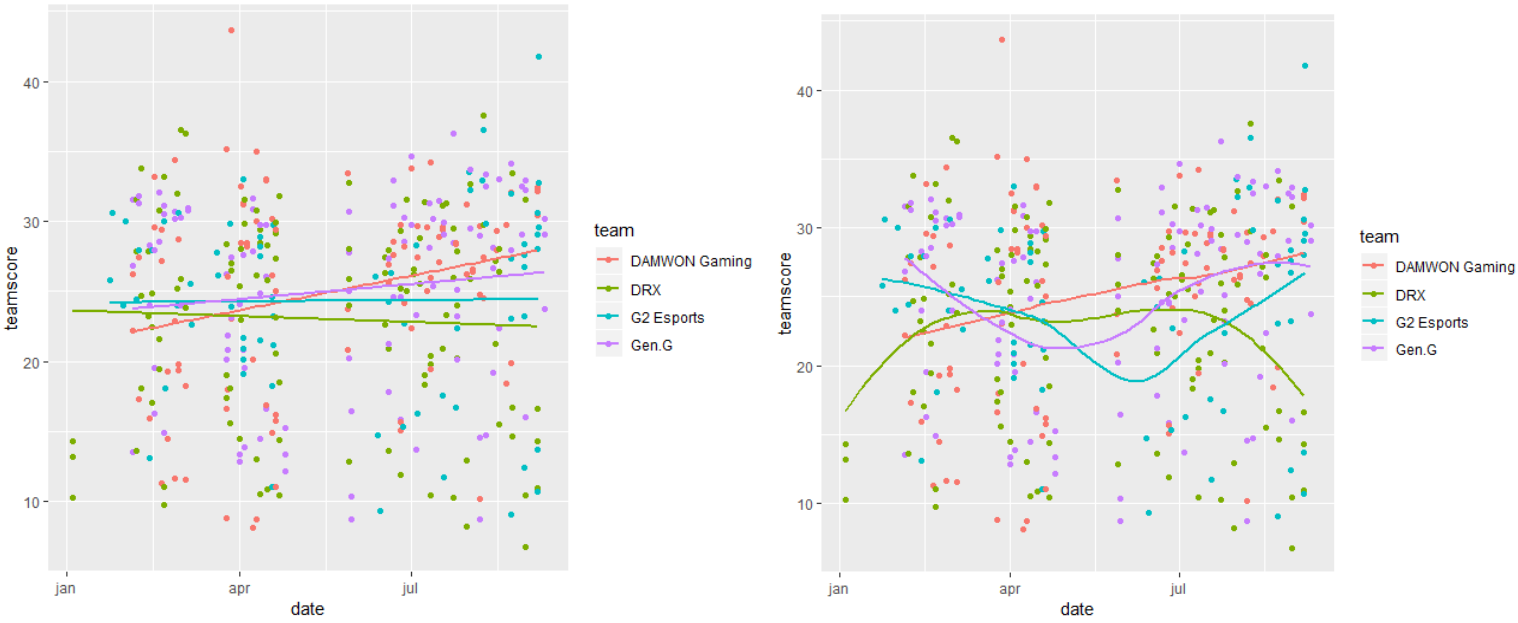

Figure 9. *Plot of lm and loess method. The left figure shows the lm method while the right shows the loess method. The plot shows the team scores of 4 teams (DAMWON Gaming, DRX, G2 Esports, Gen.G) per day they played over the course of the season.*

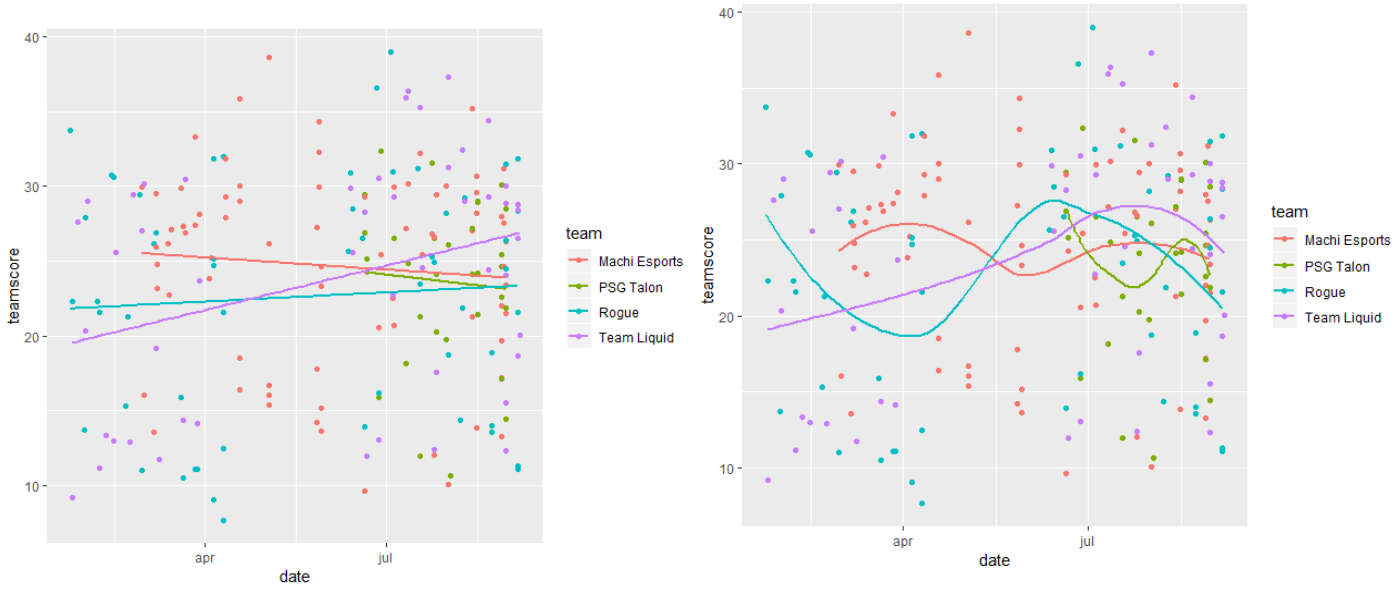

Figure 10. *Plot of lm and loess method. The left figure shows the lm method while the right shows the loess method. The plot shows the team scores of 4 teams (Machi Esports, PSG Talon, Rogue, Team Liquid) per day they played over the course of the season.*

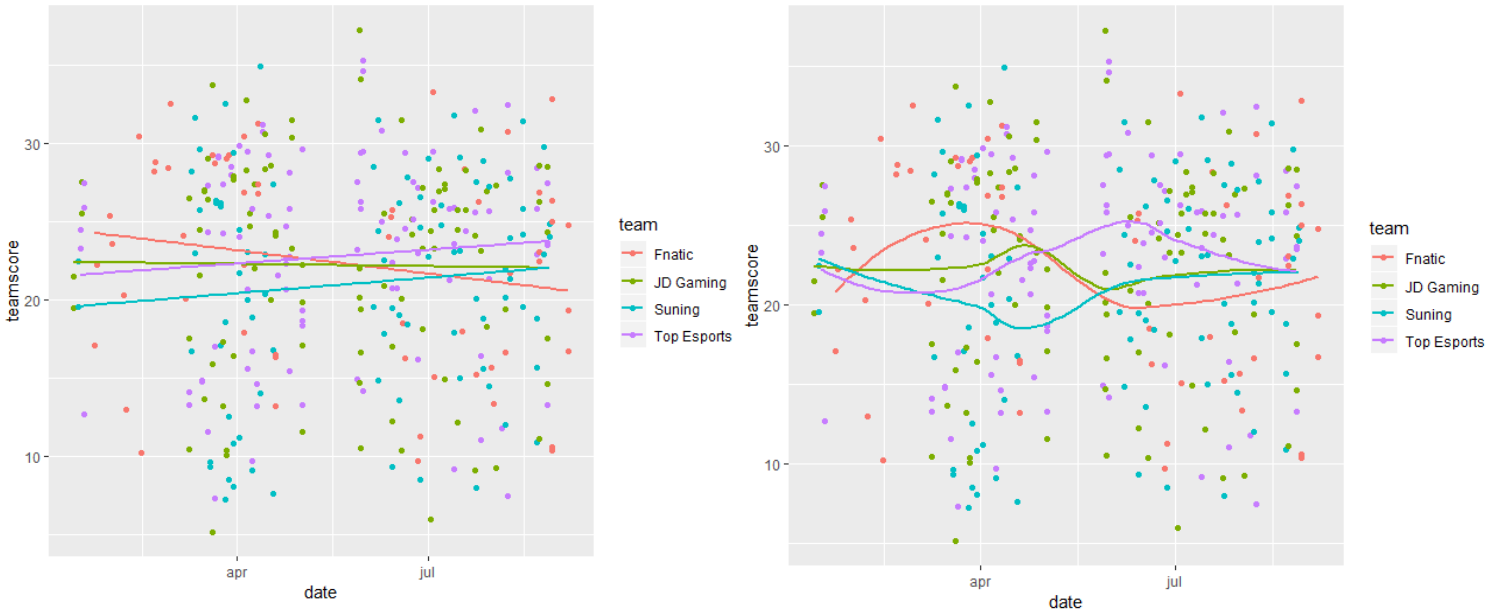

Figure 11. *Plot of lm and loess method. The left figure shows the lm method while the right shows the loess method. The plot shows the team scores of 4 teams (Fnatic, JD Gaming, Suning, Top Esports) per day they played over the course of the season.*

Team scores individual

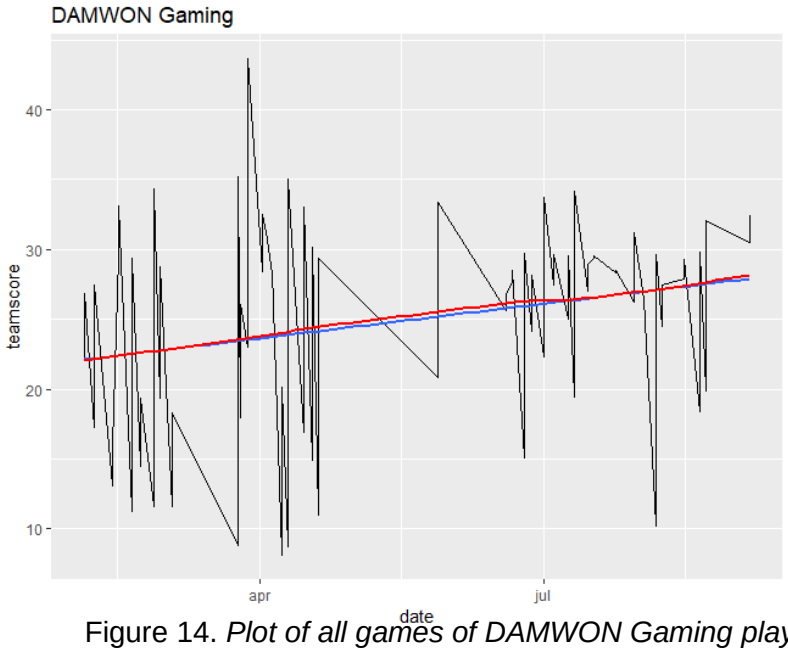

Figure 14. *Plot of all games of DAMWON Gaming played over the year with lm(blue line) and loess(red line) methods applied.*

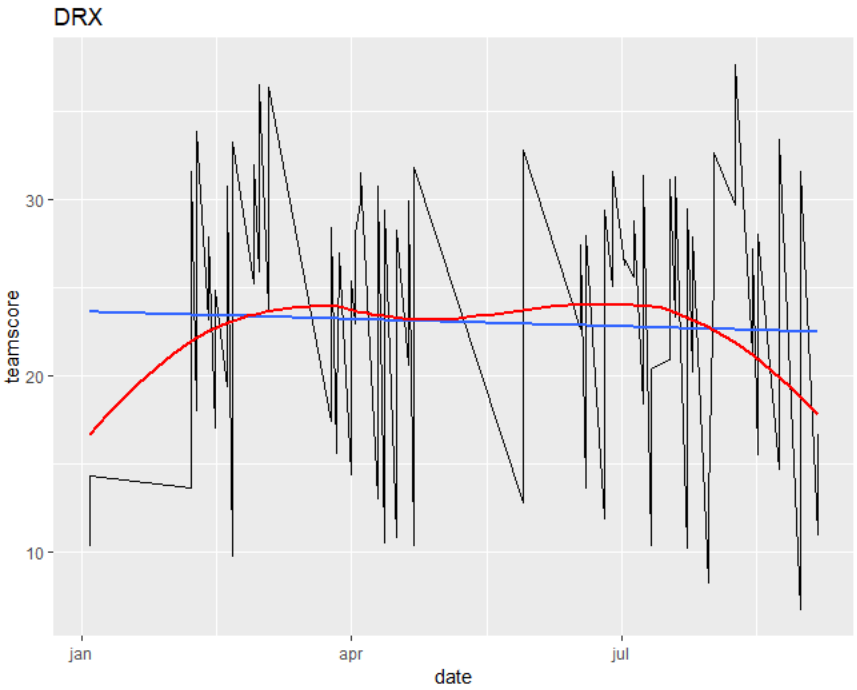

Figure 15. *Plot of all games of DRX played over the year with lm(blue line) and loess(red line) methods applied.*

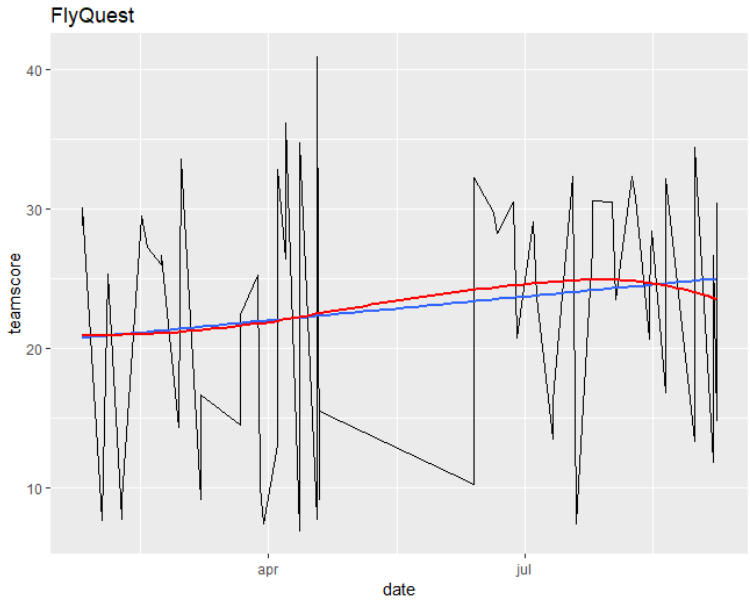

Figure 16. *Plot of all games of FlyQuest played over the year with lm(blue line) and loess(red line) methods applied.*

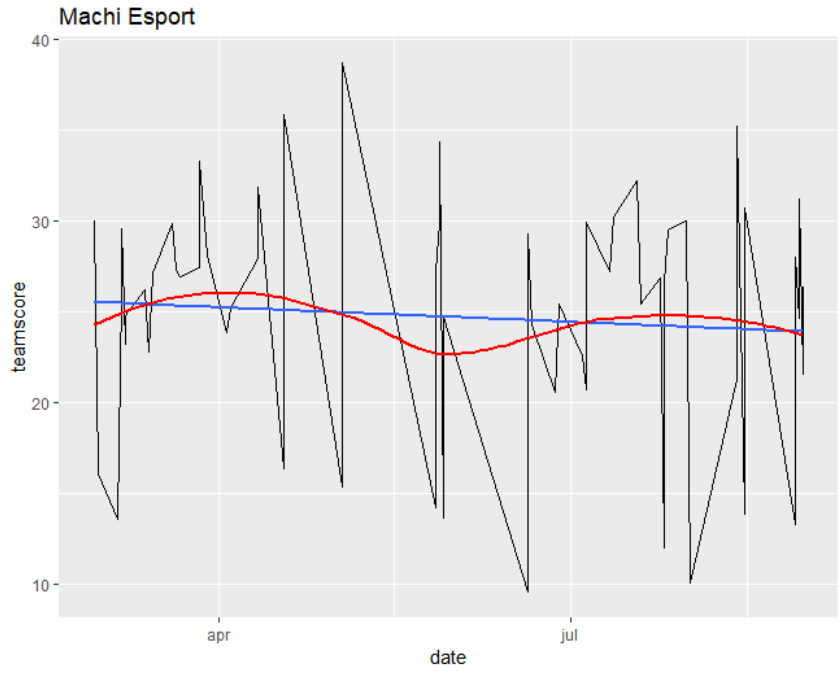

Figure 17. *Plot of all games of Machi Esport played over the year with lm(blue line) and loess(red line) methods applied.*

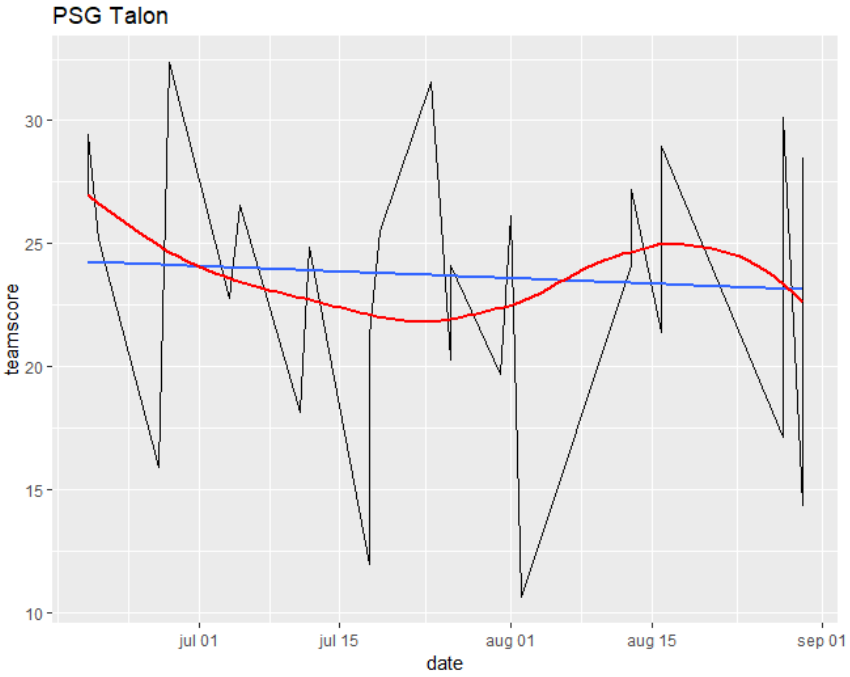

Figure 18. *Plot of all games of PSG Talon played over the year with lm(blue line) and loess(red line) methods applied.*

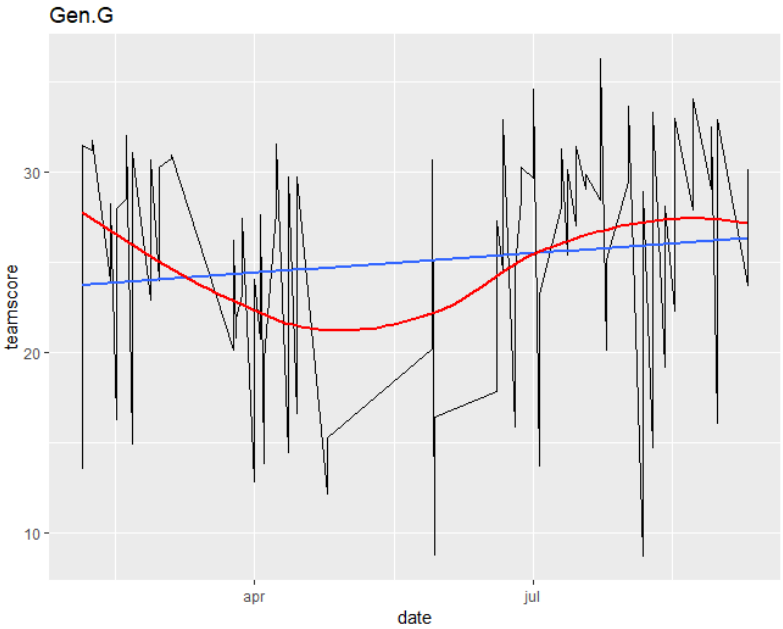

Figure 19 *Plot of all games of Gen.G played over the year with lm(blue line) and loess(red line) methods applied.*

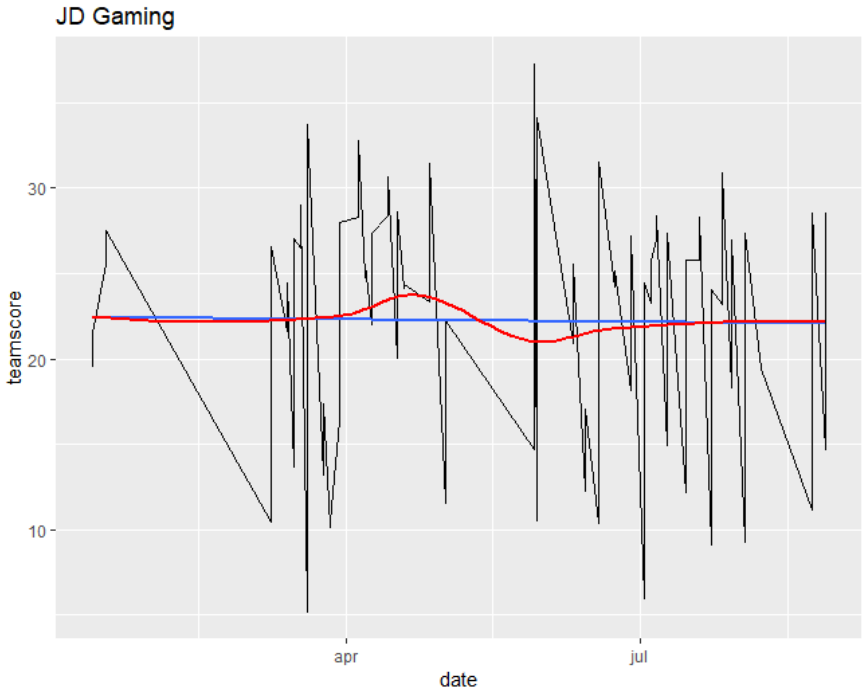

Figure 20. *Plot of all games of JD Gaming played over the year with lm(blue line) and loess(red line) methods applied.*

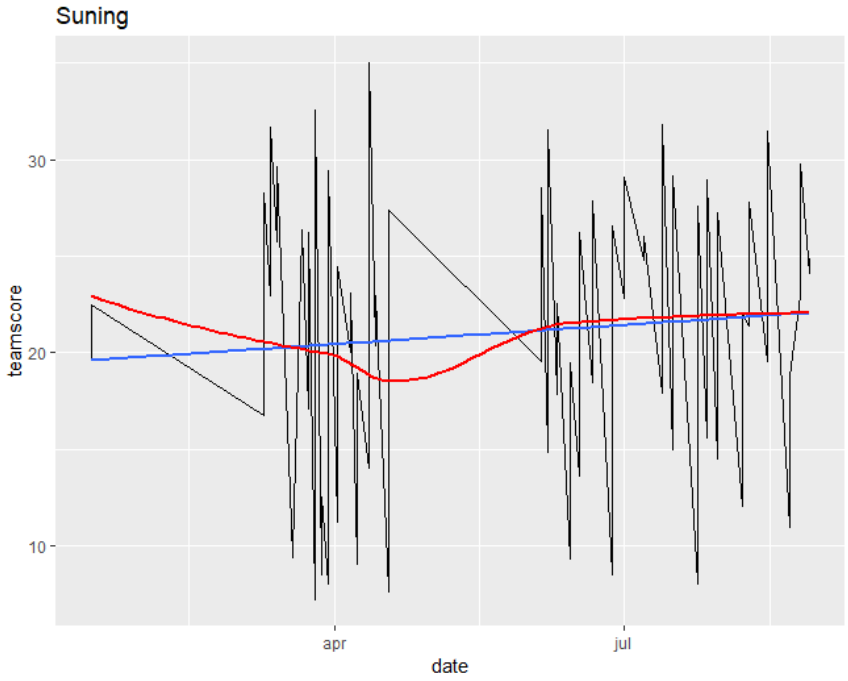

Figure 22. *Plot of all games of Suning played over the year with lm(blue line) and loess(red line) methods applied.*

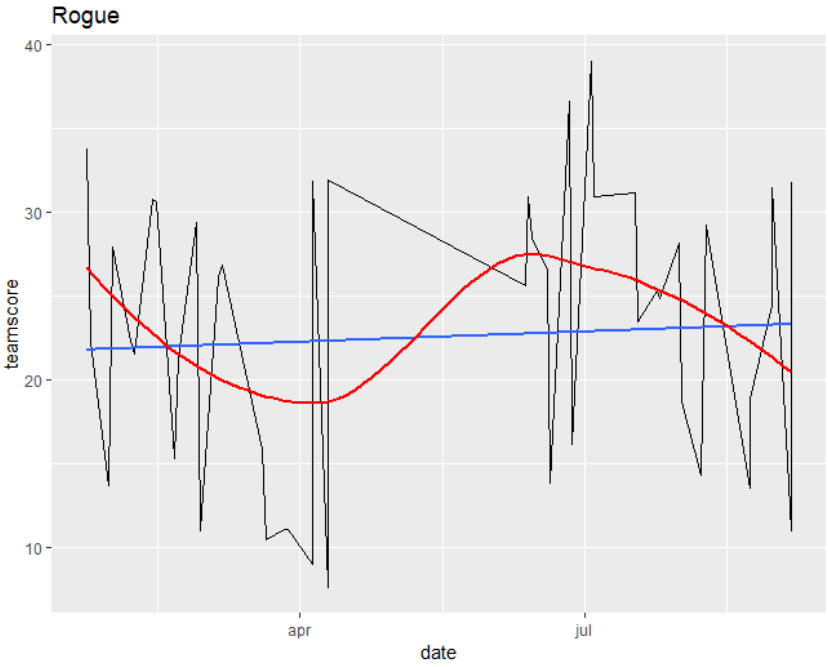

Figure 21. *Plot of all games of Rogue played over the year with lm(blue line) and loess(red line) methods applied.*

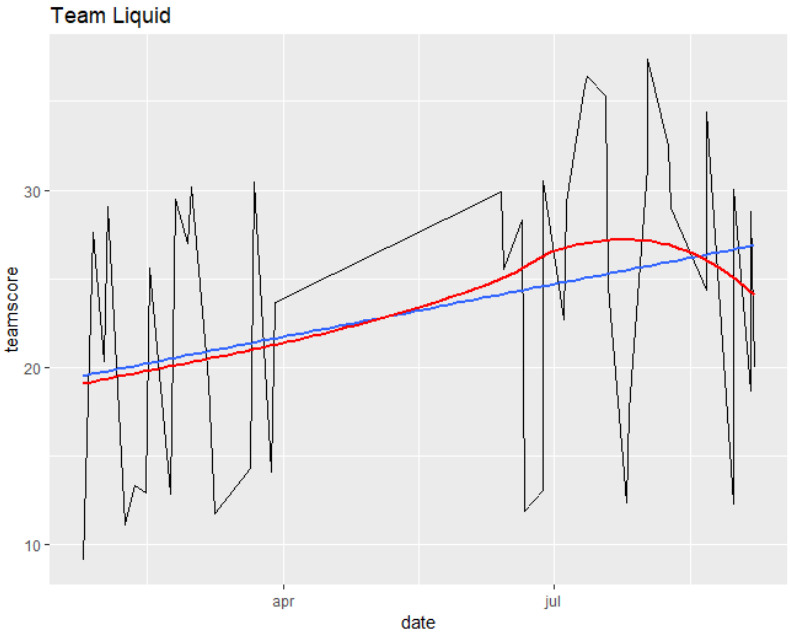

Figure 23. *Plot of all games of Team Liquid played over the year with lm(blue line) and loess(red line) methods applied.*

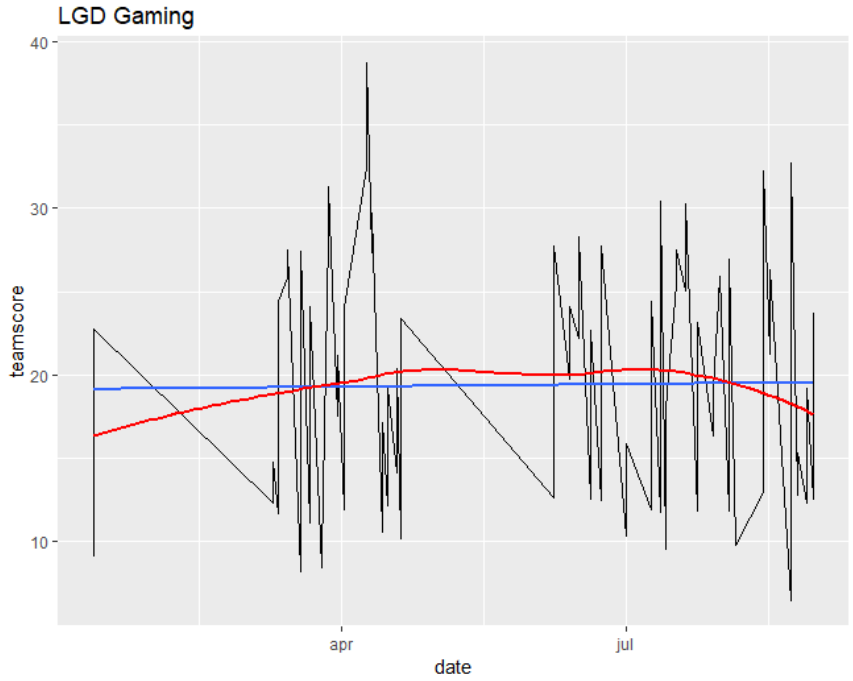

Figure 24. *Plot of all games of LDG Gaming played over the year with lm(blue line) and loess(red line) methods applied.*

Figure 26. *Plot of all games of Team SoloMid played over the year with lm(blue line) and loess(red line) methods applied.*

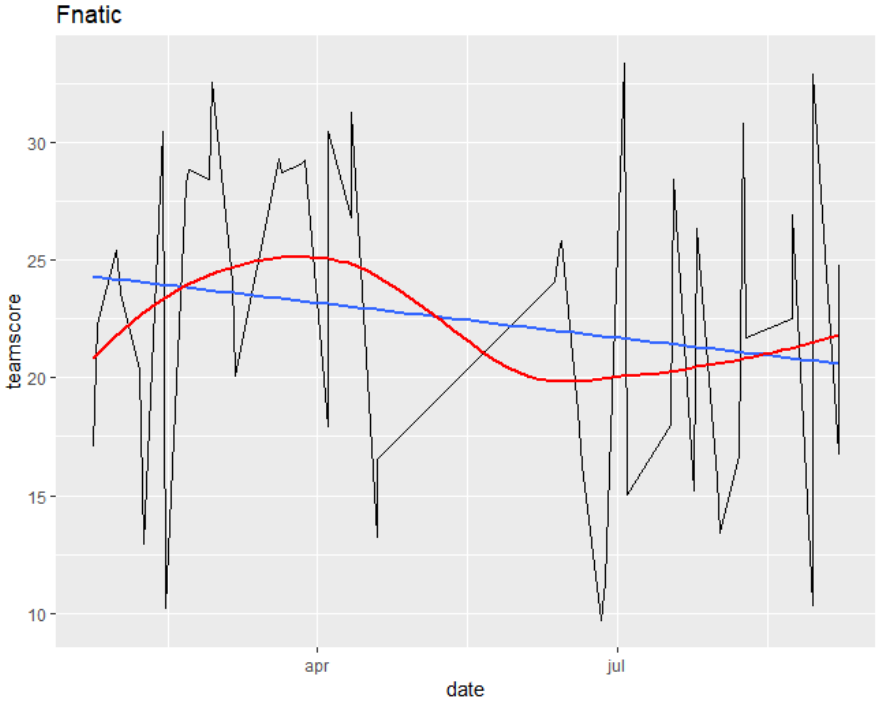

Figure 25. *Plot of all games of Fnatic played over the year with lm(blue line) and loess(red line) methods applied.*

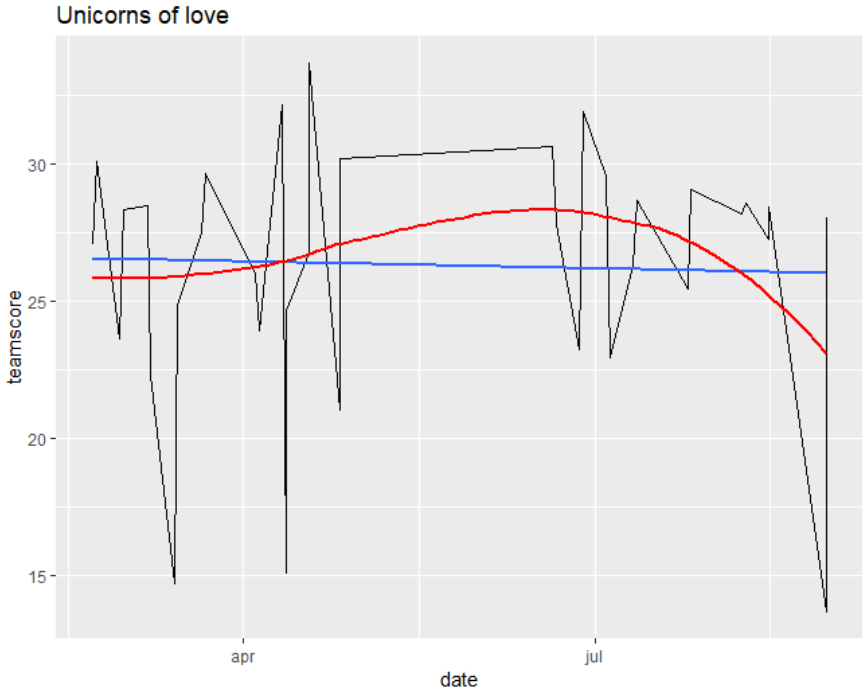

Figure 27. *Plot of all games of G2 Unicorns of Love played over the year with lm(blue line) and loess(red line) methods applied.*