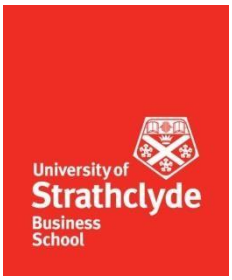


Assignment of

CS989 Big Data Fundamentals

A series of data analysis of League of Legends

Xuyang Yan

201889359

Abstract

Since recent a couple of decades year E-sport industry (electronic sports) grow up rapidly. E-sports is the use of electronic devices (like computers, game consoles, arcades, mobile phones) as the base of E-sports equipment, but the core of E-sports emphasizes on the intelligence and reaction among E-sport competitors. As an outstanding multiplayer online arena battle video game, LOL(League of Legend), especially for a competitive game, It is worth considering whether each element in games settings are roughly balanced and fair. The data from the game significant influence the balance in updated patch. whether could you control the resources in the game determines the trend of the game. The dataset in the paper provide a set of detailed information from more than 50000 matches as samples. We use statistical quantitative method, logistics regression and K-means to fit the model, which is from the unsupervised and supervised method respectively to make a win rate prediction and game elemental evaluation, it aims at supporting the game users improve their technique from dataset analysis and provide suggestion to modify elements in the game.

word count: 3134

1. Background

1.1 Introduction of research subject: League of Legend

LOL (League of Legend) is a popular multiplayer online battle game which was published by Riot Games Company. The highest number of simultaneous online user has more than 7.5 million in the world. In League of Legends, player play the role of an unseen “summoner” to control a “champion” with unique abilities. It has 10 “summoner” in one game, it has 5 persons in each team separately against each other. Teams take turns to select champions and banning the opposing team from selecting certain champions and every champion is unique in individual games. There are lots of resources need to prey on, like demolish tower and inhibitor, slay enemy champions, baron dragon, and destroy opposing team’s nexus to get final winning. All these resources will be reflected by “whole team revenue” directly. Assume that every “summoner” in the game are trying to make the best decision, therefore, in order to win the match in Legend of Legend is just like making optimization decision. The algorithm helps analyze the online game playing data, get insight about the grouping or clusters of players, and offer suggestions to new

players of the game (Braun, 2017). The game data are for supporting player's better decision making.

1.2 Introduction of dataset

Experience in League of Legend is competitive and confrontational, in competitive games it is essential that all characters are "balanced". "Balanced" play an important role when game companies design it, which is to prevent some game components too powerful cause players' game experience are not desirable enough. (Newheiser, 2009) so that is the reason why we need to focus on the game data to make game 'balanced'. This dataset in this paper contains the series of data about game indictor from 51490 matches to do some initial investigation. There are 51490 rows across 60 columns which contain various integer data.

The dataset provides a well-structured frame, which is divide into two parts, one for match information, the other one includes a series of "champion". It is convenient to connect these two parts by json (JavaScript Object Notation), which is helpful for analyzing clarity data.

.

1.3 Research question

The investigation is aimed at seeing if the data can help League of Legend's developers identify trends in how their games are played. It also concerns about which choice is appropriate to the player and their relationship. On basic principle of choosing the most directly indicator, 'whole team revenue' throughout the game from beginning to the end. So, we just choice the "First demolished Tower", "First Blood", "First Baron", "First Inhibitor", "First Dragon", "First Rift herald" as priority indicator for analysis. Game chase for the win, so win rate is the dependence variable in the investigation. Thus, the specific research question is "how these indicators affect the result? What's the relationship between these indicators?"

1.4 Challenge and problem

The most difficult part of data analysis is the unsupervised methods and the data type. Even the data type in this data all are integer, while these integer values doesn't represent an exact numerical value. For example, in "winner" and "first Blood" tuples, it only has 2 or 3 unique values in each tuple, the value '2' is point at the red team and value '1' is for the blue team. I am confused with how to apply it into unsupervised method like K-Means. I search plenty of articles about Machine Learning to find out how it works. Additionally, there are lots of different function in diverse package, and different packages have their own purpose. The format of setting parameter is completely different. Not only just in unsupervised or supervised method, but also in plotting is important to distinguish these setting.

2. Data analysis

2.1 Summary of data

According to the shape, the dataset consists of common numbers among the

51490 samples size and 61 tuples. From "game Duration" to "t2_ban5", there are variety of independent variables and "winner" is the only one dependent variables. thus, the coefficient correlation can be easily to find out, it is convenient to use the logistic regression.

Shape: (51490, 61)

	count	mean	std	min	25%	50%	75%	max
winner								
1	26077.0	1.381524	0.508752	0.0	1.0	1.0	2.0	2.0
2	25413.0	1.563412	0.515960	0.0	1.0	2.0	2.0	2.0

	gameDuration	seasonId	winner	firstBlood
count	51490	51490	51490	51490
max	4562	9	2	2

	firstTower	firstInhibitor	firstBaron	firstDragon	……
count	51490	51490	51490	51490	……
max	2	2	2	2	……

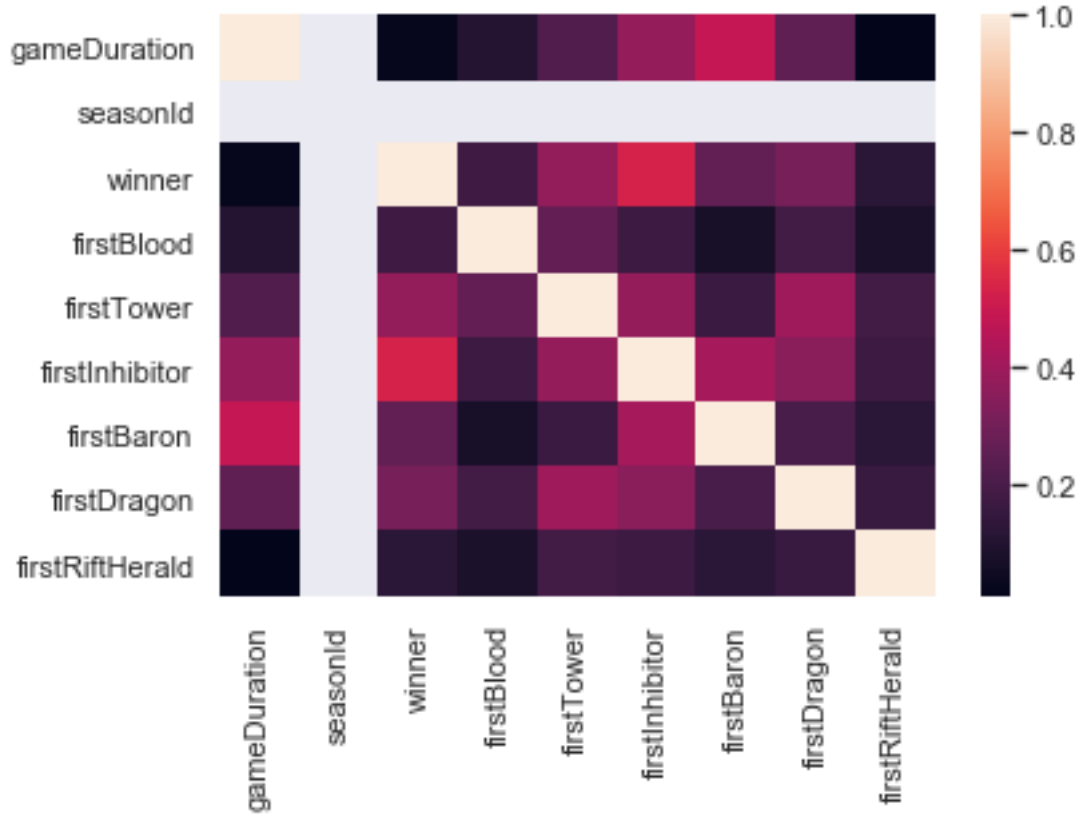
	t2_towerKills	t2_inhibitorKills	t2_baronKills	t2_dragonKills
count	51490	51490	51490	51490
max	11	9	4	6

	t2_riftHeraldKills	t2_ban1	t2_ban2	t2_ban3	t2_ban4	t2_ban5
count	51490	51490	51490	51490	51490	51490
max	1	516	516	516	516	516

From initial description, we can find something interesting. The maximum game duration last for 76minutes (4562 seconds), additionally, inhibitors were killed reach to 9 times, baron were maximum killed 6 times in a game.

Here is the heatmap for us to see the initial comparable coefficient correlation:

Figure 1: the coefficient correlation of between win rate and other indicators



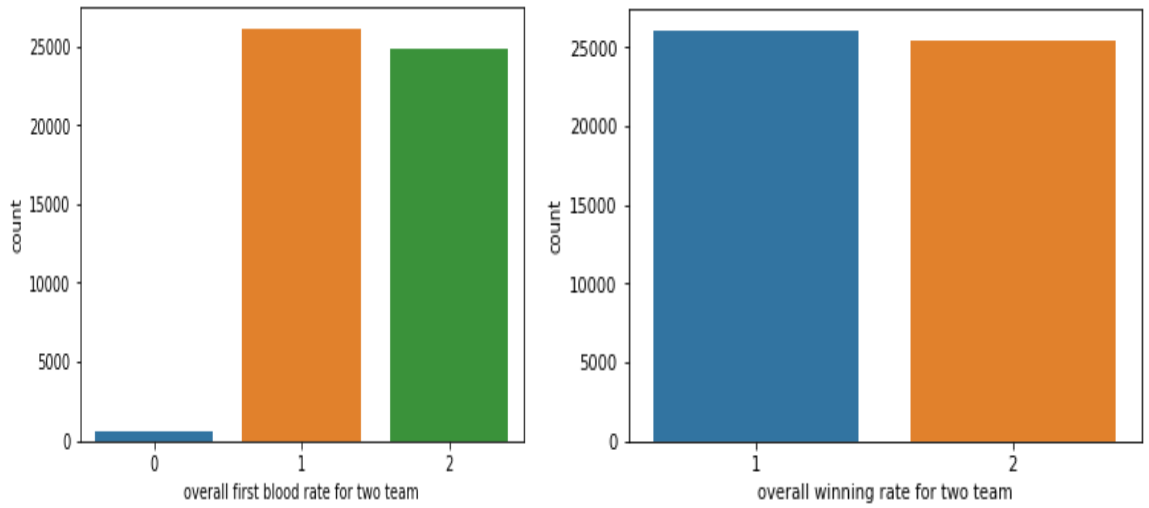
Source: Calculated from data sample

Figure 1 presents the coefficient correlation between different variable, while just like we said in beginning of summary of data (2.1), we only focus on the winner with rest of indicator because games result is whether you can win finally. Obviously, winner has a strong positive correlation with which team can get first inhibitor and first tower, but the team who can acquire first rift herald or first blood are less significant than previous two indicator to help the team get victory.

2.2 Factor analysis for different two team

In the game map of Legend of Legend, it's roughly symmetric. While it still has some different detail, which can slightly affect the game result. I will analysis it subsequently.

Figure 2.1: win rate and first blood rate for two team

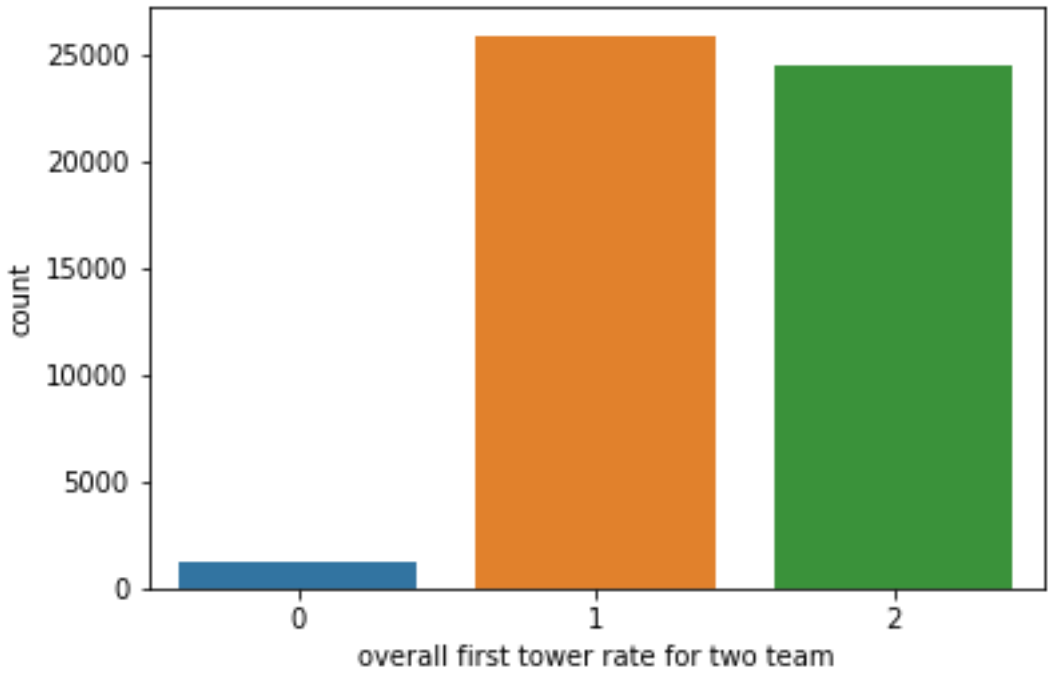


Source: Calculated from data sample

Figure 2.1 on upper left is the overall win rate from two team. ‘1’ stand for blue team. ”2” stand for red team. The blue team has higher win rate than red team.

The figure on upper right is the overall rate from two team who got first blood kills. “0” stand for neither of team got first blood kills. “1” stand for blue team. ”2” stand for red team. The blue team also has higher win rate than red team.

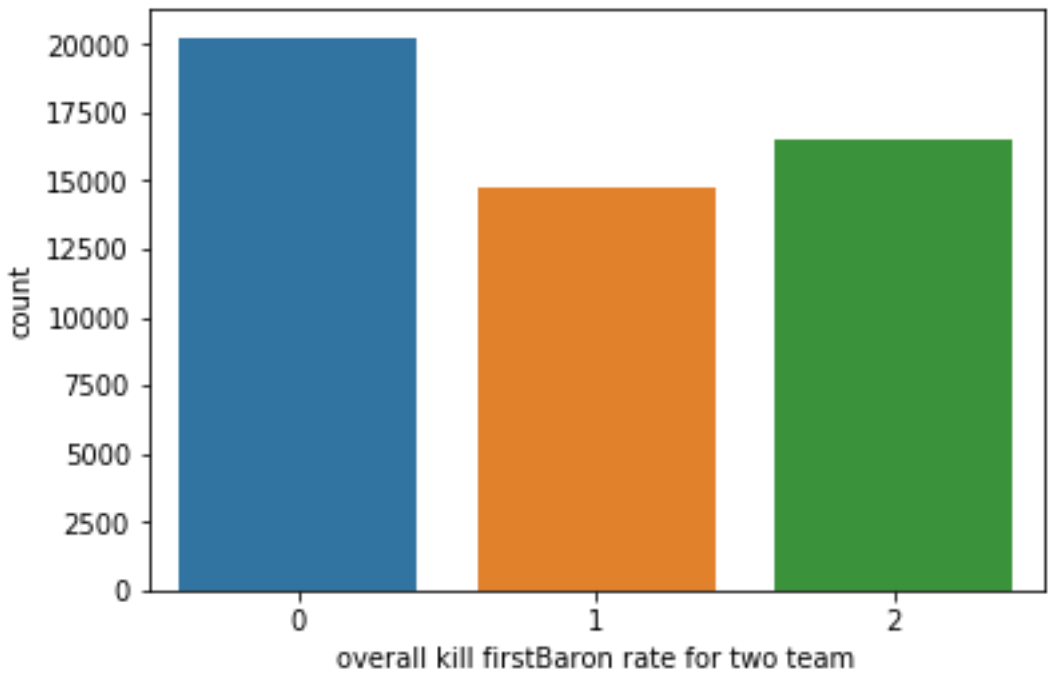
Figure 2.2: first tower rate for two team



Source: Calculated from data sample

Figure 2.2 is the overall rate from two team who got first blood kills. “0” stand for neither of team destroy first tower. “1” stand for blue team. ”2” stand for red team. The blue team still has higher rate than red team for tower destroyed.

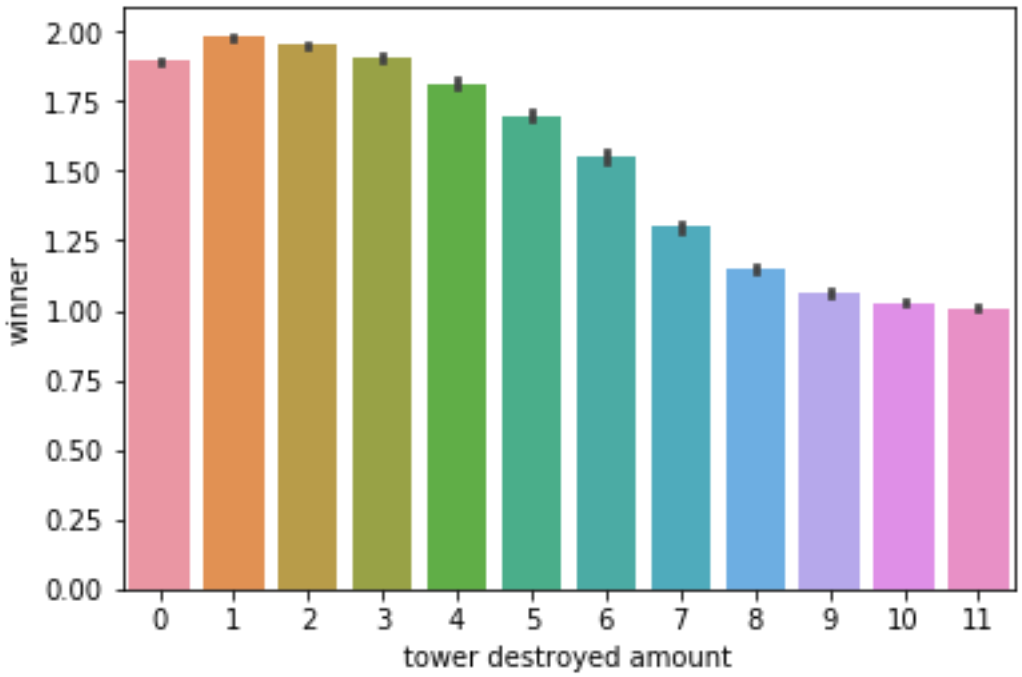
Figure 2.3: first baron rate for two team



Source: Calculated from data sample

Figure 2.3 is the overall rate from two team who got first baron kills. “0” stand for neither of team kill first baron. “1” stand for blue team. ”2” stand for red team. The blue team has higher rate than red team for slain the first baron.

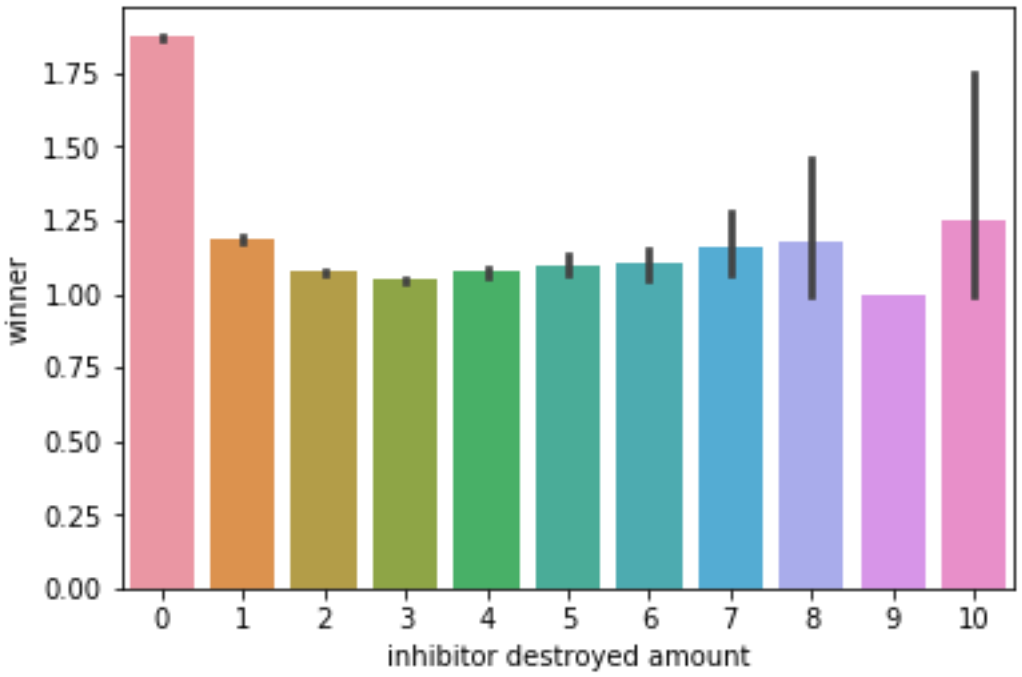
Figure 2.4: Tower destroyed amount impact on winning rate



Source: Calculated from data sample

Figure 2.4 is the relationship between tower destroyed amounts for blue team (team 1).X label for blue team tower destroyed amounts, and Y label tell us the numerical value closer to 1, the higher win rate for team 1. It shows that the more tower destroyed, the higher win rate that team 1 has. While before the first 4 towers destroyed, it is not obvious than several towers in the middle.

Figure 2.5: Inhibitor destroyed amount impact on winning rate



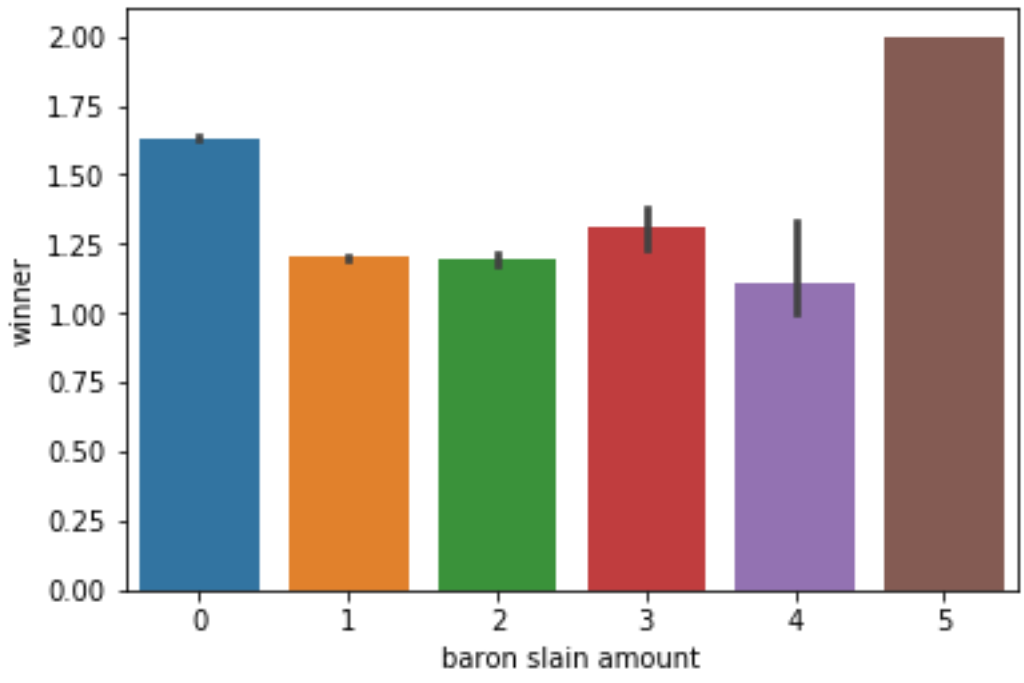
Statistical value for inhibitor kills:

t1_inhibitorKills	count	mean	std	min
0	25732.0	1.502951e+12	1.972757e+09	1.496892e+12
1	9567.0	1.502953e+12	1.942420e+09	1.496897e+12
2	8879.0	1.502867e+12	2.010882e+09	1.496894e+12
3	5104.0	1.502840e+12	2.018505e+09	1.496937e+12
4	1584.0	1.502941e+12	1.958737e+09	1.496954e+12
5	423.0	1.502996e+12	1.911817e+09	1.497030e+12

6	137.0	1.502702e+12	2.215730e+09	1.496931e+12
7	43.0	1.503204e+12	1.548848e+09	1.497720e+12
8	11.0	1.503036e+12	1.633769e+09	1.499266e+12
9	6.0	1.504128e+12	6.485525e+08	1.502842e+12
10	4.0	1.504040e+12	4.734939e+08	1.503440e+12

Figure 2.5 is the relationship between inhibitor destroyed amounts for blue team. X label for blue team inhibitor destroyed amounts. Y label is same as said above. It demonstrates that inhibitor destroyed amounts roughly have positive correlation with winning rate. However, it becomes negative correlation when inhibitor destroyed amount more than 3 times. I am curious about why there are some outliers appear. The results below the image is the count for different inhibitor kills. We will find out inhibitors kill more than 4, the sample capacities turn to decrease rapidly. That is the main reason there are some extreme value in this figure 2.5. result from small scale of data is not objective.

Figure 2.6: Baron slain amount impact on winning rate



Source: Calculated from data sample

Statistical value for inhibitor kills:

	count	mean	std	min
0	34901.0	1.502930e+12	1.980568e+09	1.496892e+12
1	14179.0	1.502914e+12	1.975232e+09	1.496924e+12
2	2251.0	1.502955e+12	1.947096e+09	1.496895e+12
3	149.0	1.502617e+12	2.140541e+09	1.496931e+12
4	9.0	1.503812e+12	7.133510e+08	1.502353e+12
5	1.0	1.502318e+12	NaN	1.502318e+12

Figure 2.6 demonstrates that slain baron is greatly helpful for win the game. While it has the same problem with previous question, the amount going down sharply, result from small scale of data is not objective and fair.

In this dataset, it was divided into two parts. One for match information, the other one saves the “Champion” information by Json (Java Script Object Notation). It is necessary to transfer the numerical value to exact “Champion” ID to see the result clearly. Otherwise it will be a lot of numerical result which is difficult for us to identify which ID matches with which definite “Champion”. Json is a lightweight data-interchange format. It is easy for humans to read and write. It is easy for machines to parse and generate(Kobayashi, 2011).

Firstly, we should transfer numerical ID to the exact “Champion” ID, and use “ID” number as index to order it

t1_champ1id t1_champ2id ... t2_champ4id t2_champ5id

0 Vladimir Bard ... Zed Thresh

1 Draven Irelia ... Yasuo Riven

2 Tristana Kayn ... Jax Ashe

3 Maokai Brand ... Riven Ashe

4 Warwick Twitch ... Kassadin Caitlyn

[5 rows x 10 columns]

t1_ban1 t1_ban2 t1_ban3 ... t2_ban3 t2_ban4 t2_ban5

0 Riven Janna Cassiopeia ... Karma Soraka Caitlyn

1 Caitlyn Darius Teemo ... Zed Caitlyn Illaoi

2 Lulu Janna Twitch ... Kha'Zix Maokai Evelynn

3 Zed Vayne Ornn ... Kayn Janna Caitlyn

4 Malzahar Lee Sin Thresh ... Braum Darius Tristana

In the beginning of a game, every player should ban and pick a “Champion” respectively. we need to combine pick list and ban list to two group, I create two vector to combine data, one from the first player pick to the fifth player pick and the other ban list is same as in pick list.

Total picklist

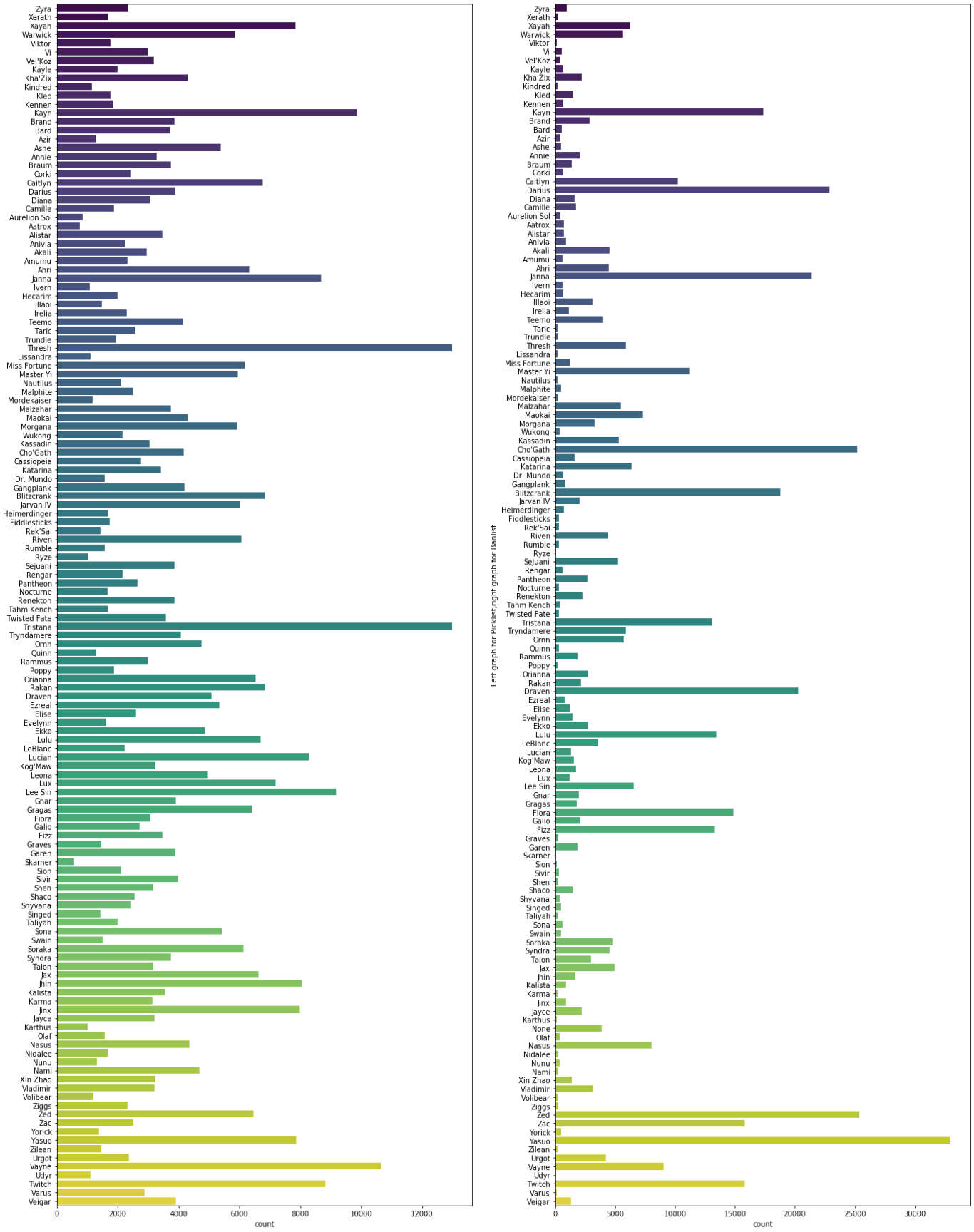
0 Vladimir
1 Draven
2 Tristana
3 Maokai
4 Warwick
5 Janna
6 Heimerdinger

7 Gnar
8 Nautilus
9 Ivern
dtype: object

Total banlist
0 Riven
1 Caitlyn
2 Lulu
3 Zed
4 Malzahar
5 Lulu
6 Yasuo
7 Lulu
8 Zac
9 Cho'Gath
dtype: object

We can see the frequency of ban and pick list:

Figure 3 Frequency of ban and pick list



From pick list, it is not hard to find out Thresh and Tristana is the most popular “Champion”. For ban list, Yasuo has been banned exceed 3000 times far more than second place. From the other hand, Skarner is the tail ender on pick list. It does not make sense to mention the tail ender on ban list, because everyone just wants to ban the most powerful and unbalance champion.

2.3 Supervised method: Logistic Regression Model

The data requires some pre-processing though plenty of the data is present in the dataset. For the section of supervised method, it is obviously that exported game data similar with binary. Because there are dummy variables in the dataset, like for “winner”, “1” represent blue team win and “2” represent red team win. Logistic regression assumes that the data obeys the Bernoulli distribution, and uses the method of maximizing the likelihood function to solve the parameters by using gradient descent to achieve the purpose of classifying the data. So, it indicates that the appropriate method is the Logistical model. We want to segment our dataset into 2 parts, one for model training and the other to validate the model. Using sk-learn, then we can segment our data. Majority of data will be trained, and test remains data. The proportion of segment will be 8:2.

I choose several of representative indicator as independent variable. (“first Blood”, “first Tower”, “first Baron”, “first Inhibitor”, “first Dragon” and “first Rift Herald”) furthermore, all of these indicators directly relate to the final victory. It can be explained that these independent variables oriented to a same dependent variable which is winning rate. Put two variables in X, Y and run the logistic regression model, print the report in the end.

	precision	recall	f1-score	support
1	0.83	0.83	0.83	20881
2	0.83	0.82	0.82	20311

avg / total 0.83 0.83 0.83 41192

The score provides the information about how well the model fit. 83% data tally with this model. Analysis from the subjective perspective of the game, these 6 indicators could directly bring players virtual game-revenue to varying extent, thus crushing the enemy with overwhelming economic advantage. It is no wonder that the model fit test data with highly significant.

2.4 Unsupervised method: K-means Clustering

If there is a category label, the clustering result can also calculate the accuracy and recall rate. classification labels also can be used as evaluation indicators for clustering results. Completeness and homogeneity score are a part of V-measure to evaluating dataset. I also chose the same X independent variable (“first Blood”, “first Tower”, “first Baron”, “first Inhibitor”, “first Dragon” and “first Rift Herald”) and Y dependent variable (“winner”) from that I have applied in logistic regression model. It not only to evaluate these indicator, but also could make a comparison between logistic regression result and K-means result

.

I set the number of cluster equal to unique amount of dependent variable(“winner”), in fact, “winner” only have two kinds result which is 1 and 2. The rest of indicator just follow default. So, the research question is:” to what extent these 6 indicators can affect the winning rate?”

The result as follow:

completeness_score: 0.22277816825627295
homogeneits_score: 0.22263336158083155

The result is not satisfactory. It is opposite to the result in logistic regression model. So, I make some change the number of cluster to 5 to see the result.

completeness_score: 0.10508187722580403

homogeneits_score: 0.2423288127852485

at last I set the number of cluster to 10:

completeness_score: 0.09791294898883415
homogeneits_score: 0.31130482752822247

All result is not accord with our imagination, moreover, it has a big difference between different number of clusters.

3. Reflections of model

After applying the dataset into two models, I found that the K-Means, the number of cluster is difficult to estimating and confirm the amount of classifications in most appropriate labels. On account of our dependent variable is a dummy variable, dummy variable makes the question's description more concise, however, its performance in K-means is not good enough, even I change the number of cluster, the performance is not stable at all.

K-means is not applicable to categorical classification. I think this is the most important reason why K-means does not make sense on my dataset, for example, in the dataset that I select the number of cluster should not to be modify because it is just like dummy

variable, it only has exact two results: win the game or be defeated (0-1 distribution). If you use K-means, it is impossible that the values of the two centers are between 0-1, so it is not in line with the actual situation. My current comprehension is that discrete distribution is not available k-means. But it works on continuous distribution.

As for the logistic regression model, it is not demanding to understand, and it is a kind of linear model, additionally, it also has a good effect on finding out a several of variables in how it has an impact on one dependent value, and logistic regression assumes that the data obeys the Bernoulli distribution, it is close to this type of dataset. It is helpful for my result.

4. Conclusion

This paper is emphasis on processing data by python and use some quantitative method to make a preliminary analysis for game winning rate, find out these indicators have positive correlation with winning rate. Besides I summarize some potential phenomenon in the game, it could improve the game players experience based on the phenomenon.

I also use two machine learning method to analyze this dataset. The first is K-Means, even it is not applicable for this dataset, but I find out some essential feature of K-means. For logistics regression, it verifies my hypothesis is highly significant to the game result, from the 83% precision of predicting result, I understand what type of dataset is appropriate for logistic regression, it is helpful to comprehend how to predict the League of Legend winning rate by big data techniques and tools.

Reference

Braun, Cuzzocrea, Keding, Leung, Padzor, & Sayson. (2017). Game Data Mining: Clustering and Visualization of Online Game Data in Cyber-Physical Worlds. *Procedia Computer Science*, 112(C), 2259-2268.

Kobayashi, N., Ishii, M., Takahashi, S., Mochizuki, Y., Matsushima, A., & Toyoda, T. (2011). Semantic-JSON: A lightweight web service interface for Semantic Web contents integrating multiple life science databases. *Nucleic Acids Research*,39(Suppl_2), W533-W540.

https://www.medcalc.org/manual/logistic_regression.php

https://www.kaggle.com/laowingkin/lol-how-to-win-the-world-championship

https://www.kaggle.com/jaytegge/league-of-legends-logistic-regression-analysis

Appendix

Python version: 3.7.0 (default, Jun 28 2018, 08:04:48) [MSC v.1912 64 bit (AMD64)]
Package used: Numpy, Pandas, Matplotlib, Matplotlib.pyplot, Seaborn,Sk-learn