Dota2 Outcome Prediction

Zhengyao Li[1], Dingyue Cui[2] and Chen Li[3]

[1] ID: A53210709, Email: zhl380@eng.ucsd.edu
[2] ID: A53211051, Email: dicui@eng.ucsd.edu
[3] ID: A53218665, Email: lic055@eng.ucsd.edu

March 13, 2017

Dota2 is a very popular video game currently. In this paper, we proposed three different models to predict the outcomes of Dota2 matches at the very beginning of the game. We discussed the data and feature exploration, which includes the analysis of collaborative and countering relationship between different heroes. In our final prediction model, we use logistic regression to predict the outcome with only hero historical information. In the other models for comparison, we use both the players and heroes historical information.

1 Introduction

Dota2 is the most popular multiplayer online battle arena (MOBA) video game in the world, where two teams of five players compete to collectively destroy a large structure defended by the opposing team known as the "Ancient", whilst defending their own.[1] Ten players each control one of the game's 113 playable characters, known as "heroes", with each having their own benefits and weaknesses. During the matches, each player can only choose different heroes.

Here are some basic rules and ideas of the Dota2 game:

1. Each player belongs to either one of the team - the Radiant and the Dire. Each team has 5 players and they fight against each other.

2. Each player controls different heroes. Each hero can enhance their strength through two means - by gaining experience (XP) to upgrade their heroes and by gaining gold to buy some useful game props, which both can get from killing enemy minions and heroes.

3. Since it is a 5vs5 MOBA game, the outcome of the game not only depends on the strength of each player, but also depends on each player's hero and the cooperation between them.

4. There are mainly two kinds of Dota2 game matches: normal matches and ranked ladder matches. The normal matches are played by random 10 players, and the ranked ladder matches are more competitive since the system will arrange both teams with players have equal competition strength.

The game is not only full of fun, but also full of competitiveness driven by money award. Premium Dota 2 tournaments (the highest level of ranked ladder matches) often have prize pools totaling millions of US dollars, the highest of any eSport.[2] And that's the reason why we want to build a good predictor to analyze the outcome of the ranked ladder matches.

2 Dataset

2.1 Dataset Introduction

We use 50,000 ranked ladder matches from the Dota2 data dump created by Opendota[3] to train our model, which contains enough data to enable us to train a good model. And we also have 100,000 ranked ladder matches with labels to test our model.

The data we used mainly contains the table 1, 2, 3, 4 and 5:

Table 1: *Dataset tables*

Table Name	Description
match.csv	General info about a match
players.csv	Info about each player in each match
test labels.csv	Outcome of a match used for test
test player.csv	Info about each player in test matches

Table 2: *match.csv - 50,000 rows*

Attribute Name	Description
match id	unique id for each match
start time	start time for each match
duration	duration for each match
radiant win	whether radiant win the match

Table 3: *players.csv - 500,000 rows*

Attribute Name	Description
match id	unique id for each match
account id	unique id for the player
hero id	id of hero that this user use
player slot	indicates the team
gold per minute	gold collected per minute
xp per minute	experience gained per minute
kills	number of kills
deaths	times of death
assists	time of assists to teammates

Table 4: *test labels.csv - 100,000 rows*

Attribute Name	Description
match id	unique id for each match
radiant win	whether radiant win the match

Table 5: *test player.csv - 1,000,000 rows*

Attribute Name	Description
match id	unique id for each match
account id	unique id for the player
hero id	id of hero that this user use
player slot	indicates the team

The first two tables describe information about 50,000 matches. It roughly includes the matches' outcome, which 10 players participated, which 10 corresponding heroes the players use, which team they belong to and some statistics about gold, experience gained and kill, death and assist times. These two tables will be used for training.

The last two tables describe information about another different 100,000 matches. They only contain those data we could know before the match actually begins, and don't contain some statistics during the match. These two tables will be used for testing.

2.2 Findings

2.2.1 Abnormal Data

After observing the dataset, we found some abnormal data. For example, one of the $hero\_id = 0$ in a match, these data could be interpreted as missing one player during a game. And for players with $account\_id = 0$, they are anonymous players who are not willing to show their account number but still have complete game information.

2.2.2 Win Rate

The most basic statistic of data is win rate, which can be calculated with the following equation.

$$\frac{number\ of\ win\ matches}{number\ of\ matches} \tag{1}$$

Win rate can be analyzed both on players and heroes. A player's win rate indicates his past performance, and a hero's win rate is able to represents if it is powerful in game.

2.2.3 KDA

There is an interesting statistic called KDA, which indicates a proportional relationship between number of kills, assists and deaths, and this feature usually stands for a player or a hero's performance during a game. The following equation shows how to calculate KDA in a match.

$$\frac{number\ of\ kills\ +\ number\ of\ assists}{number\ of\ deaths\ +\ 1} \tag{2}$$

2.2.4 GPM and XPM

The other two interesting statistics are GPM and XPM, whose full names are Gold per minute and XP(experience) per minute. These two features also represent a player's capability and performance in a match. The higher average GPM or XPM one team has, the more likely that the team could obtain victory.

3 Predictive Task

The purpose of our work is to forecast the outcome of a match for Radiant exactly at the very beginning of a match. That means we only know the ten users participate in the match, the ten corresponding heroes they choose and which five of them belongs to Radiant team.

3.1 Evaluate Method

In our work, we will use prediction accuracy to evaluate our method.

$$prediction = \frac{correct\ predictions\ number}{total\ match\ number}$$

3.2 Baseline

Since electric sports has high randomness, we use randomly prediction as our baseline.

3.3 Features

3.3.1 KDA

KDA is a useful feature to indicate a player's performance during a game. If a player obtains a high KDA in a match, it means that he made more kills and assists than deaths, and this is the criterion to judge that he performed excellently in this game. By counting all previous KDA of a player, we can figure out an average level of his capability in the game, and it will help his team to win the game. This finding can be verified in Figure 1, if radiant members have higher previous KDAs than that of dire, which means the KDA difference is positive in this figure, the Radiant is more likely to win, and vice versa.
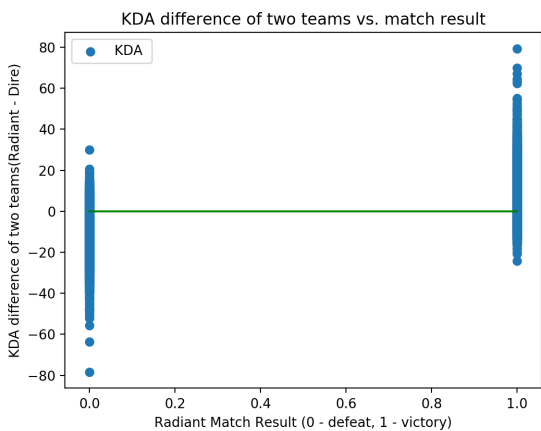

Figure 1: *Relationship between two team's KDA difference and match results*

3.3.2 GPM and XPM

GPM and XPM are also related to player's performance. If one player get high GPM or XPM during a game, it represents that he has an outstanding performance in this game. With his contributions, he and his team are able to get better game props and could upgrade faster, and obtain victory easily. And if a team's overall historical GPM or XPM is extremely higher than that of the opposites, it is safe to say that players of this team have more strong capability, and thus they are more likely to win. This finding can be confirmed in Figure 2. if radiant members have higher previous GPM or XPM than that of dire, radiant is more likely to win.
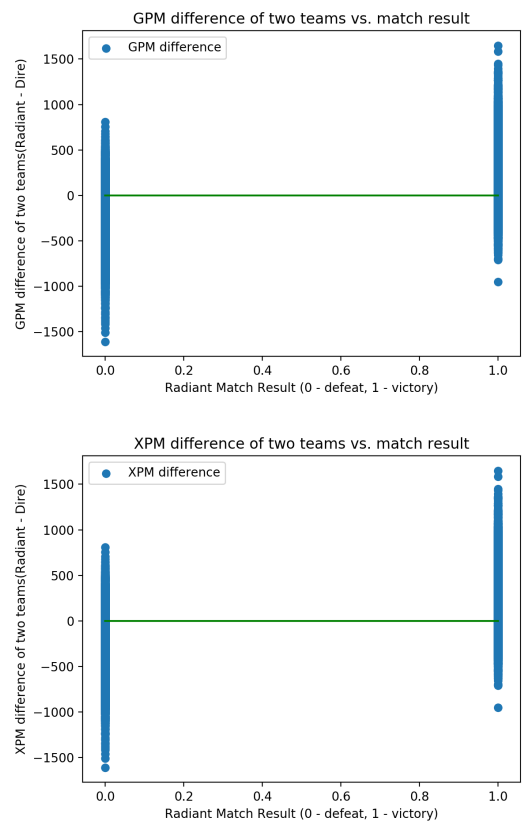

Figure 2: *Relationship between two team's GPM, XPM difference and match results*

3.3.3 Hero attribute

Besides players' performance, hero attributes could also influence the match results. There are overall 3 kinds of hero attributes: Strength, Intelligence and Agility. As we know, a team needs a full lineup to win, so whether a team has integrated battle formation can be used to help predict its win rate of a match.

3.3.4 Hero collaborate and counter

The skill collaboration and counter of heroes are also important factors of a game. If heroes can collaborate with each other in a game, they are more likely to generate extra effects. And if a hero in a team is countered by another hero in his enemy's team, it will lead a disadvantageous situation. Thus we counted the win rate when two heroes are on the same team and on the opposite team to indicate their relationship of collaboration and countering.

As the following figure shows, much of the pair heroes have a nearly 50% win rate, however, there also exists remarkable pair heroes have extremely good or bad collaboration. These data have a lot more information with a match's result, if a team has two heroes with a well cooperation, it is more likely that they will win the match.
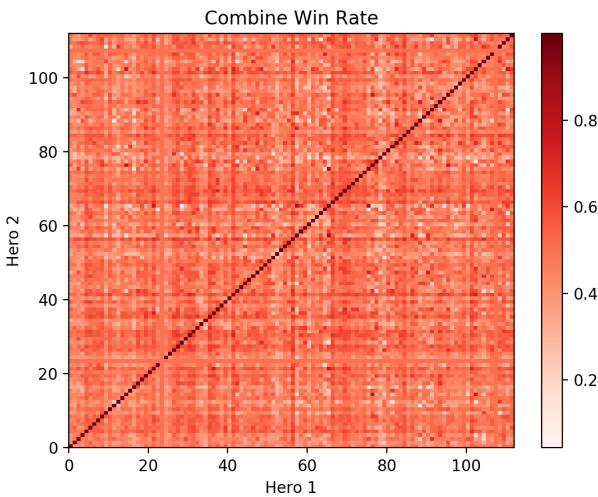

Figure 3: *Win Rate of pair heroes*

For hero counter, similar to hero collaborate, even much of the counter heroes have a nearly 50% win rate, there exists some hero pairs have extremely counter effect on each other. If these hero pairs are shown in different teams in a match, the result of the match are likely being influenced by the hero counter factors. Hero counter relationships is shown in Figure 4.
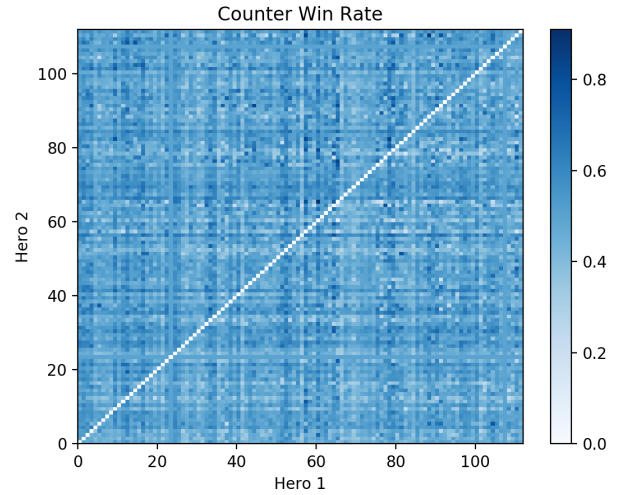

Figure 4: *Win Rate of counter heroes*

4 Model

4.1 Data cleaning

To train our data efficiently and accurately, we mainly cleaned out three types of abnormal data:

- The matches with any missing heroes, players and other important items.
- The matches with $hero\_id$ is 0, which could be interpreted as that one player hasn't chosen the hero and already left the game.
- The matches with the duration less than 900 seconds (15 minutes), which could be interpreted as that someone left the game too early.

4.2 Logistic Regression Model

Our final model uses logistic regression to predict the outcome for Radiant in a match.
In a nutshell, the model is:

$$outcome_i = \begin{cases} 1 & \text{if } X_i\theta > 0 \\ 0 & \text{if otherwise} \end{cases}$$

where $outcome_i$ represents the outcome of the i_th match for Radiant. One represents Radiant win, and zero represents Radiant lose. X_i represents the feature vector for the i_{th} match.

4.2.1 Features

Our model contains the following features:

- **Offset**
  Offset is set to be 1, which allows the model to

consider the general advantage of Radiant over Dire.

- **Cooperate Win Rate Difference**
  This feature considers the influence of the combination of heroes in each team. Because cooperation is very important in this game, so a specific pair of heroes may even control the trend of game, we cannot simply combine the effect for each hero to predict the effect when both of them appear in the same team.

  So in our model, we calculate the average historical win rate when a pair of heroes appear in the same team. And calculate the average win rate when we combine all the possible hero pairs. For each team, we will have such a value, and use their difference as the feature:

$$\text{Cooperate Win Rate Difference} = \frac{\sum_{(i,j) \ in \ Radiant} Crate_{i,j}}{\binom{5}{2}} - \frac{\sum_{(i,j) \ in \ Dire} Crate_{i,j}}{\binom{5}{2}}$$

  where (i, j) represents a pair of hero id in each team, and $Crate_{i,j}$ represents the average win rate when hero i and hero j appear in the same team.

  In a nutshell, this feature represents the extent of win rate for Radiant after both teams choose their heroes. So the larger this value, we are more confident that the win rate of the Radiant is larger than the win rate of Dire.

- **Counter Win Rate**
  We should also consider the counter relationship between heroes. Since there are some heroes could counter the others, if the Radiant team has some heroes who counter the heroes in Dire, we should be more confident that Radiant will win.

  So we calculate the average historical win rate of hero i when hero j is in the other team. Then after both teams choose their heroes, we could calculate the average win rate when each of Radiant's heroes face each of Dire's heroes:

$$\text{Counter Win Rate} = \frac{\sum_{i \ in \ Radiant} \sum_{j \ in \ Dire} Crate_{i,j}}{\binom{5}{1}\binom{5}{1}}$$

  where $Crate_{i,j}$ represents the average win rate of hero i when hero j is in the other team.

So actually the Counter Win Rate represents the extent of the Radiant win rate according to the counter relationship between heroes in both teams.

- **One-hot of Hero Attributes**
  Hero's attribute indicates its strength and weakness. So it's also important for a team to contain heroes with different attributes.

  There are totally three different attributes, and we use one-hot vector to represent the appearance of different attributes in each team. So the vector contains six bits, where the first three bits represents the appearance of three different attributes in Radiant, and the others three bits represents the appearance of three different attributes in Dire.

Table 6: *one-hot vector of hero attributes*

one-hot bit	Description
0/1	whether Radiant has strength hero
0/1	whether Radiant has intelligence hero
0/1	whether Radiant has agility hero
0/1	whether Dire has strength hero
0/1	whether Dire has intelligence hero
0/1	whether Dire has agility hero

4.3 Other Models

Other than our final logistic regression model, we also have tried the following two models to do the prediction.

4.3.1 Linear Regression

Our linear regression model uses the following features to train the model and do the prediction:

- **Offset**
  Offset is set to be 1, which allows the model to consider the general advantage of Radiant over Dire.
- **matchWinRate**
  The player's average history match winning rate.
- **GPM**
  The player's average GPM per match.
- **XPM**
  The player's average XPM per match.
- **KDA**
  The player's average KDA per match.
- **matchNum**
  The player's history total match numbers.
- **heroWinRate**
  The average winning rate of the hero used by the player in the ranked ladder matches.

- **heroShowRate**
  The average showing rate of the hero used by the player in the ranked ladder matches.

So for each match, we use linear regression to predict the win rates of two teams, and compare the results. If the Radiant team has a larger win rate than the Dire team, then we predict Radiant will win the game. Otherwise, we predict Radiant will lose this game.

4.3.2 Simple Latent Factor Model

While the heroes' abilities influence the outcome of a match, the players' level also plays an important part. That means the outcome of a match is influenced by the players and heroes. So we want to use the information about the combination of players and heroes to predict the KDA given a pair of user and hero. So we have:

$$KDA = \alpha + \beta_p + \beta_h$$

where β_p is a vector represents the KDA difference above average for each player, and β_h is a vector represents the KDA difference above average for each hero.

So when we are facing a new match, we can predict the KDA according to the information about player and hero combination to predict the KDA for this role in this match. Since each team has five players, we simply combine the predicted KDA of five roles in each team, and compare the total of KDA. If Radiant has a larger KDA than the other, we predict Radiant will win. Otherwise, we predict Radiant will lose.

4.4 Model Comparison

4.4.1 Logistic Regression Model vs Simple Latent Factor Model

The strength of logistic regression model is that it takes the combination of heroes into account. Because this game is a team contest, each hero has its own strength and weakness, the cooperation of heroes are the key point to win the match. So considering the effect of different combinations of heroes is import to predict the outcome. That's why we use logistic regression as our final model. But it doesn't consider any information about players. So sometimes even some heroes have high win rate, but the player cannot perform well, the outcome of the match may also be influenced largely.

The latent factor model considers both the level of players and heroes. Even if a player didn't use a hero before, we could still predict the KDA of it. However, this model also has some problems. Because the number of players is very huge, many of players are those we haven't seen in test data. So it's hard to predict the KDA of those players in the test data, and may also cause a higher wrong predictions rate.

4.4.2 Logistic Regression Model vs Linear Regression Model

Linear Regression is our first model, and we separate the match into two parts: predict the win rates for two teams and compare them. Its performance is close to latent factor model, because we all predict match outcome separately. However, in our logistic regression model, we use the difference of win rate as a feature, which means we consider the whole match during the training.

5 Related Works

Our dataset comes from the Opendota. In the state-of-the-art paper *Real-time eSports Match Result Prediction*[4], the researchers used the same dataset like us to do the real-time outcome prediction with the combined model of the logistic regression and the Attribute Sequence Model. At the mean time, they also predict the win rate at the very beginning of a match, which generated similar prediction result with the result our work. They used logistic regression with different features to do the predictions, and got the accuracy of 58.26%, while our accuracy is 59.713%.

Some researchers used a similar dataset that is gotten from Steam Web API[6]. This dataset contains only the regular matches data, while ours focus on ranked ladder matches. The report *Dota2 Win Prediction*[5] uses this dataset to do the outcome prediction before the contest starts with the logistic regression model too. They use one-hot vector contains more than 200 bits to represent the appearance of each hero in both teams, and also consider the combination of heroes, but they didn't consider the attributes of the heroes, which is somewhat unreasonable.

6 Results

6.1 Performance of different models

Table 7: *Results*

Model	Prediction Accuracy
Randomly prediction	49.873%
Linear Model	52.618%
Latent Factor Model	53.772%
Logistic Regression	59.713%

Our baseline performs a nearly 50% prediction accuracy, which is reasonable on such sports predictions. The linear model and the latent factor model have similar performance which is slightly better than baseline, but it is already a big progress for prediction with strong randomness. And our final model logistic regression made a significant improvement.

6.2 Interpretation of Parameters

Table 8: *Feature and Parameter for Logistic Regression*

Features	Parameters
Offset	-2.73
If radiant has Strength hero	-5.16×10^{-3}
If radiant has Agility hero	3.39×10^{-3}
If radiant has Intelligence hero	1.24×10^{-1}
If dire has Strength hero	-6.26×10^{-2}
If dire has Agility hero	-7.92×10^{-2}
If dire has Intelligence hero	-1.54×10^{-1}
Hero collaborating win rate difference	10.42
Hero counter win rate difference	5.93

For offset, we can interpret this parameter as a subordinate condition on radiant team. For hero attributes, intelligence hero have more positive influence when it is in the Radiant team and more negative influence when it is in dire team. Comparing with intelligence heroes, strength and agility heroes have fewer effects. For hero collaborating and counter, if the sum of the Radiant hero collaborating win rate sum is higher than that of the Dire team, it is more likely that the Radiant will win the game, and it is the same as hero counter win rate.

6.3 Feature Performance

During the regression, we found that even if hero attributes could make some improvement, using the feature hero collaborating and counter could improve the prediction performance a lot more, which means these two features are the most useful.

7 Conclusions

So far, we have finished three models to predict the outcome of a Dota 2 match at the very beginning of it. Our best model is logistic regression trained by 50,000 different matches, and achieved an accuracy of 59.713% in the test set that contains 100,000 different matches.

The accuracy maybe not good if we just look at the number. However, our test set is very huge, even larger than our train set. That means there must be several outliers that are very hard to predict. Also, a match is very unstable, not to mention that our match data comes from ranked ladder matches, which has players in different teams with close capability. These facts contribute to that a player's history data is not useful as we expected in latent factor model or linear regression model. Also, so many facts during the match may influence the outcome, which makes it harder to predict the outcome just after heroes picked. So the task itself is very hard and unstable if we just look at the limited information before a match begins.

However, our model still has a lot of meanings. It beats the baseline by 10%, which means that we are more confident to predict the outcome of a match after we know the heroes picked. And also useful to give recommendation for users to pick heroes.

8 Future Work

Though we have achieved a rather high accuracy of the outcome prediction in the test set, due to the time constraints we still have some promising improvements to make:

- By now, the algorithm is so slow, it takes several hours to clean and organize the data, and another several hours to train the model and to do the test.
- Due to the dota2 game is too popular, thousands of matches happen every day and millions of users play this game every day. Even though our data set has over 15,0000 match entries, the time span of the matches is rather small and the repeated coverage of users is rather limited.
- The dota2 is a vigorous game and will update monthly, and accordingly the game balance, map, heroes will dramatically change. On the other hand, since the dota2 is a competitive game, a player's competitive state will change over time as well. Therefore, for a more general predicting model, we do believe for each user feature training process, that to use a sliding window match history is way better than to use all the match history.

Except the above points need to be achieved in the future, our model just predicts the outcome before the game starts, but a reliable real-time predictive model is also very useful in terms of match prediction. How to convert and revise our model to accurately do this kind of predictions is also we want to do in the future.

Reference

[1] *McDonald, Tim. "A Beginner's Guide to Dota 2: Part One - The Basics". PC Invasion. Retrieved 1 August 2016.*
[2] *https://en.wikipedia.org/wiki/Dota_2*
[3] *https://www.opendota.com/*
[4] *Yang Y, Qin T, Lei Y H. Real-time eSports Match Result Prediction[J]. arXiv preprint arXiv:1701.03162, 2016.*
[5] *https://cseweb.ucsd.edu/jmcauley/cse255/reports/ fa15/018.pdf*
[6] *https://developer.valvesoftware.com/wiki/Steam_Web_API*