Predicting the outcome of Dota 2 Matches using Elo

Petar Kostic
Utrecht University
4075897

Mark Berentsen
Utrecht University
5511941

***Abstract*—In this paper, we present a prediction model for the popular computer game Dota 2. We detail previous efforts of predicting the outcome of matches for the game and introduce a new concept for approaching this problem using the Elo rating system. Furthermore, we provide insight into our data collection.**

Finally, we look back on our findings and conclude that Elo ratings on their own could be used for predictions given some different methods. Although, due to the nature of the game itself, a prediction model supported by more information will be more accurate.

I. Introduction

Dota 2 is an immensely popular computer game, made by Valve in 2013. It has an active player base of over 12 million unique players, as of February 2017 [1]. Even though a lot of users play it casually, there is also a professional side to the game, consisting of tournaments with massive monetary payouts. For example, Valves recent Dota 2 tournament, ”The International 2016”, had a prize pool of over $20 million [2].

A Dota 2 match is played in two teams, each consisting out of five players. Before a match begins, each player is able to choose a ”hero”, which is an unique character to fight as. As of now, there are a grand total of 113 heroes, which makes sure that each match is unique [3]. The player uses their hero to defeat enemy heroes and minions that guard the enemy base. As the game progresses, players are able to spend their acquired gold on items that will make them more powerful.

With so many matches that are being played constantly, it is no wonder that there is a high interest in predicting outcomes of matches. Although controversial, there are betting sites to be found for exactly this purpose. This phenomenon eventually led us to conduct this research: is it possible to predict the outcome of a Dota 2 match?

In ”Dota 2 Win Prediction” [4], the authors provided a model that is able to predict the outcome of a match with 73% accuracy, solely using the hero picks as information source. While this is impressive, it solely provides predictions based on hero choices. Skill level of players is left out of consideration.

In ”To win or not to win? A prediction model to determine the outcome of a Dota 2 match” [5], the author presents an improvement to a simple logistic regression model. It is, once again, based on hero combinations in teams.

In ”Predicting the winning side of DotA2”[6], the authors tried to reproduce the above method with a smaller dataset, but found out that the test error was too high. To improve this, stepwise regression and feature selection were used to improve their predictions. They also mentioned that *”Future work should take the player into account to make the prediction more accurate.”*

It seems that most of the related work follows this structure. We aim to provide a prediction model that is able to predict outcomes of matches based on player skill level instead. To achieve this, we will make use of the Elo rating system [7], which we will give to each player in our data. We exclude hero picks from our research scope, so we can test our model in its purest form. We hope to achieve a prediction model that is able to make fairly accurate predictions, based solely on player skill.

II. dataset

We used Valve’s web API [8] to pull data for 1.111.106 matches with 1.234.166 unique players between 21/09/2016 and 25/11/2016. The scraped data is subject to the following requirements:

- To make sure we are only dealing with human players, we only allowed matches with 10 players present in them. This way our Elo ratings will not get influenced by AI controlled players.
- The game mode of the match is either random, random draft, captain’s draft, captain’s mode, single draft, all pick, or least played. We have chosen to solely include these modes since they are closest to the true competitive vision of Dota 2.
- No players have left the match before it is finished. These matches have unbalanced teams, which would interfere with the distribution of correct Elo ratings.
- We only take matches into consideration where all players in it have chosen to expose their public match data [9]. Without this, we cannot identify unique players, which would impact the Elo rating distribution in a negative way.

Every match is structured as a single line in a JSON format, which includes a plethora of information. In the end we only used a fragment of it to gather the unique player id’s, the player slot, and which team eventually won.

Prior to our analysis, the data is shuffled to prevent over-fitting to particular timezone or regions of the world.

III. Methodology

The main objective in Dota 2 is to destroy the opposing team’s base. One team is called the ”Radiant”, while the other team is called the ”Dire”. These terms are roughly analogous to ”home” and ”away”, as they only determine the starting point of each team on the game world map, which is roughly symmetric. Each game can only end in a win or loss for a team.

A. An adapted Elo rating system

We will briefly describe our (somewhat naive) modification of the logistic curve Elo rating system. Every new player we encounter in our dataset starts with an Elo rating of 1000. Based on the results of a set of preceding matches, each player's Elo rating R_{player} is updated as an average measure of the team's current rating that he is in. R_D is the average rating of the players in the dire team and R_R is the average rating of the players in the radiant team. The expected score for the Dire team is calculated as follows:

$$E_D = \frac{1}{1 + 10^{(R_R - R_D)/400}}$$

$$\Delta R = K(\alpha - E_D)$$

$$where\ \alpha = \begin{cases} 0 & if\ radiant\ wins\ the\ match \\ 1 & if\ dire\ wins\ the\ match \end{cases}$$

The Elo ratings are updated after the match and the new rating for each player in the dire team is.

$$R_{player} = R_{player} + \Delta R$$

The new rating for the radiant team is calculated in the same way. The rating follows the player, and is updated after each match. We used an empirically determined value of $K = 50$ in all of our rating calculations.

B. Adjusting the dataset for predictions

According to a blog post by the Dota 2 Dev team[10], the rating for players with no prior experience in the game converges around 500 games played, while the rating for experienced players converges around 50 games played. In order to take the Elo convergence of all players in our dataset into account, we only selected matches where all 10 players have at least played 100 matches for predictions and statistical analysis. The selected matches will always have the most recently updated ratings. As a result, a large chunk of our dataset is used to compute the Elo ratings for all players in our dataset. Out of 1.111.106 matches, 24.347 were selected for analysis.

C. Predictions and statistical procedures

Since the outcome of a Dota 2 match can be represented as a dependent (dichotomous) variable, we made use of logistic regression [11] in favor of the radiant team. To use the Elo ratings for making predictions, the model was fit with a single covariate: the rating difference in favor of the radiant team for each match:

$$R_{diff} = R_R - R_D$$

Unlike linear regression with ordinary least squares estimation, there is no R^2 statistic which explains the proportion of variance in the dependent variable that is explained by the predictors. However, there are a number of pseudo-R^2s that could be of value. There is some discussion around the significance of these so we only used them as a rough indicator of how well we are predicting the outcome.

We assessed the predictive ability of our model by calculating the mean of the outcome, using the following decision boundary against our test set.

$$y = \begin{cases} 1 & if\ P(y = 1|X) > 0.5 \\ 0, & otherwise \end{cases}$$

Finally, we ran our model on the test set, plotted an ROC curve and calculated the AUC (Area Under the Curve), which is a typical performance measurement for a binary classifier

IV. Results

TABLE I. RESULTS AND ANALYSIS OF LOGISTIC REGRESSION MODEL

Parameter estimates				
Predictor	Coef.β	SE(β)	z-value	**p**
Intercept	0.1179609	0.0146489	8.053	8.11e-16
R_{diff}	0.0040826	0.0001556	26.232	<2e-16
Pseudo-R^2		Wald test		
McFadden	0.0284187	Predictor	df	χ^2
Cox & Snell	0.0385636	R_{diff}	1	688.14
Nagelkerke	0.0514603	Decision boundary accuracy = 0.504		

The Coef.β in the second column of Table I is the coefficient associated with the variable listed to the left. It is the estimated amount by which the log odds of "Probability of winning" would increase if the R_{diff} would be one unit higher.

In the next column, we see the standard error (SE(β)) associated with these estimates. The z-value in the next column is obtained by dividing the Coef.β by the SE(β). This quotient is assumed to be normally distributed with our sample size. Next to the z-values are the two-tailed p-values that correspond to those z-values in a standard normal distribution.

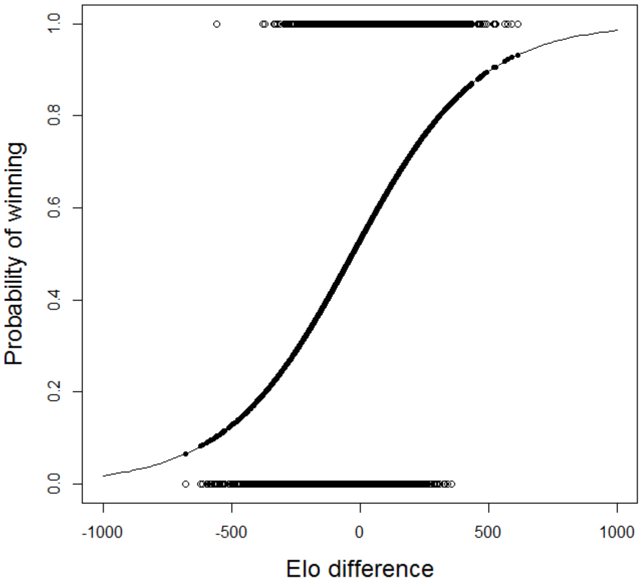

Fig. 1. Logistic model probabilities of a radiant win as a function of the rating difference. Note overlap of the black dots on the top and bottom, these are the individual R_{diff}s of observations placed at appropriate outcomes (top = 1.0 = radiant win, bottom = 0.0 = radiant loss)

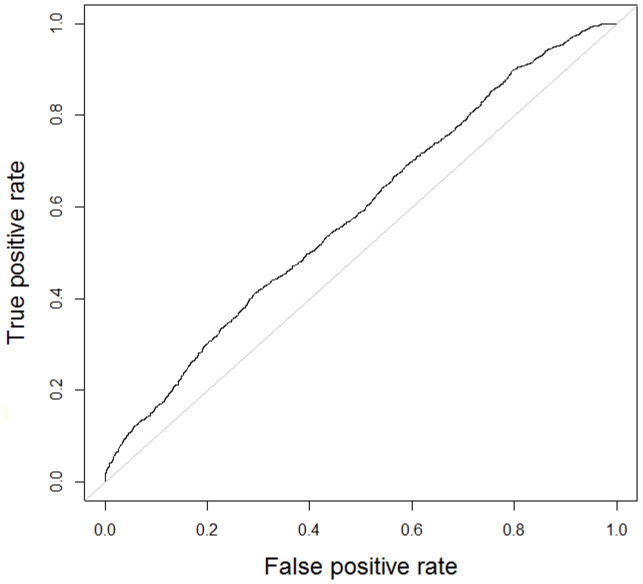

Fig. 2. The ROC curve of ran using the model seen in Table I and Fig. 1 on the testing set, AUC = 0.58

We used a Wald test [12] to evaluate the statistical significance of the R_{diff} predictor in our model.

V. Discussion

The first result of value in Table I is the p-value for R_{diff}. Such a low p value indicates significant at the 5% level. The null hypothesis that the predictor's coefficient is equal to zero (Intercept), which indicates that there is no association between the predictor and the response, can thus be rejected. The large χ^2 distribution with one degrees of freedom (df) of the Wald test [12] supports this result. Although this showed promise at first, the decision boundary accuracy on our test set in the last row of Table I is roughly equal to 50%. Which shows no predictive accuracy at all. This is supported by the low pseudo-R^2 values. These measures range from 0 to just under 1, with The McFadden of the model reflecting the fact that the R_{diff} covariate does not enable accurate prediction of the individual binary outcomes.

In contrary, the area under the ROC curve in Fig. 2 gave a somewhat appealing interpretation. The AUC is the probability that if you were to take a random pair of observations, one with $Y = 1$ and one with $Y = 0$, the observation with $Y = 1$ has a higher predicted probability. In short, the AUC gives the probability that the model correctly ranks such pairs of observations. Although our AUC still too close to 0.5, the ROC curve in Fig 2 is showing a trend. Indicating that the R_{diff} is a step in the right direction for a predictive model for Dota 2.

There is no one 'best' skill rating system. Every one has its flaws. Every game has a different adaptation of the Elo rating system where the margin of victory is different. Valve decides which players get teamed up and which teams get placed against each other based on more information than solely player's MMR rating system [10]. As a result, it is more likely that players with similar skill level will be present when acquiring match data. The game also gives players the option to play with friends which do not have to be of the same true skill level. The overlap of the black points between the top and bottom in Fig. 1 and our results indicate that our model cannot make accurate distinctions. This is supported by the slight S-curve in Fig. 1. A curve with a more steep coefficient is preferred. Disregarding all the different variables during these matches themselves, Our Elo calculations combined with Valve's matchmaking system influences the training data and thus the prediction model.

A. *Future Work*

The accuracy given Table I and the AUC of Fig. 2 are dependent on the manual split of the dataset. For a more precise score, it is recommended to evaluate the model using K-fold cross validation to give a better indication of the predictive performance on different subsets of data. It might be possible that we were quickly over-fitting in our results. Additionally, Learning curves in machine learning [13] can give insights on improving the manual split even better. We think that the use of a logistic regression learning curve for the training and testing iterations can be very insightful to get better predictive accuracy of the model.

Despite our results, we believe it is very likely to make a much more accurate predictions when combining Elo rating differences in a logistic regression model with more than one covariate (e.g. combining the Elo rating difference with the hero selection data that was already used in related work).

In retrospect, the methods we used could also be improved in multiple ways:

- By acquiring a bigger dataset to follow the players' Elo rating more accurately and monitoring the convergence, which should give better Elo difference measurements.
- Additionally, tests of predictive accuracy are better employed with independent datasets. Ideally, the same players should not be present in both the training and test set.
- In essence, Elo rating was invented for 1 vs. 1 games. A more accurate system for estimating a team's rating, rather than taking the average Elo of all players is highly preferred to improve prediction accuracy.
- Changing the fixed K-value to a dynamic K-value in the rating calculation based on some other values or a formula will most likely give a better representation of a player's true skill.
- Finally, disregarding the Elo rating system altogether and replacing it with a rating system more suited for teams containing different players each match might also be considered as an option (e.g. like Microsoft's TrueSkill Ranking system [14], although an implementation must be made since it is closed-source.)

The performance of our filtering was quite slow. It could be improved with parallelization across multiple CPU cores for the Filestream. Ultimately, we believe there are many promising possibilities in this particular area.

Our error values in Table I show that the reproducibility of our model with datasets of equal size. (ref. Supplementary Online Material)

VI. Conclusion

In this paper we present a prediction model to predict match outcomes for the game Dota 2. We used a logistic regression model to approach this challenge, which lead to a step in the right direction of getting predictions.

Given the results of our logistic regression model, we can conclude that solely looking at the difference of the average Elo ratings of teams, which are being updated based on the binary outcome on a match basis, is not enough information to make reliable predictions. Since we only see a slight raise in accuracy compared to a random guess, we can speculate that better methods and/or more information is needed: whether in the form of more sample data or more variables. This not only has to do with nature of matchmaking processes in Dota 2, but also with in-game variables.

In Dota 2, the rating difference is a highly significant predictor, however it remains an open question as to which covariates are needed in an logistic regression model in order to make accurate predictions. Recommendations based on our findings have been included in the discussion.

References

[1] Valve. Dota 2 blog. http://blog.dota2.com/, 2017. [Online; accessed 28-March-2017].

[2] TeamLiquid. The international 2016. http://wiki.teamliquid.net/dota2/The_International/2016, 2016. [Online; accessed 28-March-2017].

[3] Dota 2 Wiki. Heroes. http://dota2.gamepedia.com/Heroes, 2017. [Online; accessed 29-March-2017].

[4] Nicholas Kinkade and Kyung yul Kevin Lim. Dota 2 win prediction. https://cseweb.ucsd.edu/~jmcauley/cse255/reports/fa15/018.pdf, 2016. [Online; accessed 28-March-2017].

[5] Kaushik Kalyanaraman. To win or not to win? a prediction model to determine the outcome of a dota2 match. https://cseweb.ucsd.edu/~jmcauley/cse255/reports/wi15/Kaushik_Kalyanaraman.pdf, 2016. [Online; accessed 28-March-2017].

[6] Kuangyan Song, Tianyu Zhang, and Chao Ma. Predicting the winning side of dota2. http://cs229.stanford.edu/proj2015/249_report.pdf. [Online; accessed 25-March-2017].

[7] Arpad E. Elo. *The rating of chessplayers, past and present*. Arco Pub., New York, 1978.

[8] Valve. Dota 2 api documentation. https://dota2api.readthedocs.io/en/latest/, 2014. [Online; accessed 29-March-2017].

[9] Dotabuff. Frequently asked questions. https://www.dotabuff.com/pages/faq, 2017. [Online; accessed 29-March-2017].

[10] Valve. Dota 2 blog - matchmaking. http://blog.dota2.com/2013/12/matchmaking/, 2013. [Online; accessed 20-March-2017].

[11] J.M. Hilbe. *Logistic Regression Models*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2009.

[12] William H. Greene. *Econometric Analysis*. Prentice Hall, Upper Saddle River, NJ, 5. edition, 2003. pg. 55 - 161 and 211 and 229.

[13] C. Sammut and G.I. Webb. *Encyclopedia of Machine Learning and Data Mining*. Encyclopedia of Machine Learning and Data Mining. Springer US, 2017. pg. 578 - 589.

[14] Microsoft Research. Trueskill ranking system. https://www.microsoft.com/en-us/research/project/trueskill-ranking-system/, 2005. [Online; accessed 19-March-2017].

VII. Supplementary Online material

Supplementary material of our statistics, result discussion and source code (for scraping, elo calculation and analysis) can be found on Petar's Github: https://github.com/Snookik/INFOB3OMG