**The odds are with us. Or are they?**

**A regression approach to predicting a winner in Dota2**

Caleb Holdener

Belmont University

Email: [Caleb.Holdener@pop.belmont.edu](mailto:Caleb.Holdener@pop.belmont.edu)

**Abstract – In this paper multiple models were created and presented in order to predict the outcome of an multiplayer online battle arena computer game called Dota 2. Similar work has been done, however most of these models and datasets are extremely old at this point, causing them to be inaccurate for the current game state.**

**The details associated with building these models are presented, along with how the data was gathered and processed. By presenting more current issues and problems associated with modern Dota prediction algorithms, further progress can be made to develop more accurate prediction algorithms.**

1. Introduction

Dota 2, also known as Defense of the Ancients 2, is a MOBA (multiplayer online battle arena) game, which is a subgenera of strategy games. In Dota there are two teams of five players called Radiant and Dire. Each player controls, in most cases, a single unit called their Hero in order to destroy the opposing teams base, also known as their ancient. Players can choose from 112 different Heroes, each with different designs, strengths and weaknesses. Players select their Hero during a drafting phase at the beginning of the game. Once a Hero is picked, then no other player can choose that Hero. Because of this, and the uniqueness of each hero each match is unique. The goal of the project is the generation of a machine learning algorithm, that will be able to predict the outcome of a Dota 2 match based only on the Heroes that were drafted for each team.

1. Related Work

In the past, others have attempted to create a prediction algorithm, in their studies they achieved an accuracy rate of 60-70% on the high end (during earlier studies) and on the upper ends of 50% during later studies.

Various organizations, programs, and websites have also attempted to create a prediction algorithm. One of the most notable organizations/websites is OpenDota, which API was utilized in my model. OpenDota runs in collaboration with OpenAI to help run prediction algorithms during the International, which is the most prestigious Dota tournament of the year. While OpenDota does not typically have these predictions for public games, this website and API was still useful to verify match data.

1. Dataset
2. *Data Collection*

In order to perform predictive analytics, data must first be collected. The data that had to be collected must have included which players are playing on each team and what Heroes those players are playing. This was accomplished by first gathering a large number of match ID numbers. We gathered a range of match IDs (5214334377 to 5214534376). These ID numbers can be used to further query the needed features. This could be accomplished with either the OpenDota query API, or the Steam (the company who owns Dota) API. The OpenDota API was utilized for this dataset. The advantage of OpenDota API platform, is the rate limit of 1200 calls per minute, versus the rate limit of 60 calls per second with the Steam API. By using the Requests Python Library, calls were able to be made to the API in order to return full match data. during a single patch. This gave a raw dataset that could then be processed.

1. *Data Processing*

This data then had to be filtered and the correct features appended to a new dataset that would be used in the final dataset. Because the Hero data is categorial, and technically a part of 2 groups (the two different teams), in order to perform one singular regression, a new dataset containing dummy variables was created. This gave a matrix, where each row was a match and the X represented a possible hero. This is typically described as one hot encoding.

1. Methodology

A logistical regression and a deep learning model were utilized to predict the match output based on the binary matrix described in section III B. This matrix was used to train our model.

1. *Logistic Regression*

A logistic model was chosen due to its ability to predict a binary output, a simple model was created to accomplish this based on this.

* 1. *Model Creation and Tuning*

The model was creating by utilizing the scikit library. Initially by using the default cross validation, upwards of 1.2% accuracy score was lost during validation. In order to better train our model a variety of parameters were added to attempt to increase the accuracy, specifically by increasing the C parameter, and by changing the cross validation method to StratifiedKfold. However, the highest validation score I could achieve was within 1.1% of the model.

* 1. *Analysis*

Based on the model created we can conclude that currently in Dota, pure draft without any other factors (features) is not very important to the outcome of the game. While some heroes had a high impact on the win rate, because of the lower sample, with both individual heroes, and overall matches.

1. *Deep Learning*

A deep learning model was chosen because it is an emerging field of computer science research. TensorFlow is a leader in deep learning model creations and their Keras API allows for highly customized models when compared to other libraries.

* 1. *Model Creation and Tuning*

The model was created using Keras API in Tenserflow 2.0. A network of two hidden dense layers was created and an output layer using the Sigmoid Activation function, which gives a range between 0 and 1, perfect for predicting the probability of a binary output. When analyzing initial model prototypes, it became obvious that the model was overfitting the training data, therefor leading to lowered validation accuracy. To better tune our model and prevent overfitting, an early stopping callback which monitored validation loss was introduced. In addition, model complexity was lowered, leading to a more accurate and faster model.

* 1. *Analysis*

Using a tuned deep learning model, I was able to achieve, on average, an even – if not slightly higher, accuracy compared to the logistic regression mode. Even with a relatively simple model, we were able to achieve a consistent 56% accuracy, which was 4% higher than blind guessing the winner.

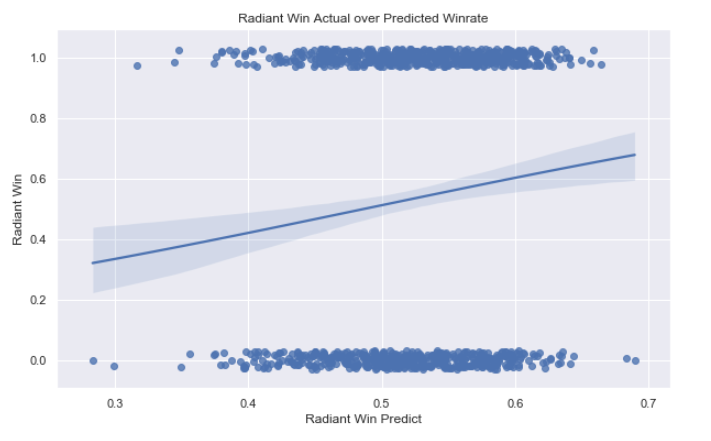


Fig. 1. Radiant Win Actual over Predicted Winrate

1. Conclusion

In this report, logistic regression and deep learning models were used in an attempt to predict match outcomes based only on the heroes picked at the start of the game. Using regression, an accuracy score of ~55% was achieved, while deep learning gave an accuracy score of ~56%.

The accuracy score of ~55-56% was only 3-4% higher than predicting the Radiant each time (which had an average win rate of 52%). This is lower than similar models utilizing older datasets. Two main options that could cause this, the dataset used was flawed, or in more recent patches Dota has become more balanced between both the heroes and teams. Given the results of our model however, it can be concluded that the current model (given the current dataset) is only an average model for predicting win-rate.

1. Future Work