**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)

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 with five players each. Each player controls, in most cases, a single unit called their Hero in order to destroy the opposing teams base. Players can chose from over 110 different Heroes, each with different designs, strengths and weaknesses. Players select their Hero during a drafting phase at the beginning of the game. The goal the project, is the generation of a linear based model, 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% 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, the API was useful to verify match data.

1. Dataset
2. *Data Collection*

In order to perform predictive analytics, data must first be collected. The data that will be collected must include which players are playing on each team and what Heroes those players are playing. This was accomplished using an existing database of matches that is collected and stored by this company called OpenDota. This data is open to the public and can be easily gathered through their database explorer. Matches were able to be queried through this explorer by utilizing SQL (Structured Query Language). I pulled 20000 matches that were ranked all-pick games (meaning every hero was able to be picked). This gave the least biased dataset. The matches pulled from the explorer however, didn’t have any of the required features I needed, so I needed additional processing on the data.

1. *Data Processing*

The pulled data contained identifying ‘Match ID’ numbers that unique to every match played. This ID number 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. This data 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.

1. Methodology

A logistical regression model was 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. Model Creation

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.

1. Model Tuning

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. Conclusion

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

The accuracy score of ~55% was only 3% higher than predicting the Radiant each time (which had an average win rate of 52%). This is lower than 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.