Dota 2 Match Outcome Classifier Using Hero Selection

Josh Morgan
joshmorg1007@uvic.ca

December 13, 2022

Abstract

In this work several different models are trained using a variety of machine learning algorithms and methods with the aim of creating an accurate classification model for determining the winner of Dota 2 match based solely on each team's selection of 5 heroes. Both a newly generated dataset of matches, and an original dataset from a paper tackling the same problem in 2013 [1]. For both datasets, the best performing model was a simple one layer neural network with an accuracy of 59.42% for the new dataset, and 71.07% for the original dataset. The model for the original dataset outperforms the best performing model of the original paper, which achieved an accuracy of 69.8% using logistic regression. The discrepancy between models trained on the original dataset, and the new dataset, though inconclusive, is likely due to 9 years of gamplay updates reducing the importance of hero selection when determining the outcome of a match.

1 Preface

Dota 2 is a popular multiplayer online battle arena (MOBA) game developed and published by Valve Corporation. The game is played by two teams, of five players, called "Radiant" and "Dire", with each team occupying and defending their own separate base on the map. The goal of the game is to destroy the enemy team's base, called the "Ancient", while defending your own. Each player controls a unique character from a pool of 123, called a "hero", with their own set of abilities and playstyle. The game is known for its high skill ceiling and deep strategic gameplay. It also has a thriving professional esports scene, with large tournaments offering millions of dollars in prize money, with the latest tournament having a prize pool of over 18 Million USD [2]. Dota 2 continues to be a popular and engaging game for both casual and competitive players.

Each player on a given teams plays a certain role. The roles are as follows Cary, Mid, Offlane, Soft Support and Hard Support. They are often referred to as a number from 1 to 5, using the order mentioned before, where the number loosely refers to that players priority in terms of allocated game resources.

1.1 The Importance of Hero Selection

The drafting phase in Dota 2 is an incredibly important part of the game, as it allows players to strategically choose their hero and counter the enemy team's picks. A well-planned draft can give a team a significant advantage and set the stage for a successful game. A model that accurately predicts the outcome of a match based on the hero selections of each team could prove beneficial to both casual an competitive players. Tools using such a model, like a hero recommender, or a draft simulator, could be incredibly beneficial. These tools could provide suggestions for hero combinations and counter picks, as well as allowing players to practice and refine their draft strategy. This could not only improve individual performance, but also enhance team coordination and communication during the drafting phase. Overall, a model that predicts the outcome of a match based solely on the hero selection of both teams could prove extremely beneficial.

1.2 Game Complexity

Dota 2 is well known for being an incredibly complex game and is classified as such by the machine learning community [3]. With this in mind, hero selection, although important, can only contribute so much towards the outcome of a given match. There are many other variables that contribute, such as individual players skill/experience/comfort with their selected hero, the players choice of itemization from the pool of 200+ items [1], the order in which a player chooses to level up their skills, or the players choice of talent [2].

2 Problem

In this paper, ten different classification models were trained with the goal of providing an accurate prediction of a Dota 2 match based on each teams choice of 5 unique heroes from the pool of 123 using data obtained from Valve's publicly available match data API.

As many of the heroes in Dota 2 are very versatile, meaning that they can can be played in multiple different roles. Valve's API doesn't provide information as to which role a given hero was being played in, therefore a third party API provided by STRATZ [4] was utilized in order to get this information. The STRATZ API parses the replay of a given match in order to extrapolate extra information, such as what role a given hero was playing. The hope of having this extra role information contained within the dataset is to further the accuracy of the trained models.

3 Related Work

Dota 2 has been the subject of a number of machine learning research. The most high profile being OpenAi's Five project [3] where a team of 5 neural networks were trained in order to play the game and able to compete with amateur players, although with some restrictions in place. A hero recommendation engine using a classification model to use hero selection to determine the outcome of a match has also been explored previously by Kevin Conley and Daniel Perry[1]. In their paper they claim that they were able achieve 69.8% accuracy using logistic regression and 67.43% accuracy using K-nearest neighbours.

The goal of this paper was to replicate the findings of Conley and Perry and produce a model at least as accurate, while reflecting the current state of the game. Since the original paper was published in 2013, there have been 17 new heroes added to the game, 41 major gameplay balance updates, as well as many more small gameplay balance updates.

[1] Items are bought by the player using in game currency earned throughout the match called "gold" and are used to cover the player's hero's weaknesses, enhance their strength, defend against an enemy hero's strengths, or exploit an enemy hero's weakness

[2] a hero's talent is an additional improvement to a hero's base stats, or ability stats that the player gets to choose upon reaching certain level milestones. Upon reaching one of these milestones, the player gets to choose one from a set two talents

4 Dataset

Initially, only the new data gathered using Valve's API and the STRATZ API was going to be explored, however after achieving results substantially lower than the findings of Conley and Perry, their original dataset was also explored for comparison of models using the same methodologies.

4.1 Dataset Structure

The dataset consists of 10 categorical features and a binary label indicated whether the Radiant team was the winner of the match. Each of the categorical features in a given sample is the name of one of the 123 heroes, or 106 heroes in the case of the original dataset, within the game. The libraries used in order to train these models requires that categorical features be One-Hot Encoded, which ballooned out the dimentionality dramatically, increasing the runtime of some of the algorithms used to train the models.

In order to minimize some of other external factors, such as player experience/skill with the game, matches are only considered if they are tagged as "Very High Skill" [3] by the Valve API. At the time when the original dataset was collected, Conley and Perry claim that this tag roughly indicates the top 8% [1], however now this tag is closer to indicating the top 12% [5]. Additionally, a match was discounted from the dataset if a player had quit the match before completion.

4.2 Differences Between Original Dataset and New Dataset

The original dataset, follows the structure outlined above, with each categorical features having 106 different classes representing each of the heroes in the game at the time. The first 5 categorical features (before One-Hot Encoding) represent the Radiant teams, and the last 5 features represent the Dire team. Between the 5 features for both teams, there is no order between them. There are around 48,000 samples, with 61.5% having Radiant labeled as the winner, and 38.5% having Dire labeled as the winner. This distribution is curious as it doesn't seem reflect the distribution of actual matches at the time, at least in from the limited records still available. This could indicate a error in the way that the original dataset was sampled.

The newly collected dataset also follows the structure outline above, but with each categorical feature having 123 different classes representing each of the heroes currently in the game. The features are ordered in the same way as the original dataset, however, the order within a team's features now holds additional role information gathered by the STRATZ API. Due to time and API constraints, the new dataset only has around 23500 samples, with 53.4% having Radiant labled as the winner, and 47.6% having Dire labeled as the winner.

4.3 Training and Validation Splits

For both the new and old dataset, the following methodology was used to train and validate each model. Initially 10% of the dataset was put to the side, reserved for the final validation of the models. For each of type of model trained, the remaining 90% was split into training and testing data using a 70% - 30% split. After all hyper parameters were selected, the models were evaluated using the 10% put to the side.

[3] This tag indicates that the average MMR or ELO ranking of the players in the match is above 3800.

Table 1: Best Decision Tree Model Performance

	New Dataset	Original Dataset
Training Accuracy	0.6571	0.6571
Validation Accuracy	0.5478	0.6258

Table 2: Best Random Forest Model Performance

	New Dataset	Original Dataset
Training Accuracy	0.6176	0.6571
Validation Accuracy	0.5309	0.6258

5 Models

Learning curve figures omitted in order to meet 6 page requirement. The relevant figures will be submitted along side report and code.

5.1 Decision Tree

The best performing Decision Tree model was selected by using sklearn's GridSearchCV class using 5-fold cross validation. The best performing model for new dataset was trained using entropy as the criterion, and having a max depth of 30. The best performing model for the original dataset was trained using gini as the criterion, and having a max depth of 10. The training and validation accuracies are summarized in Table 1.

5.2 Random Forest

The best performing Random Forest model was selected by using sklearn's GridSearchCV class using 5-fold cross validation. The best performing model for new dataset was trained using gini as the criterion, and having a max depth of 20, with 100 estimators. The best performing model for the original dataset was trained using gini as the criterion, and having a max depth of 20, with 10 estimators. The training and validation accuracies are summarized in Table 2.

5.3 Logistic Regression

The best performing Logistic Regression model was selected by using sklearn's GridSearchCV class using 5-fold cross validation. The best performing model for new dataset was trained using a C of 0.5, Class Weight set to None, using the newton-cg solver . The best performing model for the original dataset was trained using a C of 0.5, Class Weight set to None, using the saga solver. The training and validation accuracies are summarized in Table 3.

5.4 Naive Bayes

Three different algorithms for Naive Bayes were considered, Multinomial Naive Bayes, Bernoulli Naive Bayes, and Gaussian Naive Bayes. For both datasets, Multinomial and Bernoulli Naive Bayes performed

Table 3: Best Logistic Regression Model Performance

	New Dataset	Original Dataset
Training Accuracy	0.7771	0.7314
Validation Accuracy	0.5754	0.7126

Table 4: Naive Bayes Model Performance

	New Dataset			Original Dataset		
	Multinomial	Bernoulli	Gaussian	Multinomial	Bernoulli	Gaussian
Training Accuracy	0.9101	0.9005	1.0	0.7392	0.7388	0.7002
Validation Accuracy	0.5793	0.5771	0.4835	0.7123	0.7145	0.6745

very similar, while Gaussian Naive Bayes performed alright for the original dataset, although still worse than Multinomial and Bernoulli, but worse than the baseline of 53% for the new dataset. The training and validation accuracies are summarized in Table 4.

5.5 Neural Networks

Four different neural network architectures were explored, with their hyper parameters manually tuned. All models used Binary Cross Entropy as their loss function, and the Adam optimizer with varying learning rate as their optimizer. The first was a simple one hidden layer network, using a learning rate of 0.0002, 100 nodes in the hidden layer, and using the sigmoid activation function. The second model expanded to 2 hidden layers, still using a learning rate of 0.0002 and the sigmoid activation function, however the nodes per hidden layer were expanded to 200, and a dropout layer was added after each hidden layer with a dropout rate of 0.2. The third model is identical to the secound, other than the learning rate was reduced to 0.00002 and the activation function used for the hidden layers was the selu function. The final model was an ensemble model of 3 of the one hidden layer models described, one trained purely on the Radiant team's hero choice, one trained purely on the Dire team's hero choice, and one trained on the full data. The training and validation accuracies are summarized in Table 5

Table 5: Neural Network Model Performance

	New Dataset		Original Dataset	
	Training Accuracy	Validation Accuracy	Training Accuracy	Validation Accuracy
One Layer	0.7023	0.5671	0.7345	0.7098
Multi-Layer Sigmoid	0.6302	0.5721	0.7134	0.7021
Multi-Layer Selu	0.7039	0.5842	0.7123	0.7010
Ensemble	0.7049	0.5809	0.7230	0.6983

Table 6: Final Validation of Models

	New Dataset Final Accuracy	Original Dataset Final Accuracy
Decision Tree	0.5556	0.6433
Random Forest	0.5449	0.6370
Logistic Regression	0.5807	0.7059
Multinomial NB	0.5934	0.7059
Bernoulli NB	0.5892	0.7059
Guassuab NB	0.4717	0.6692
One Layer NN	0.5943	0.7107
Multi-Layer Sigmoid NN	0.5598	0.7098
Multi-Layer Selu NN	0.5862	0.7061
Ensemble NN	0.5926	0.7098

6 Final Evaluation

Using the 10% of each dataset that was set aside, the final evaluation of the models accuracies are summarized in Table 6. The best performing model on the new dataset was the one layer neural network, achieving an accuracy of 59.43%, the best performing model on the original dataset was also the one layer neural network, however it was able to achieve an accuracy of 71.07%. Even when accounting for the difference in the baseline accuracy due to the dataset label distribution, the original dataset models outperforms the new dataset by around 3.5% for the best models. This discrepancy could be due to a few factors. Firstly it could be due to the potential sampling error mentioned in section 4.2, where some bias could have been introduced into the original dataset. Another reason is that is could be simply due to the difference quantities of training data, the original data has over double the number of samples of the new dataset due to time and API limitations. The final, and most likely, explanation is that since 2013 the game's balance has shifted to a point where hero selection is less important than other factors when determining the outcome of a match due to the 41 major and many more minor gameplay balance updates.

References

[1] Daniel Perry Kevin Conley. *How Does He Saw Me? A Recommendation Engine for Picking Heroes in Dota 2*. `https://kevintechnology.com/2013/12/30/using-machine-learning-to-recommend-heroes-for.html`. 2013.

[2] Liquipedia. *The International 2022*. `https://liquipedia.net/dota2/The_International/2022`. 2022.

[3] OpenAI. *OpenAI Five*. `https://blog.openai.com/openai-five/`. 2018.

[4] STRATZ. *STRATZ API*. `https://stratz.com/api`. 2022.

[5] esportstales. *DOTA SEASONAL RANK DISTRIBUTION AND MEDALS - OCTOBER 2022*. `https://www.esportstales.com/dota-2/seasonal-rank-distribution-and-mmr-medals`. 2022.