E-Sports Game Winner Prediction : Dota2

Sandeep Mysore*
skm62@pitt.edu
University of Pittsburgh
Pittsburgh, PA, USA

Yangyin Wang†
yaw79@pitt.edu
University of Pittsburgh
Pittsburgh, PA, USA

ABSTRACT

This paper is a documentation of the application of Data Mining Techniques and Algorithms to predict the winner of a MOBA (Multiplayer Online Battle Arena) E-Sports Game, Dota2 using in-game statistics. The paper details the methodology used to tackle this problem from initial Data Sourcing from Kaggle to Data Cleaning, Manipulation and Preparation.It also details the modelling, evaluation and validation methodologies used in the project.

CCS CONCEPTS

• **Computing methodologies** → **Classification and regression trees**; **Support vector machines**; • **General and reference** → **Evaluation**; **Validation**; *Performance*.

KEYWORDS

dataset, prediction, classification, clustering , regression, validation, data mining

ACM Reference Format:
Sandeep Mysore and Yangyin Wang. 2020. E-Sports Game Winner Prediction : Dota2. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

E-Sports are the new fad especially with the onset of the pandemic. MOBA (Multiplayer Online Battle Arena) games are a way to interact and have fun with friends while still being quarantined at home. These MOBA games are a subgenre of strategy video games in which each player controls a single character with a set of unique abilities that improve over the course of a game and which contribute to the team's overall strategy. The ultimate objective is for each team to destroy their opponents' main structure, located at the opposite corner of the battlefield. However, MOBA games can have other victory conditions, such as defeating every player on the enemy team. Dota2 is played in matches between two teams (Radiant and Dire) of five players, with each team occupying and defending their own separate base on the map. Each of the ten players independently controls a powerful character, known as a "hero", who all have unique abilities and differing styles of play. During a match players collect gold, experience points and items for their heroes to successfully defeat the opposing team's heroes in player versus player combat. A team wins by being the first to destroy the other team's "Ancient", a large structure located within their base.

*Both authors contributed equally to this research.
†Both authors contributed equally to this research.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 PROBLEM STATEMENT

The problem at hand is to predict the winner of a Dota2 game based on in-game statistics like gold and experience (Xp) accumulated by each team at different time intervals.

3 APPROACH

We have opted to use Data Mining Techniques to build a model with the aim of tackling our problem. The steps for our approach:

- Source the data
- Clean and manipulate the data
- Build our model
- Train and test our model
- Evaluate and Validate the results The details of our approach can be seen below.

4 DATA

The dataset we used has been sourced from Kaggle. The link to the dataset is given below. https://www.kaggle.com/devinanzelmo/dota-2-matches.

4.1 Data Description

The data sourced is approximately of size 1.45 GB consisting of 20 csv files each with different aspects of the game. We, to tackle our problem have mainly used two of the csv files, i.e match.csv and player_time.csv.

4.2 Data Manipulation

The reason we have opted to consider the files match.csv and player_time.csv in our dataset is primarily due to the fact that these files consist of in-game statistics and information regarding the gold and experience of each team along with the details of the winning side. The reason we have not considered the other files are due to the fact that we do not intend to use player skill levels or team communication details for prediction as it is a known fact that higher skilled players usually end up winning the game. This was a KDD (Knowledge Driven Decision). Also, this has already been worked on before and we intended to not let player experience affect our prediction.

4.3 Data Cleaning

- Data corresponding to Ranked matches has been considered and all other data has also been removed
- We have removed all 0 or NAN values in our chosen files
- We have removed attributes like Hero Damage, KDA stats (Kill-Death-Assist) and tower damage as they directly correspond to the gold and experience accumulated by the players
- Player Votes has also been removed as they are influenced by rage and don't affect the game outcome

4.4 Data Preparation

- The gold distribution of members in each team is calculated
- The Experience distribution of members in each team is calculated
- The gold advantage is calculated
- The Experience advantage is calculated
- The Team advantage is therefore calculated
- Match Status at the 15 minute time interval is found
- All the calculated data is used and merged with the match result

4.5 Data Split

- Training using 60% of the data randomly sampled
- Preliminary testing using 40% of the data randomly sampled

4.6 Data Visualised

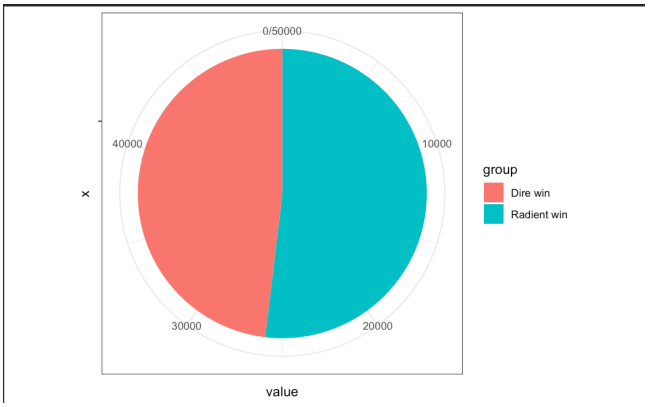

Figure 1: Data Split Visualisation

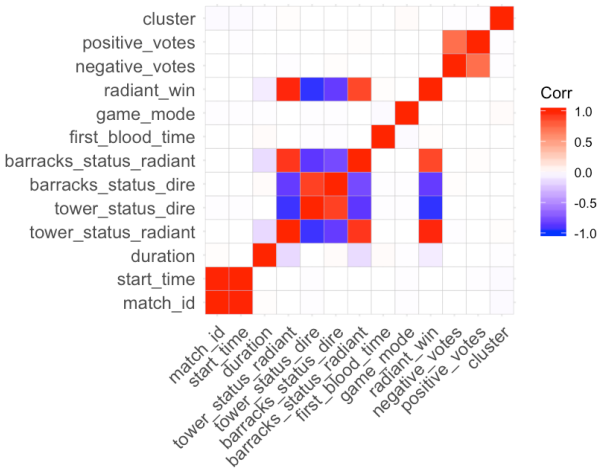

Figure 2: Corr plot of initial Data

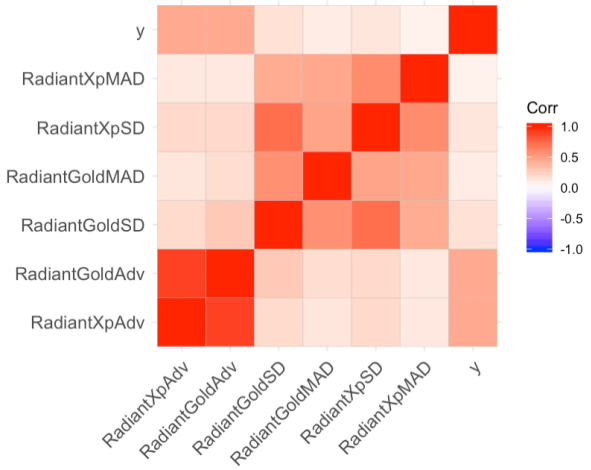

Figure 3: Corr Plot After Cleaning

5 BUILDING OUR MODEL

We built multiple models to compare performances of our prediction

5.1 Models

- Logistic Regression
- K-Nearest Neighbor (KNN)
- Naive Bayes
- Decision Tree
- Support Vector Machine (SVM)

5.2 Model Validation

- 10 fold Cross Validation is used to validate the model

5.3 Model Evaluation Metrics

- 1 - Error
- Precision
- Recall
- F-Score
- Area Under the ROC curve (AUC)

6 RESULTS

The results of each of our models based on our Evaluation Metrics are tabulated and the ROC curves for each of the models is displayed.

Table 1: Model Evaluation Results

Models	1 - Error	Recall	Precision	F-Score	AUC
L-Regression	0.3691	0.6261	0.7182	0.5945	0.7619
Naive Bayes	0.3747	0.6207	0.7067	0.5879	0.7507
KNN	0.5002	0.4998	0.4406	0.4552	0.4991
Decision Tree	0.4140	0.6029	0.6298	0.5777	0.6797
SVM	0.3762	0.6194	0.6955	0.5858	0.7479

6.1 Result - ROC Curves

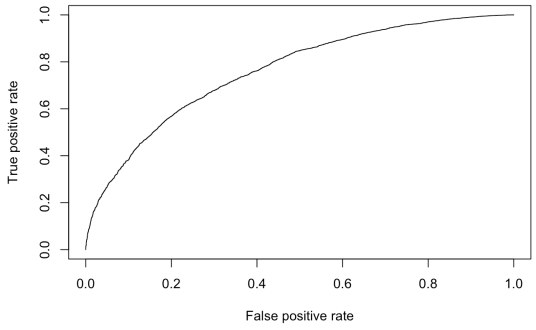

Figure 4: Logistic Regression ROC Curve

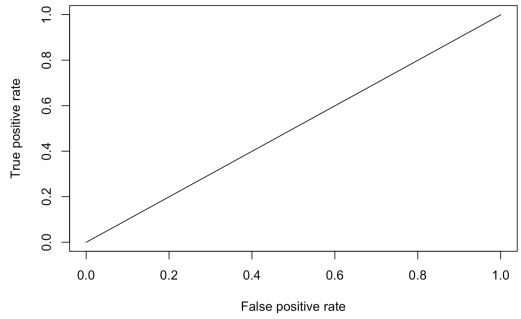

Figure 5: K-Nearest Neighbor ROC Curve

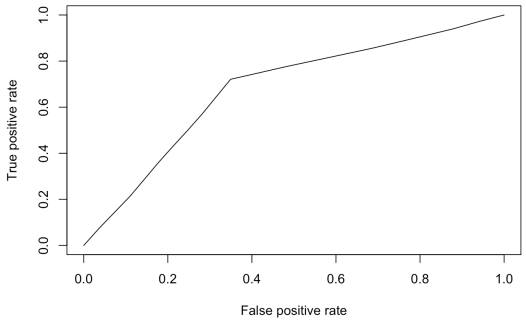

Figure 6: Decision Tree ROC Curve

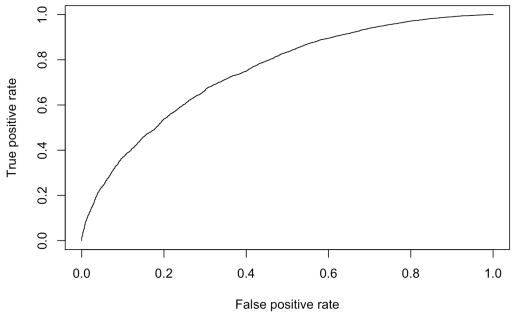

Figure 7: Naive Bayes ROC Curve

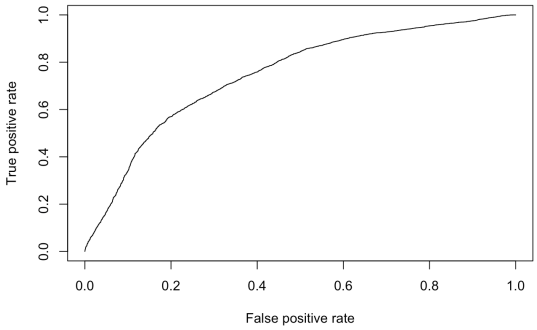

Figure 8: Support Vector Machine ROC Curve

7 CONCLUSION

Through extensive Modelling, Testing, Evaluating and Validating we have arrived at the conclusion that the Logistic Regression Model performs the best giving an Area Under the ROC Curve (AUC) value of 76.19% and a Precision value of 71.82%. In a game with high variability and high intricacy, depending on player skills,

this model arrives at a good result considering the fact that player skill was not taken into consideration as we aimed at predicting solely on in-game statistics alone. We would like to conclude by quoting a hero in the game *Invoker*

> "I am a beacon of knowledge blazing out across a black sea of ignorance"

The reference to the knowledge is the use of Data Mining Techniques to better understand and predict the game result in a vast expanse unpredictably variable game mechanics with high intricacies.

8 FUTURE WORK

In the future, we can see that the model can be optimised by adding data like tower status and First Blood (First kill of the game) Stats to the training of the model.Adding player skill rating and player communication data can also help in optimising the model. We aimed at solving the problem in a particular way, and we have accomplished the task at hand with the method decided upon.

REFERENCES

[1] https://www.kaggle.com/devinanzelmo/dota-2-matches
[2] https://www.kaggle.com/parshakov/does-communication-matter
[3] https://www.kaggle.com/brassmonkey381/first-team-to-kill-a-barracks-wins
[4] https://www.kaggle.com/abenahmed1/setting-up-a-prediction-problem-dota-2
[5] https://www.kaggle.com/semenoffalex/does-match-duration-affect-dire-radiant-winrate
[6] https://www.kaggle.com/davidmercury/hero-picking-and-match-result
[7] https://www.kaggle.com/devinanzelmo/a-quick-look-at-dota-2-dataset
[8] https://en.wikipedia.org/wiki/Dota_2
[9] https://www.reddit.com/r/DotA2/comments/7x6xi9/whats_your_favorite_quote_in_dota_2/
[10] https://mattscradle.com/top-dota-2-heroes-impressions-during-battles/
[11] https://en.wikipedia.org/wiki/Multiplayer_online_battle_arena