Exploring Key Variables Influencing the Outcome of Esports Matches

Submitted by

M. Hisham Zahid
23L-7827

Supervised by

Dr. Asif Mahmood Gilani

A thesis submitted in partial fulfillment of the requirements for the degree of
Masters (Computer Science)
at National University of Computer & Emerging Sciences

School of Computing
National University of Computer & Emerging Sciences

Lahore, Pakistan.

September 2024

Plagiarism Undertaking

I take full responsibility of the research work conducted during the Masters Thesis titled *Exploring Key Variables Influencing the Outcome of Esports Matches*. I solemnly declare that the research work presented in the thesis is done solely by me with no significant help from any other person; however, small help wherever taken is duly acknowledged. I have also written the complete thesis by myself. Moreover, I have not presented this thesis (or substantially similar research work) or any part of the thesis previously to any other degree awarding institution within Pakistan or abroad.

I understand that the management of National University of Computer and Emerging Sciences has a zero tolerance policy towards plagiarism. Therefore, I as an author of the above-mentioned thesis, solemnly declare that no portion of my thesis has been plagiarized and any material used in the thesis from other sources is properly referenced. Moreover, the thesis does not contain any literal citing of more than 70 words (total) even by giving a reference unless I have the written permission of the publisher to do so. Furthermore, the work presented in the thesis is my own original work and I have positively cited the related work of the other researchers by clearly differentiating my work from their relevant work.

I further understand that if I am found guilty of any form of plagiarism in my thesis work even after my graduation, the University reserves the right to revoke my Masters degree. Moreover, the University will also have the right to publish my name on its website that keeps a record of the students who plagiarized in their thesis work.

____

M. Hisham Zahid

Date: ______________________

Author's Declaration

I, M. Hisham Zahid, hereby state that my Masters thesis titled *Exploring Key Variables Influencing the Outcome of Esports Matches* is my own work and it has not been previously submitted by me for taking partial or full credit for the award of any degree at this University or anywhere else in the world. If my statement is found to be incorrect, at any time even after my graduation, the University has the right to revoke my Masters degree.

M. Hisham Zahid

Date:

Certificate of Approval

It is certified that the research work presented in this thesis, entitled "Exploring Key Variables Influencing the Outcome of Esports Matches" was conducted by M. Hisham Zahid under the supervision of Dr. Asif Mahmood Gilani .

No part of this thesis has been submitted anywhere else for any other degree.

This thesis is submitted to the School of Computing in partial fulfillment of the requirements for the degree of Mastersin Computer Science

at the

National University of Computer and Emerging Sciences, Lahore, Pakistan

September' 2024

Candidate Name: M. Hisham Zahid Signature: ____________________

Examination Committee:

1. Name: Examiner 1 Signature: ____________________
   Professor, FAST-NU Lahore.
2. Name: Examiner 2. Signature: ____________________
   Professor, FAST-NU Lahore

Dr. Kashif Zafar ____________________________
Head of the School of Computing, National University of Computer and Emerging Sciences, Lahore, Pakistan.

Abstract

Dota 2, a premier title in the esports landscape, presents a complex and strategically rich environment where the interplay between team compositions, individual player performances, and in-game decisions critically impacts match outcomes. As the game's meta evolves, understanding the key variables that contribute to victory has become increasingly important for teams, analysts, and the broader esports community. This research leverages a dataset of approximately 180,000 professional matches (from 2014 to 2024) to analyze these variables, employing machine learning and statistical methods to uncover patterns and predictors of success. By developing predictive models that incorporate both pre-game factors (such as hero drafts) and in-game dynamics (such as player metrics, first bloods, and net worth), this study aims to provide actionable insights that can enhance decision-making in drafting and gameplay. Although centered on Dota 2, the variables and methods explored in this research are common across many MOBA (Multiplayer Online Battle Arena) games, making the findings potentially relevant to a broader context. The results have the potential to inform better strategies and contribute to the ongoing development of data-driven approaches in competitive esports.

Acknowledgements

I would like to express my heartfelt gratitude to my supervisor, Dr. Asif Mahmood Gilani, for their guidance and support throughout this thesis. I am also thankful to my family and friends for their unwavering encouragement, and to all who contributed to this research.

Dedication

This thesis is dedicated to my beloved family, whose unwavering support and encouragement have been my greatest source of strength. To my friends and mentors, thank you for your inspiration and guidance throughout this journey.

Contents

Chapter 1

Introduction

Dota 2, one of the most prominent titles in the esports industry, presents a complex strategic environment where team compositions, player actions, and in-game decisions significantly impact match outcomes. This research seeks to analyze these factors by leveraging a comprehensive dataset of professional Dota 2 matches, applying advanced machine learning and statistical techniques to predict match results. While the study is focused on Dota 2, the variables we examine—such as player performance metrics, in-game events, and team dynamics—are not unique to this game. These variables are commonly found across other MOBA (Multiplayer Online Battle Arena) games, making our findings potentially relevant and applicable beyond just Dota 2.

1.1 Importance

The ability to predict match outcomes in Dota 2 is of immense strategic value. Accurate predictions can influence key decisions during drafting, team coordination, and in-game strategy, potentially giving teams a competitive edge. As the esports industry continues to grow, data-driven approaches like the one explored in this research are becoming indispensable for teams looking to optimize their performance.

1.1.1 Strategic Value

In high-stakes professional matches, the difference between victory and defeat can hinge on minute decisions and strategies. By identifying and understanding the variables that most strongly correlate with match outcomes, teams can make more informed choices, leading to improved consistency and success.

1.1.2 Data-Driven Decision-Making

The integration of data analytics into esports strategy is transforming the way teams approach competition. This research leverages a large dataset of professional Dota 2 matches, focusing on variables that are significant within the game and that are also common in other similar games, providing insights that are theoretically significant and practically applicable in real-world scenarios.

1.2 Related Works

Research into outcome prediction in Dota 2 and similar games has explored a variety of machine learning models and statistical methods. Previous studies have demonstrated that variables like team compositions, early-game performance, and player statistics can be powerful predictors of match outcomes.

1.2.1 Machine Learning Approaches

Several studies have applied machine learning models to predict outcomes in Dota 2 matches. Wang et al. (2018) [1] employed Random Forests and Support Vector Machines (SVMs) to capture complex interactions between variables like team compositions and player statistics, demonstrating high accuracy in predicting match results. Similarly, Akhmedov and Phan (2021) [2] explored various algorithms including SVMs and Gradient Boosting to analyze in-game features and historical data, focusing on feature engineering to enhance model performance.

1.2.2 Predictive Modeling Techniques

Different modeling techniques have also been used to analyze Dota 2 matches. Zhang et al. (2020) [3] introduced a bidirectional Long Short-Term Memory (LSTM) model for improving hero lineup recommendations. This model accounted for the sequential nature of the draft phase, offering a nuanced understanding of how hero picks influence subsequent selections. Song et al. (2019) [4] developed a predictive model using logistic regression and decision trees to determine the winning side of matches, highlighting the role of metrics like gold and experience differences.

1.2.3 Naïve Bayes and Team Draft Analysis

Wang and Shang (2017) [5] utilized a Naïve Bayes classifier to predict match outcomes, emphasizing the model's simplicity and interpretability. Their study focused on features such as hero pick rates and early game performance. Additionally, Grutzik et al. (2017) [6] examined the draft phase using machine learning models to predict match outcomes based on hero selections, underscoring the strategic importance of team compositions.

1.2.4 In-Game Data and Death Prediction

Katona et al. (2018) [7] developed a model to predict imminent player deaths in Dota 2 based on real-time in-game data. Their research highlighted the importance of situational awareness and reactive decision-making in avoiding deaths and maintaining team stability.

Chapter 2

Methodology

This chapter outlines the research methodology employed to analyze the dataset of approximately 180,000 professional Dota 2 matches. The methodology involves data preprocessing, exploratory data analysis, model development, and validation, with a focus on variables that are broadly relevant in the MOBA genre.

2.1 Dataset Description

The dataset used for this research consists of a comprehensive collection of approximately 180,000 professional Dota 2 matches, each represented by a wide array of variables. These variables encompass various aspects of the game, from match metadata to detailed player performance metrics. The dataset includes the following key columns:

- **league_id, league_tier, league_start_date_time, league_end_date_time:** These columns provide information about the league in which the match took place, including the league's unique identifier, its competitive tier, and the start and end dates of the league.
- **series_id, series_type:** These fields denote the series in which the match occurred, including the series' unique identifier and type (e.g., best of three).
- **match_id, match_start_date_time, match_duration_seconds:** This set of columns captures the specific match details, including the unique match identifier, the start time of the match, and its duration in seconds.
- **first_blood_time_seconds:** This column records the time, in seconds, at which the first blood (first kill) occurred during the match.

- **radiant_team_id, radiant_team_name, dire_team_id, dire_team_name, winner_id:** These columns provide details about the teams involved in the match, including their identifiers, names, and the winning team.
- **radiant_kills, dire_kills:** The total number of kills achieved by the Radiant and Dire teams during the match.
- **radiant_player_X_id, radiant_player_X_name, radiant_player_X_hero_id, radiant_player_X_hero, radiant_player_X_position, radiant_player_X_lane, radiant_player_X_role, radiant_player_X_kills, radiant_player_X_deaths, radiant_player_X_assists, radiant_player_X_networth:**
  These fields detail the performance metrics of each of the five players on the Radiant team. They include each player's ID, name, chosen hero, in-game position, lane, role, and statistics such as kills, deaths, assists, and net worth.
- **dire_player_X_id, dire_player_X_name, dire_player_X_hero_id, dire_player_X_hero, dire_player_X_position, dire_player_X_lane, dire_player_X_role, dire_player_X_kills, dire_player_X_deaths, dire_player_X_assists, dire_player_X_networth:**
  Similar to the Radiant team, these columns capture the same performance metrics for each of the five players on the Dire team.
- **game_version_id:** This column indicates the version of the game during which the match was played, which is essential for understanding the context of the match, as gameplay mechanics and balance can change between versions.

This dataset provides a rich source of information for analyzing the factors that influence match outcomes in professional Dota 2 games. By leveraging these detailed variables, this research aims to develop robust predictive models that can accurately forecast match results.

2.2 Data Preprocessing

Data preprocessing is a critical step in ensuring the accuracy and reliability of the analysis. This involves cleaning the dataset by handling missing values, normalizing variables, and encoding categorical data. Additionally, feature engineering will be conducted to create new variables that may offer additional insights into match outcomes.

2.2.1 Handling Missing Data

Missing data will be addressed using imputation techniques or by excluding incomplete records where necessary. The goal is to maintain a robust dataset that accurately represents the variables of interest without introducing bias.

2.2.2 Normalization and Encoding

Normalization of numerical variables ensures that they are on a comparable scale, which is crucial for the effectiveness of machine learning models. Categorical variables, such as hero picks and item choices, will be encoded using techniques like one-hot encoding, allowing them to be used in the analysis.

2.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) will be used to identify key features and patterns that influence match outcomes. This step includes visualizing data distributions, examining correlations between variables, and identifying outliers or anomalies.

2.3.1 Feature Selection

Key features that are likely to influence match outcomes, such as gold difference, experience difference, and hero win rates, will be selected for further analysis. This process will involve both domain expertise and statistical techniques like correlation analysis and principal component analysis (PCA) to ensure that the most relevant features are considered.

2.3.2 Pattern Recognition

Patterns in the data, such as trends in hero picks or the impact of early-game events, will be analyzed using visualization techniques like heatmaps, scatter plots, and time-series analysis. Identifying these patterns is crucial for understanding the factors that drive match outcomes and for guiding the development of predictive models.

2.4 Model Development

Various statistical and machine learning models will be developed to predict match outcomes. These models will include traditional methods like logistic regression and decision trees, as well as more advanced techniques such as Random Forests, SVMs, and Gradient Boosting Machines (GBMs).

2.4.1 Model Selection

Multiple models will be trained and evaluated based on their accuracy, precision, recall, and other performance metrics. The most effective models will be selected for further refinement and testing, with a focus on their ability to generalize to unseen data.

2.4.2 Feature Importance

The importance of different features in predicting match outcomes will be assessed using techniques such as feature importance scores and SHAP values. This analysis will highlight which variables have the most significant impact on match results, providing valuable insights for both the development of the models and their practical application.

2.5 Validation

The predictive models will be validated using a subset of the data that was not used during training. Cross-validation techniques, such as k-fold cross-validation, will be employed to ensure that the models are robust and perform well on unseen data.

2.5.1 Cross-Validation

Cross-validation will involve splitting the dataset into multiple folds to test the models on different subsets of the data. This technique helps prevent overfitting and ensures that the models generalize well to new matches.

2.5.2 Model Evaluation

Model performance will be evaluated based on metrics like accuracy, precision, recall, F1-score, and area under the receiver operating characteristic (ROC) curve. These metrics will provide a comprehensive assessment of how well the models predict match outcomes, guiding further refinement and application in the context of professional esports.

References

[1] N. Wang, L. Zhang, F. Li, and X. Liu, "Outcome prediction of dota 2 using machine learning methods," in *Proceedings of the 2018 International Conference on Mathematics and Artificial Intelligence*, (Beijing, China), 2018.

[2] K. Akhmedov and A. H. Phan, "Machine learning models for dota 2 outcomes prediction," *arXiv preprint arXiv:2106.01782*, 2021.

[3] L. Zhang, C. Xu, Y. Gao, Y. Han, X. Du, and Z. Tian, "Improved dota 2 lineup recommendation model based on a bidirectional lstm," *Tsinghua Science and Technology*, vol. 25, no. 6, pp. 712–720, 2020.

[4] K. Song, T. Zhang, and C. Ma, "Predicting the winning side of dota 2," *Journal of Computer Science and Technology*, vol. 34, no. 4, pp. 567–580, 2019.

[5] K. Wang and W. Shang, "Outcome prediction of dota 2 based on naïve bayes classifier," in *2017 IEEE/ACIS 16th International Conference on Computer and Information Science (ICIS)*, (Kobe, Japan), pp. 123–130, IEEE, 2017.

[6] P. Grutzik, J. Higgins, and L. Tran, "Predicting the outcome of professional dota 2 matches using team draft data," in *Proceedings of the 2017 Conference on Computer Vision and Pattern Recognition*, (Honolulu, HI), 2017.

[7] A. Katona, R. Spick, V. J. Hodge, S. Demediuk, F. Block, A. Drachen, and J. A. Walke, "Time to die: Death prediction in dota 2 using deep learning," in *2019 IEEE Conference on Games (CoG)*, (London, UK), pp. 341–348, IEEE, 2019.