A Comprehensive Analysis of Factors Influencing Dota 2 Match Outcomes using Machine Learning Techniques

Qichang Dong

Introduction

Dota 2 is a popular multiplayer online battle arena (MOBA) game, has gained widespread attention in the e-sports community due to its complex gameplay and strategic depth. With an increasing number of players and spectators, understanding the factors that influence match outcomes has become increasingly important for players, coaches, and analysts. This study aims to explore the various factors influencing match outcomes in Dota 2 and develop a predictive model using machine learning techniques to enhance decision-making and performance in the game.

Dota 2 Overview

Dota 2 is a highly strategic, team-oriented game that involves two squads, Radiant and Dire, each with five players, competing to destroy the opposing side's ancient structure. Boasting a diverse roster of over 100 unique heroes and a vast selection of in-game items, Dota 2 offers immense variety for devising innovative strategies and tactics. The game's complexity is further amplified by frequent updates introducing new heroes, balance changes, and features. This intricate and dynamic nature has captivated researchers who explore various aspects of the game, such as team compositions, hero synergies, and counter-picks, to gain a deeper understanding of its mechanics and strategies.

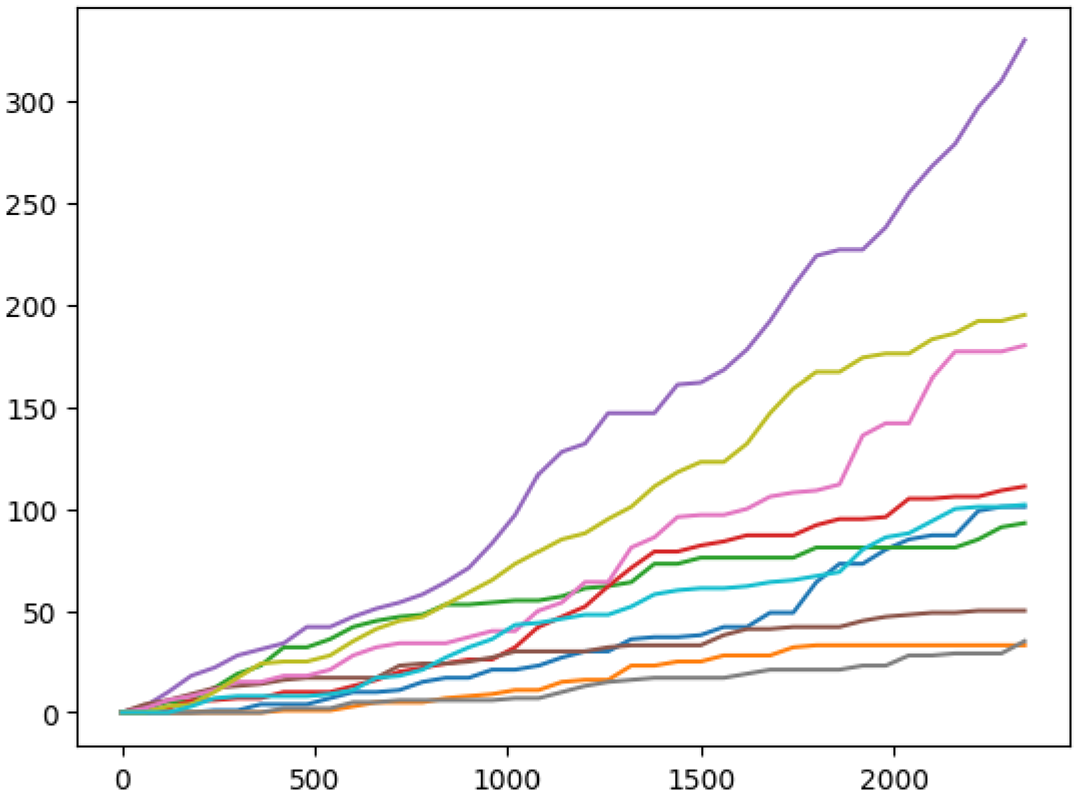
Data Collection

Data for this study will be collected from publicly available Dota 2 match data through Kaggle. mlcourse.ai - Dota 2 - winner prediction Dataset | Kaggle The dataset will include match data from professional, semi-professional, and high-ranked public games, ensuring a diverse range of skill levels and strategies. The dataset will consist of over 100,000 matches to provide a robust sample for the analysis.

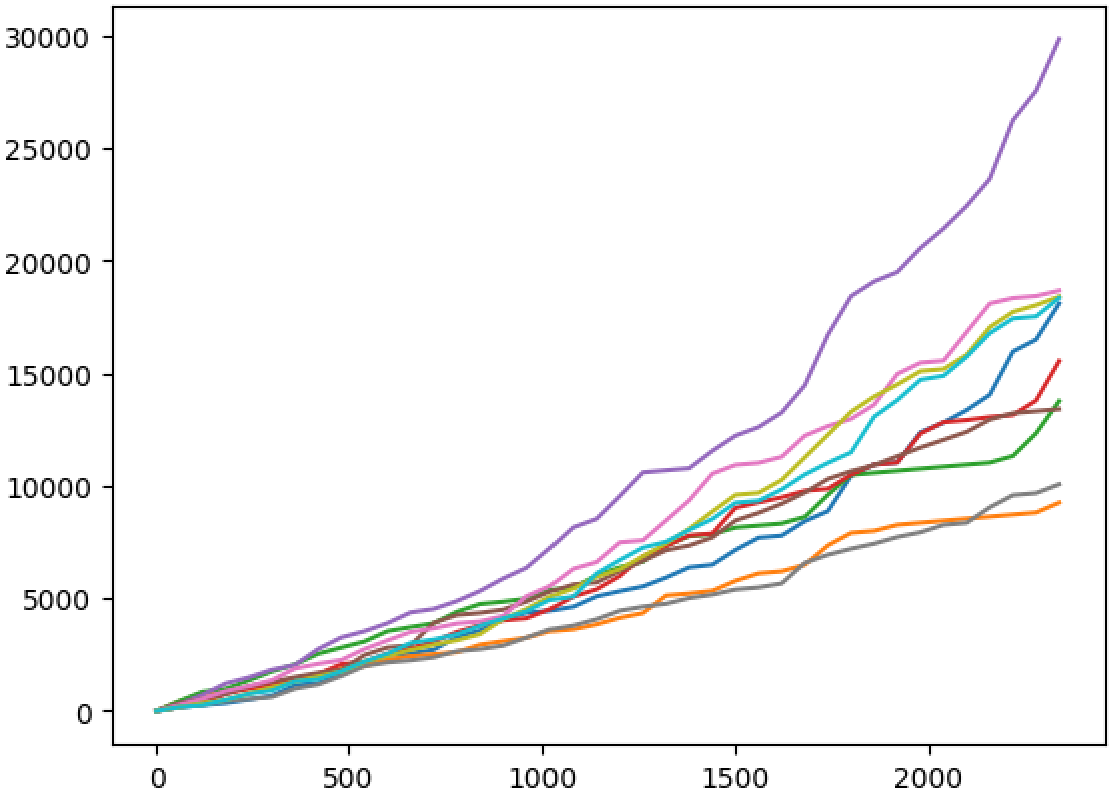
Data Preprocess and Exploring

Initially, I imported all necessary data for the model and examined the contents of the dataset. Given the complex nature of Dota 2, it became evident that all features would significantly impact the model's performance. To prepare the training data, I removed duplicate columns from the train_target.csv file and combined it with the train_features.csv file. I then utilized StandardScaler to scale all data, excluding the radiant win rate, resulting in the final dataset for the training process.

Subsequently, I conducted data exploration to better understand the dataset. Utilizing the JSON file provided with the dataset and Matplotlib, I created a line graph to visualize the relationship between Time and Last Hits in the game. In Dota 2, the last hits are crucial as they translate to gold, which ultimately allows players to acquire better items. This makes the last hits a significant factor in determining a team's victory or defeat. Interestingly, the graph revealed that, contrary to the expectation that competing teams should maintain a similar economy, there was consistently one player who secured substantially more last hits than others throughout the Game.



Last Hit for all players



Gold for all players

Model Selection

In this study, various machine learning algorithms were considered for the prediction task of Dota 2 match outcomes based on the provided features. Logistic Regression was ultimately chosen as the preferred model due to its interpretability and simplicity. As a simple and easily interpretable model, Logistic Regression allows for direct analysis of the relationships between features and the predicted outcome, which is essential for understanding the significance of individual factors in the prediction process and enabling stakeholders to make informed decisions. Additionally, its linear nature works well when the relationship between features and the outcome is approximately linear, making it a suitable choice for our dataset.

The models will be evaluated using metrics such as accuracy, precision, recall, and F1 score. Additionally, the area under the receiver operating characteristic (ROC) curve (AUC-ROC) will be used to assess the model's performance in distinguishing between winning and losing teams. The model with the best performance will be selected as the final predictive model for Dota 2 match outcomes.

Conclusion

The logistic regression model applied to the Dota 2 dataset demonstrated a reasonable ability to predict match outcomes, achieving a 72% accuracy and an ROC AUC score of 0.72. These results indicate that the model can provide insights into factors that contribute to winning or losing a match. However, given the complexity and dynamic nature of Dota 2, further research and refinements are necessary to enhance the model's performance. This could involve exploring additional features, experimenting with alternative machine learning models, or fine-tuning hyperparameters. Ultimately, a more accurate and robust predictive model could provide valuable insights for players, coaches, and the broader Dota 2 community, enhancing strategies, team compositions, and gameplay.

Work Cited

Dota 2, https://www.dota2.com/home.

Biswas, Sushma. “Mlcourse.ai - Dota 2 - Winner Prediction Dataset.” *Kaggle*, 8 Sept. 2019, https://www.kaggle.com/datasets/sushmabiswas/mlcourseai-dota-2-winner-prediction-dataset.