Research Proposal: Predicting eSports Wins Utilizing Real-Time Data

Daniel Craig

CUNY - DATA698

Dr. Arthur O’Connor

2/26/24

# Introduction

ESports is a rapidly growing industry, currently evaluated at $4.3 billion dollars in 2024, with a 7% growth rate over the next 4 years and expected value of $5.8 billion dollars by 2028 [1]. In the realm of competitive gaming, Dota 2 is one of the most lucrative titles a professional eSports player can compete in. Viewership for Dota 2’s global tournament, The International, garnered nearly 1.5 million viewers in 2023 [2]. The International holds 7 of the top 10 highest prize pools in eSports history with the highest record of $40 million dollars and Dota 2 as a game holds 8 of the top 10 spots [3]. The game holds roughly a 200,000 player count at any point in the day with peaks of up to 600,000 [4]. The game is in a 5v5 team format in a top-down perspective, such as real-time strategy games StarCraft or Command & Conquer, with players selecting from a pool of 100+ heroes with unique abilities with the goal of destroying the enemy team’s “Ancient”, taking players 35 to 45 minutes to complete the game. The “Ancient” is the furthest building from the enemy team’s spawn location and serves as an equivalent to a King in chess. The players compete for two primary resources; gold and experience. These resources can be gained by being near or killing enemy players or units. Players are placed on a square map with three distinct areas called “lanes.” Non-controllable units called “creeps” run down the three lanes where they crash against enemy creeps. Players compete for killing more creeps than their enemies to gain gold and experience until they are strong enough to take the enemy’s base. There are several defensive buildings called towers that are placed in each lane that serve as deterrents to keep enemies from destroying a team’s Ancient. Teams contend with each other using the unique abilities of their hero, typically with millions of dollars on the line.

As a heavily viewed eSport, statistics, predictions, and match favorability are presented during live feeds of professional matches with an entire production crew serving live overlays and stat blurbs during gameplay. One feature that was embraced and developed by the game’s developer, Valve, is a live feed of win favorability, in the form of percentages, throughout the game. Yu et al. (2018) [5] found the feature to be 68% accurate at the half-way point in games. This favorability can drastically change throughout the course of the game swinging from 80% favorability to as low as 30% at times. While a formal explanation of Valve’s methodology for predicting a win during a live game has not been made public, there have been several forays into predicting wins.

# Literature Review

Song et al. 2015 [6] predicted wins based on only the heroes selected during the drafting stage of a match before any play across 3000 matches. A logistic regression approach was utilized with features chosen as heroes drafted and hero combinations for a testing accuracy of around 60%. Akhmedov and Phan [7] created a game – state integration app to pull data during a game live for roughly 100 games, performed a sensitivity analysis around the variable *player.gold* to determine highly correlated features to create a new dataset of variables for use in a multi-step forward prediction and fed into Linear Regression, Neural Net, and Long Short-Term Memory (LTSM) models for impressive 82%, 88%, and 93% accuracies, respectively. Yang et al. 2016 [8] combines pre-game features in player ranking and player hero statistics which results in a 70% accuracy in prediction and later combines these features with gold, death, and experience variables gathered from live games for use in an Attribute Sequence Model (ASM). The ASM would then predict the transition probability of the variable *player.gold* to quantify likelihood of gold change. These outputs from the pre-game logistic regression model and the ASM transition probability metric are combined into another logistic regression model for a final prediction. A comparison of prediction accuracies between pre-game, real-time, and combined models are compared over game duration with the combined models sitting between 75-80% within the first 20 minutes of a game and 85-95% after the 25 minute mark. Most live prediction papers focus on treating features in time windows or time series.

As some studies progressed into deep learning utilizing recurrent or convolutional neural networks, others returned to classic machine learning algorithms such as V.J. Hodge et al. (2021) [9]. V.J. Hodge et al. (2021) focused on a Weka Logistic Regression, Weka Random Forest, and Microsoft’s LightGBM algorithms for prediction and parsing features using a replay parser. Specific features were not detailed outside of being labeled as in-game metrics. These models were tested during a live event, ESL One Hamburg 2017, in coordination with the event handler to generate live stats and enhance commentary and analysis by the talent panel. Game stats were retrieved using another Game State Integration app. Accuracy for predictions in professional games were placed at 75%.

# Research Goals and Methods

Proposal for this research would attempt to bridge a specific gap related to feature selection that is not highlighted in aforementioned documents, namely, the focus on in-game gold progression attributed to specific heros and the impact of team fights in determining outcome. While most research focuses on a gold differential between teams to evaluate as a feature, this research will focus the relationship between allocation and growth of gold with interactions in team fights. Ultimately, predicting a win in the game is determining which team is performing better, assuming a well-balanced game. In recent years, Valve’s development team have improved data storage to include metrics at time points. A log of team fights, including which hero killed whom, and individual hero gold progression by the minute have been added. The theory is that the increased detail in the features will contribute to a higher win prediction accuracy by revealing important relationships between gold acquisition and time of game as well as allowing team skill to be expressed by kill interactions during fights, rather than a simplified gold differential. The reasoning behind this expectation is due to an intended balance feature of the game. Heros are designed to be strong at different points in the game, as well as having different strengths such as being proficient at dealing damage to enemies or excelling at dealing tower damage. Heros typically receive a spike in power upon purchasing certain items at specific timings with a follow-up play by the team to capitalize on said timing. Using a simple gold differential between teams does not account for latent implications between gold amounts on specific heros and its implication on the upcoming fight or interactions between heros in fights. A goal of this paper is to evaluate the efficacy of including hero gold acquisition on minute marks and team fight interactions to more accurately judge which team is performing better to hopefully increase accuracy in live game predictions.

Anticipated models to be used for this are a combination of Support Vector Machine, Random Forest, and Recurrent Neural Network models. Challenges in this research will be finding techniques to utilize the anticipated time-series data of gold gain and team fights as features for models to work with, similar to the research performed in [8] and [9]. A main differentiator between this research and others is the lack of a Game State Integration used by [7], [8]. and [9] to collect all data available from a game for training purposes, and instead relying on the recorded minute mark values of variables in the historical database to re-create a “live-game” scenario for the model to train on. Papers focusing on live-game statistics with a type of game state integration tended toward smaller sample sizes due to the work involved with data collection. The advantage of using the recorded minute mark values of variables that this paper suggests is access to hundreds of thousands of samples that can be used as training data. The data will be collected using the OpenDota API that gives access to Valve’s database storing Dota 2 game metrics. The database is updated per game and thousands of games are played every day.

# Citations & References

[1] *DOTA 2 TI viewers 2019 | Statista*. (2019). Statista; Statista. https://www.statista.com/statistics/518201/dota2-championship-viewers/

[2] *DOTA 2 TI viewers 2019 | Statista*. (2019). Statista; Statista. https://www.statista.com/statistics/518201/dota2-championship-viewers/

[3] *Top eSports tournaments by prize pool 2022* | *Statista.* (2022). https://www.statista.com/statistics/517940/leading-esports-tournamets-worldwide-by-prize-pool/#:~:text=The%20eSport%20tournament%20that%20has

[4]Djundik, P. (2013, July 9). *Dota 2 Steam Charts* [Review of *Dota 2 Steam Charts*]. Steam Charts - SteamDB. https://steamdb.info/app/570/charts/#1y

[5]Yu, L., Zhang, D., Chen, X., and Xie, X., “*MOBA-Slice: A Time Slice Based Evaluation Framework of Relative Advantage between Teams in MOBA Games*”, arXiv e-prints, 2018. doi:10.48550/arXiv.1807.08360.

[6] [Song, K., Zhang, T., & Ma, C. (2015). *Predicting the winning side of DotA2.* https://cs229.stanford.edu/proj2015/249\_report.pdf](https://cs229.stanford.edu/proj2015/249_report.pdf)

[7] Akhmedov, K. and Phan, A. H., “*Machine learning models for DOTA 2 outcomes prediction*”, <i>arXiv e-prints</i>, 2021. doi:10.48550/arXiv.2106.01782.

[8] Yang, Y., Qin, T., and Lei, Y.-H., “*Real-time eSports Match Result Prediction*”, <i>arXiv e-prints</i>, 2016. doi:10.48550/arXiv.1701.03162.

[9] V. J. Hodge, S. Devlin, N. Sephton, F. Block, P. I. Cowling and A. Drachen, "*Win Prediction in Multiplayer Esports: Live Professional Match Prediction*," in IEEE Transactions on Games, vol. 13, no. 4, pp. 368-379, Dec. 2021, doi: 10.1109/TG.2019.2948469.