Machine_Learning_Project_Proposal(2)(2)

November 14, 2025

1 Project Proposal Title:

Predicting Match Outcomes in the Game DoTA

1.1 Teammates

- Wyatt Churchman (jdr357)
- Segundo Sanchez (sas458)
- Ryan Kerlick (rak88)

1.2 Project Abstract

In this project, we want to predict the outcome of DoTA 2 matches using real game data from the OpenDota JSON Data Dump. Each match contains a huge amount of numeric information, including player statistics, gold and XP graphs, combat logs, item usage, time-series data, and other gameplay metrics. Our plan is to convert these JSON objects into a large, high-dimensional numeric dataset (well over 10 million floats) and then build machine learning models on top of it.

We will start with simple baseline models to see how well raw features predict whether the Radiant team wins. After that, we'll use dimensionality-reduction and HD-curse mitigation techniques—like PCA, Random Projection, and APP—to see if they improve performance or reveal hidden structure in the matches. The goal is not just to predict the winner, but also to learn something interesting about playstyles, match patterns, and how high-dimensional features interact.

1.3 Problem Statement

- *State the problem you propose and why it's important.*

In competitive online games like DoTA, it's extremely important to match players of similar skill levels. When a low-skill player gets paired against someone who is much more experienced, the outcome is almost always one-sided, which leads to frustration and a bad gameplay experience. At the same time, high-skill players don't get much enjoyment from a match that offers no challenge. Over time, poor matchmaking can reduce player engagement and hurt the game's long-term health and revenue.

Being able to predict match outcomes based on pre-match and early-match features could help improve matchmaking systems by identifying patterns that separate balanced matches from mismatched ones. Understanding these patterns also helps explain what actually influences a fair, competitive game.

- *Give a clear and complete statement of the problem.*

Can we use DoTA match data alongside ML algorithms to predict match outcomes?

- *What benchmark will you use, and why?*

We will use accuracy, F1 score, RMSE for win probability. These benchmarks are directly relevant to our data, and will show us how close our predictions are to the true outcome.

- *Where does the data come from, and what are its characteristics?*

Our dataset comes from the OpenDota JSON Data Dump, which contains over one billion recorded matches of DoTA. Each match includes information about the match itself (duration, game mode, region), the ten participating players (kills, deaths, assists, gold/min, XP/min, damage, item logs, and more), and the skill bracket assigned by Valve. This dataset is extremely large, high-dimensional, noisy, and multi-source. After flattening the JSON, each match produces hundreds of numeric features, and using only a small subset of the full dataset gives us well over the required 10 million floats.

- *What informal success measures do you plan to use?*

Beyond the formal benchmarks, we will also look for signs that the model is capturing known gameplay patterns, such as:

- Higher-skill players tending to have stronger early-game stats
- Shorter matches at higher skill brackets
- Early gold/xp advantages being strong predictors of winning

If the model reflects patterns that experienced players already recognize, that's an informal confirmation that our model is learning meaningful structure.

- *What do you hope to achieve?*

We hope to produce a reliable model that delivers predictions with high accuracy, and that picks out the most important predictors of match results. We hope to use this model to improve matchmaking to grow the player base.

1.4 Dataset

https://blog.opendota.com/2017/03/24/datadump2/

This dataset contains data from the video game DoTA. It houses information on over a billion matches inside the game, divided into three torrents, each housing different information about the matches.

matches: Match level data (start time, cluster, game mode)

player_matches: Data for each player (kills, deaths, gold per minute)

match_skill: Skill bracket assigned by Valve, the developer of the game (Normal, high, very high)

1.5 Methodology

- *What AI/ML tools will be used as a baseline and what as an improvement?*

Our approach starts by flattening the raw OpenDota JSON match data into a large numeric dataset by extracting player stats, match stats, time-series summaries, and aggregated logs. Once the data is cleaned and standardized, we build baseline models such as Logistic Regression and Random Forest to predict match outcomes. After establishing a baseline, we apply high-dimensionality reduction techniques—including PCA, Random Projection, APP, and feature selection—to improve model stability and reveal hidden structure in the data. We then retrain our models on the reduced feature sets and also explore clustering (like K-Means) and anomaly detection to better understand different match patterns. Throughout the process, we compare model performance before and after dimensionality reduction and visualize our findings to uncover the most important gameplay features influencing match results.

1.6 Teaming Strategy

Who does what? When and how often do you meet?

We plan to check in with each other every day we have class to talk about progress, divide up new tasks, and make sure everyone knows what they're working on. Outside of class, we'll stay in contact on Discord and update each other on anything that needs attention.

We'll also work together on the shared parts of the project like the presentation slides, the written report, and connecting our individual pieces into one full pipeline.

To make collaboration smoother, we'll use:

- Google Colab for shared notebooks
- GitHub for version control and code sharing
- Discord for communication and quick problem-solving
- Google Docs/Slides for the write-ups and final presentation

1.7 Role Assignments and Commitment Matrix

- *Project assignments and completion plan.* Role Assignments

Ryan : Data Pipeline & Feature Engineering Lead - JSON parsing, flattening, numeric feature extraction, constructing the >10M float dataset

Wyatt : Modeling Lead - Build baseline models, dimensionality reduction (PCA, RP, APP), clustering and prediction

Segundo : Visualization & Reporting Lead - Produce plots, PCA projections, cluster visualizations, write results & final report/presentation - *Collaboration tools and how you will ensure success.*

We will be using multiple softwares and tools to collaorate. One software will be Google Colab, to create notebooks and write reports together. Github and Discord are other software we will use to communicate and shar files.

1.8 Mitigation Plan

- *What will be the data/task/method alternative?*

If the full JSON dataset ends up being too big or slows down our progress, we'll switch to a more compact version by focusing only on the player-level stats or by using the OpenDota player_matches table, which is already flatter and still gives us well over 10 million floats.

If predicting match outcomes turns out to be too noisy, we'll switch to an unsupervised task like clustering different types of matches or detecting outlier matches (for example, games with extreme gold swings or unusual player behavior).

- *What if a teammate is MIA?*

If someone becomes inactive, we'll redistribute their tasks evenly among whoever is available, and we'll notify the instructor like the rubric requires. We may also scale down the project slightly (for example, fewer dimensionality-reduction methods or fewer match records) so we can still finish the work on time.

- *What if your baseline is GIGO?*

If our baseline model turns out useless because the raw data is too messy or inconsistent, we'll clean up the feature set, try dimensionality-reduction techniques (PCA, Random Projection, APP), and engineer more stable aggregate features. These steps should improve model quality even if the raw metrics are noisy.