Predicting winners in esports tournaments (Super Smash Bros Melee)

Executive Summary

Team: Dan Ursu, Jaspar Wiart
Github: https://github.com/jjaw89/fall-2024-smash-melee-top-8-prediction

Overview: Super Smash Bros Melee is a fighting game released for the Nintendo GameCube in 2001. While several newer installments in the franchise have been released since then, Melee continues to regularly be played in tournaments offering tens of thousands of dollars in prize money, and pull in hundreds of thousands of viewers online.

Objective: Our aim was simple: *predict the winner*. Specifically, we first wanted a model to predict the winner of individual matches, and a second model that would look at the top eight finalists in a tournament, and predict the overall winner. In both instances, we already had a fairly sophisticated baseline model to try to beat, described below.

Baseline: In zero-sum games, the Elo rating system or some variant is used to compute numeric values meant to indicate the relative skill levels of players (with the Glicko-2 rating system being the most popular modern version). Because of all of the work that has already gone into creating these widely-used rating systems, the baseline model of simply predicting whoever has the highest rating as the winner is already fairly sophisticated, and any small improvement upon this model would be considered a success.

Key Performance Indicators (KPIs): Accuracy score was our primary performance metric.

Methodology: Data on previous tournaments was easily obtained from this github repo, provided by smashdata.gg and containing information on players, tournaments, matches, and occasionally individual game data (such as characters played, etc…). The dataset was fairly complete, and minimal cleanup was required. From here, we computed weekly Glicko-2 ratings of each player, and with this were able to establish our baseline models.

Super Smash Bros Melee is a fighting game where players can choose individual *characters* to fight with, and it is very plausible that different players will play better or worse purely depending on their opponent's character. Our first goal was therefore engineering features that took this into account, and in the end, we settled on three different modified versions of the standard Glicko-2 rating. This is where a substantial portion of our effort went.

Additional features, mostly miscellaneous information related to the ratings (both default and engineered), character usage, and previous matches between players, were also added.

For our single-match predictor, we trained XGBoost on all of our engineered features and used Optuna to perform hyperparameter tuning. Unfortunately, we had several ideas for a model to predict the winner out of the top eight finalists, but it never outperformed a slight modification to the baseline.

Results: We saw a genuine improvement in our model for predicting individual matches, when compared to the baseline of "whoever has the higher rating". A second baseline of training XGBoost on ratings only (to account for strange behaviour near default/extreme values) is also given. Confidence intervals of 95% are also given next to the accuracy scores, and are quite small due to the large dataset.

Model (individual matches)	**Accuracy** (on all matches)	**Accuracy** (on top 8 matches)
Whoever has the higher rating	77.56 ± 0.16	73.89 ± 0.36
XGBoost trained on rating only	79.05 ± 0.16	74.04 ± 0.36
XGBoost trained on all engineered features	79.89 ± 0.16	75.03 ± 0.35

In particular, testing on all matches, and restricting to top 8 matches, we clearly see a statistically significant improvement in accuracy of approximately 1%. Outputting the feature importance of our final model also shows that almost all of our engineered features, especially character-specific ratings, were used substantially in the model.

No top 8 predictor that we created outperformed the baseline of simply choosing the player with the highest rating out of the four finalists from the "winners' side" brackets. This baseline had an accuracy of approximately 70.1 ± 1.3 (as opposed to 67.6 ± 1.3 for the highest rated player in the entire top 8), and our more sophisticated models all performed similarly.

Conclusion: The main achievement in our project is showcasing how the standard Glicko-2 rating system can be used in novel ways to generate not only the default skill ratings for players, but character-specific skill ratings as well (among other features), which altogether increase the predictive accuracy of our models above a baseline of just choosing the player with the highest default rating. These ideas are not exclusive to Super Smash Bros Melee, and are likely applicable to other fighting games as well.

Future work: XGBoost was chosen among a few other more basic models as seemingly giving the highest accuracy, but it is quite possible that applying and properly tuning more sophisticated models, such as neural networks, might increase performance. It is also quite possible that there are many other useful features that could be engineered, both for Melee and other esports. Finally, despite the fact that no top 8 predictor that we designed ultimately outperformed the modified baseline accuracy, more analysis could be done to see if one of them is still useful in other aspects, such as predicting upsets.