Combat Conundrum: Enhancing Dungeons & Dragons Encounters

Executive Summary

Ingrida Semenec, Noah Waller, Alxandr Kane York, Saswat Mishra, Doug Stauffer

1 Overview

Dungeons and Dragons (D&D) is a renowned table-top role-playing game noted for its complex combat encounters. These encounters determine the adventure's trajectory and vary in difficulty. However, the existing system to gauge encounter difficulty, known as challenge-rating, has significant limitations and fails to address the nuances of combat dynamics. To remedy this, we embarked on a data-driven exploration using FIREBALL: a dataset comprising 25,000 real gameplay combat encounters collected via the Avrae Discord bot. We aimed to develop a more sophisticated tool to assist Dungeon Masters (DMs) in crafting balanced and engaging encounters.

2 Data Cleaning

The FIREBALL dataset was intricate and broadly unstructured. We encountered many challenges in data cleaning, which consumed a significant portion of our effort. We spent about 85% of our time organizing the raw data and implementing necessary filters. We ultimately curated a set of 10,000 combats. The process involved removing irrelevant data, such as non-combat sessions, instances that suggested "homebrew" (custom, non-standard) rules, and excluding anomalous features.

3 Approach

Our initial exploratory data analysis (EDA) showed no clear linear relationships between individual features and our designated measure for combat difficulty: total post-combat party health-point (HP) ratio. Our approach explored a select set of features including weighted total monster level, total player level, number of monsters, and number of players. Classical models, such as Linear Regression and Decision Trees, were employed but proved unsuccessful. We expanded our approach to include a broader range of features and leveraged a Neural Network approach, which aimed to better capture the nuances of combat encounters and enhance our predictive capabilities.

4 Results & Strategies

Classical models, such as Linear Regression, Decision Tree Regression, Random Forest Regression, Gradient Boosting Regression, and Poisson Regression, revealed high bias and poor predictive performance. Although Random Forest Regression and Gradient Boosting Regression showed slightly better performance due to their ability to capture non-linear correlations, all classical models ultimately failed to provide satisfactory predictions. These findings confirmed that neither basic challenge ratings nor weighted challenge ratings could accurately predict combat difficulty, consistent with our project premise.

The Neural Network approach yielded mixed results. Despite incorporating dropout and early stopping techniques to mitigate over-fitting, our model's predictions remained sub-optimal. While the model architecture was modified to distinguish between monster and player features to capture latent relationships, and SHAP analysis revealed shifts in the feature importance order suggesting that the chosen features are insufficient for accurate predictions.

Additionally, we explored predicting the probability of Total Party Kills (TPKs; an indicator of extremely difficult combat encounters) as a crucial aspect of combat dynamics. After reintroducing TPK instances and addressing data imbalances through random under-sampling, we trained Logistic Regression and Decision Tree models for binary classification. Both models exhibited predictive power, with the Logistic Regression model outperforming the Decision Tree.

5 Future Improvements

We did not find a clear predictor of combat success. Future directions include refining the TPK predictor classifier by gathering more TPK-related data. Combat difficulty predictors could benefit from datasets with stricter parameters (perhaps mediated by D&D's parent company, Wizards of the Coast) and the inclusion of diverse metrics like monster statistics. Exploring item analysis, incorporating a larger sample of combat encounters, and utilizing hierarchical models could provide deeper insights into D&D combat dynamics. Additionally, our development of a web app prototype could provide practical implementation of this combat prediction tool, pending improved prediction capabilities.

In conclusion, our work has highlighted the complexities and challenges involved in predicting combat outcomes and estimating TPK probabilities in D&D. We found that challenge ratings do not reliably predict combat difficulty but struggled to identify improved predictive features. Given the limitations of the current dataset, there are promising opportunities for further research and innovation in this area.