Executive Summary

Combat Conundrum: Enhancing Dungeons & Dragons Encounters

1 Overview

Dungeons and Dragons (D&D) is a renowned table-top role-playing game that involves multifaceted 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, proves limited 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. Approximately 85% of our time was dedicated to collating the raw data and applying relevant filtration. Ultimately, we curated a set of 10,000 combats. The process involved eliminating noise, such as non-combat sessions, instances that suggested "homebrew" (custom, non-standard) rules, and filtering out anomalous features.

3 Approach

Our initial exploratory data analysis (EDA) did not reveal clear linear relationships between individual features and our outcome 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 improve our predictive capabilities.

4 Results & Strategies

Classical models including Linear Regression, Decision Tree Regression, RandomForest Regression, Gradient Boosting Regression, and Poisson Regression revealed high bias and poor predictive performance. RandomForest Regression and Gradient Boosting Regression showed slightly better performance due to their ability to capture non-linear correlations, but overall, all classical models failed to provide satisfactory predictions. However, these models showed that basic challenge ratings and weighted challenge ratings could not accurately predict combat difficulty, which aligned with our project background.

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 feature importance orders, the overall predictive performance remained unsatisfactory.

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

In sum, a clear predictor of combat success was not found. 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.