Modeling Proposal

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

April 12, 2024

1 EDA and Classical Methods

Initial exploratory-data analysis (EDA) did not reveal any clear linear relationships between potential individual features and our outcome measure of interest (total party HP ratio; calculated via total party post-combat HP divided by total party pre-combat HP). Given the possibility of a combination of these features to predict the outcome measure, our first modeling approach leveraged features we hypothesized to be most impactful: weighted total monster level (weighted according to the D&D 5th Edition Handbook), total player level, number of monsters, and number of players. We used these features to train LinearRegression, DecissionTreeRegressor, RandomForestRegressor, GradientBoostingRegressor, PoissonRegressor, and MLPRegressor. All models performed similarly (as scored by Mean Squard Error [MSE]), but the GradientBoostingRegressor and RandomForestRegressor performed slightly better due to the non-linear nature of the input correlations. Model performance was visually inspected via plots of predicted values versus actual values. All models showed high bias and a poor ability to model the data, despite low MSEs. A RandomForestClassifier model was used to classify segmented ranges of total party HP ratio and showed similarly poor performance (49 - 8 percent accuracy depending on predicted total party HP ratio range). Ultimately, these classical models informed us that more EDA was needed to identify outliers, segment distinct data trends, and incorporate other potential features in future models. The poor performance of these models highlights the relative inability of the current D%D Challenge Rating system to predict combat success in this actual combat-state dataset.

2 Dimensional Reduction

Given the challenges encountered with classical models, our next approach involves the use of Partial Least Squares (PLS) and Principal Component Analysis (PCA) to enhance our predictive capabilities. The application of these techniques is motivated by several factors inherent to the data and the nature of our prediction problem.

Partial Least Squares (PLS)

PLS is particularly suitable for our context for several reasons: Handling Multicollinearity: Our dataset suffers from multicollinearity, given the interdependence among features like the weighted total monster level, total player level, number of monsters, and number of players. PLS handles multicollinearity efficiently by projecting the original features into a new space of latent variables that summarize their predictive information.

Dimensionality Reduction with a Focus on the Response Variable: Unlike PCA, which only considers the explanatory variables, PLS also takes into account the relationship between predictors and the outcome variable (total party HP ratio). This makes PLS especially useful for our dataset, where the goal is to improve prediction accuracy by capturing the most relevant variance.

Improving Model Interpretability: By reducing the feature space to a smaller set of latent variables that are most relevant to our response variable, PLS simplifies the complexity of our models. This aids in better understanding the underlying patterns and relationships, contributing to a more interpretative modeling process.

Principal Component Analysis (PCA)

Noise Reduction: Our initial EDA suggests the presence of considerable noise within the dataset. PCA helps in reducing this noise by focusing on the principal components that capture the most variance in the data, thereby improving the signal-to-noise ratio.

EDA Enhancement: PCA will serve as an advanced EDA tool. By visualizing the principal components, we can uncover hidden patterns, trends, and clusters that were not apparent before. This can inform the selection of features and the direction of more targeted modeling efforts.

Baseline for Comparison: PCA will also serve as a baseline to evaluate the effectiveness of PLS. By comparing models built on principal components with those built on PLS components, we can better understand the benefits of incorporating the response variable into our dimensionality reduction process.

In summary, the incorporation of PLS and PCA into our approach aims to address the limitations observed with classical models by efficiently handling multicollinearity, reducing dimensionality while retaining relevant variance, and enhancing model interpretability and prediction accuracy.

3 Deep Learning Method

To better incorporate our features, efficiently process our unstructured dataset, and attempt to model hidden relationships/patterns among our features and outcome measure, we will use a Neural Network modeling approach. This model is designed to handle two different types of input data through separate pathways: one for monster-related features and the other for player-related features. Each pathway will be composed of several linear layers with ReLU activation functions and the final layers will be designed to produce outputs of specified dimensions. These final layers will then be combined and further processed through additional ReLu-activated layers. Finally, we will use a sigmoid function to predict the HP ratio as a proxy for the result of the combat. The training loop includes gradient zeroing, forward pass, loss computation, backward pass, and parameter updates. It also includes early stopping based on validation loss to prevent over-fitting and a plotting function for visualizing the training loss over epochs. Finally, we will evaluate the model and quantify its performance using the Mean Squared Error (MSE) between the predicted and actual outcomes, followed by a visual comparison.