Amani Akkoub (ctf3un), Charli Ashby (arv2vp), Jonah Cicatko (zvh3zz), Emily Hunter (brw6pe), Anne Kumashiro (yxt7ue), Jack Nickerson (rze7ud)

DS 3001 - Foundations of Machine Learning

11/03/2024

*Pre-Analysis Plan*

The goal of this project is to predict the results of the college football 2024-25 season, focusing on the ACC conference. Specifically, we aim to identify common characteristics of successful teams from past seasons to predict outcomes in the current season, culminating in predictions for the ACC Championship game. The central research question is: Who is most likely to win the ACC Championship game?

An observation in this study is a single football game that has already occurred, with various features such as team statistics that act as our independent variables that could influence the outcome of a game (ex. offensive success rate, game location, weather conditions, etc). The target variable will be the outcome of the game, which can be represented as a binary variable, win or loss. This study will use supervised learning with a focus on regression models. The goal is to predict the binary outcome related to the championship game, win or loss.

The models and algorithms to be used include linear regression and decision trees. Linear regression will be used to capture the relationships between continuous game features and outcomes, providing insight into the linear impact of each feature on the outcome. Decision trees will be applied to capture potential nonlinear interactions among variables, such as interactions between location, weather, or team and player statistics. Regularization techniques, such as Lasso and Ridge, help manage multicollinearity and prevent overfitting in statistical models. Multicollinearity arises when predictor variables are highly correlated, making it challenging for the model to distinguish the individual impact of each predictor. This can lead to unstable estimates and a model that doesn't generalize well to new data. Regularization works by adding a

penalty term to the model's objective function, which discourages large coefficients. Lasso (L1 regularization) can reduce some coefficients to zero, effectively selecting only the most important predictors, while Ridge (L2 regularization) shrinks coefficients toward zero without eliminating them entirely. Both techniques reduce overfitting by simplifying the model, making it more robust to new data and less reliant on noise or irrelevant patterns in the training data.

We will measure the success of the model by how accurately it predicts the actual victor of a football game. We can additionally use metrics that quantify binary classification model performance such as the $F_\beta$ score in our assessment of the success of our model.

The major anticipated weakness of this model is that teams change every season due to factors like graduations, drafts, and transfers. These changes can cause variability that may reduce the power of the model because of non-representative historical data. We can potentially include features that account for team changes, such as a new player acquisition or team experience levels, and we will regularly update the dataset with the mode updated data to ensure that the model adapts to the changes made mid-season. If this approach fails, we know that historical data alone is not enough for an accurate prediction because of the dynamic compositions, indicating a more complex model is necessary for an accurate prediction.

To prepare the data for the analysis, we will use one-hot encoding to convert our categorical data into a numerical format that algorithms can process, for example, variables like team names, game locations, and weather conditions will need to be converted. We can use principal component analysis (PCA) if we find high correlation between our numerical variables in order to avoid multicollinearity, which is probable to arise from different player statistics/metrics.

The results from this analysis can be shown with a table of regression coefficients for linear models to interpret each independent variable, as well as the visualizations of decision trees to explain decision rules. Comparisons between different models (e.g., linear regression vs decision trees) will also be presented using these metrics to determine which approach provides better predictive accuracy. For example, decision trees may be better suited to represent the non-linear factors like weather or game location, while linear regression will work better to measure variables, such as offensive success rate, that have a more linear relationship. By periodically updating the dataset and adjusting features mid-season to reflect team changes, we aim to increase the model's adaptability to current team dynamics, thus improving its predictive power.