Capstone Two Proposal -
AUDL Ultimate Frisbee Winning Model(s)

Background, Context

The AUDL (2012-current) is a men's professional Ultimate Frisbee league in North America, currently consisting of 24 teams across four regional divisions. By now, there are 10 years of team, game, and player statistics available on their website and via their API. It is likely some teams are utilizing their data to develop and optimize winning strategies in-house, but to my knowledge there are not publicly available, widely-adopted winning or points-added models.

For this project, I will imagine the stakeholders to be strategic decision makers of a hypothetical team. I will look for historical trends in the data that I can use to guide my client's team-building and game-strategy.

Problem Statement

Can basic game summary statistics (throws, completions, blocks, turnovers) be used to predict a game's outcome? If so, how do a team's stats contribute to their likelihood of winning?

Data Collection

As mentioned previously, I will use the AUDL API for initial data retrieval and store it locally. There are endpoints available for a number of statistics:

1. **Game scheduling and scores**
2. **Game summary statistics for each team**
3. Game events (play-by-play)
4. Player history (teams, years active)

I will use the endpoint (1) to get a list of all gameIDs, and then use those to collect each game's summary statistics using the endpoint (2). In my initial work, I have had to work around a few issues. I will detail these in the notebook submissions.

I will also start the process of parsing the play-by-play data into a much larger data set through which one could follow the flow of a game. An example can be seen in the **Field Map** visualizations the AUDL provides in their "Advanced Stats" page for a game. This data and its analysis will be outside the scope of the Capstone Two model.

Constraints

Not all statistics are recorded for every game. Therefore, I will have to decide between eliminating a feature or a decent amount of entries. For example, “Red Zone” and “Huck” statistics are not available before 2019. Will it be better to drop that many games, the associated features, or find a way to impute missing data?

I will discuss these decisions in the submission notebooks. Imputation choices (whether or not, how) may be influenced by background knowledge. I have the current feeling that I will be dropping more entries than filling in values.

Initial Scope

I will engineer and use the features available for each game summary to develop two models. One will be a regression model to predict the margin of victory for a certain team, given their and their opponent’s statistics. Another will be to simply classify a winner for a game, based on the two teams’ numbers, or to determine the probability of a certain team winning. I imagine there may be some changes along the way, too. Ideally, the models will not be too complex, and “real-world” strategies may be elucidated from their parameters. A variety of linear/logistic, bagging, boosting, and nearest-neighbors algorithms will be used for prediction. Models will be evaluated based on their goodness of fit and with a variety of aggregate error metrics. After evaluation, a single model will be provided for each prediction.

Deliverables

After completion of work, results will be summarized into slides and a report to be delivered to the client(s), as well as the required cleaned data and top-performing model parameters. Future work building on the delivered models and data procedures will be discussed.

Future Ideas *(not part of Capstone Two deliverables)*

- **Data collection**
  - Current statistics collected are only for game summaries, collect granular data for each game
  - Include roster information for games and points
  - Formalize pipeline for API data retrieval, cleaning+conversion, and persistence in a dedicated repository
- **Data Analysis**
  - Analyze new datasets
  - Include player-based analysis, current scope is limited to teams
- **Models**
  - Expand score margin/win probability models with more data
  - Create model to predict inputs for score margin/win probability models using team's recent games and opponent, then use to predict a game's score
  - Unsupervised clustering of player data/individual player contribution to success
  - Event-based, "points added" metric