Micah Lee, Oliver Yang, Kimberly Low, Akash Bhatnagar

BA305 Project Proposal

Initial Ideas

The team discussed a number of different categories as we have different interests, including:

- Entertainment: Movies & Music
- Sports: Football, Soccer, Baseball
- Oil
- Real estate
- Stock market
- Hotels

Potential Datasets

We decided to focus on the area of sports. We used the database resources- and other external sources- to search for relevant datasets. We also attempted to come up with questions we could ask for each.

- MLB: Predict the winning teams in MLB and determine how many times each team will win and lose in a given season.
- NFL (Github)
- NFL (Kaggle): Predict the type of offensive play to be used in a given situation; will the team run the ball, or throw a pass?
- FIFA: Predict the number of goals a team will score and/or predicting the winner in 2026.

Preliminary Questions

Out of the datasets above, we found that the NFL data on Github is the most suitable for this project. Our dataset of interest is named *player_stats_kicking_2021.csv*. It includes kickers' data from each team for each game of the 2021 season. We are thinking of creating a predictor model using the NFL kickers stats dataset in order to help predict relative kicking success against specific teams.

Commented [1]: fantasy football point system- aggregation of the variables in here (will need to merge two data sets)

each player against each team

put together all player data: score = regression between dummies for each player and opposing team

try to use 2-3 models from class, skim through syllabus

Preliminary Findings

Looking at the data, we discovered that:

- There are over 30 variables in the dataset but not all of them are useful. We will delete certain variables which we don't think are necessary to this prediction.
- Although most of the useful variables are metrics, some variables such as *fg_made_0_19* to *fg_made_60_* can be transformed into categorical variables using dummies to help run certain numerical prediction models. We could also convert numerical variables to categorical using discretisation if necessary.
- *Fg_made / fg_missed / fg_pct* and *pat_made / pat_missed / pat_pct* are most likely highly correlated. We will also execute the dimension reduction by running the PCA to avoid high correlation between variables.
- There appear to be a couple of nulls, which will have to be taken care of.

Micah Lee, Oliver Yang, Kimberly Low, Akash Bhatnagar

Our New Question

- Predicting player position based a series of variables

https://nycdatascience.com/blog/student-works/data-study-to-predict-nba-player-positions/

Sports to look at

- Soccer
- Hockey

Possible Datasets to use

https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics?select=transfermarkt_fbref_201920.csv

https://www.kaggle.com/https://www.kaggle.com/datasets/kriegsmaschine/soccer-players-values-and-their-statistics?select=transfermarkt_fbref_201920.csvdatasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv

https://www.kaggle.com/datasets/batuhandemirci/fifa-2021-team-and-player-dataset?select=players.csv