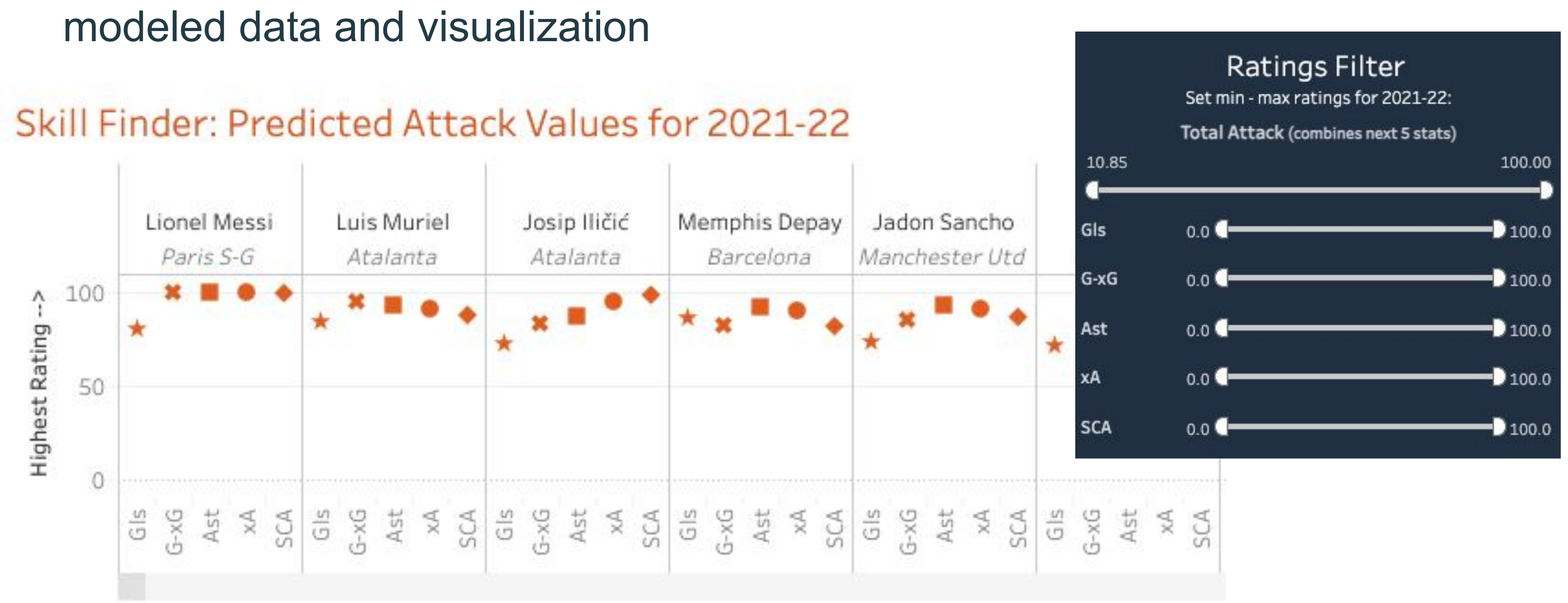
AFC Richmond Analytics

Adam Gushansky, Dan Hislop, Ethan Agranoff,
Jeremy Piech, Sri Asuri, Terry Ballou-Crawford

Problem Definition

- Team AFC Richmond's project focuses on soccer players in Europe's "Big 5" Leagues, which concentrates 48% of soccer's worldwide wealth.
- Different leagues and teams have unique play styles which affect player performance.
- Despite the criticality of these factors, public models rarely account for them when predicting player performance.
- **Goal: Our group aims to integrate team play style and league, along with historical player performance, to predict player performance more accurately.**

- **Audience: Why does it matter?**
- Our player predictions will directly impact pro clubs selecting players to transfer by providing more accurate predictions rooted in specific team and league play styles.
- Soccer fans will use our visualizations to research for betting, fantasy sports, and fun.
- Both teams and fans should be able to narrow their search for players using our modeled data and visualization



Data Characteristics

- Data source: FBref.com, which provides stats for all of the "Big 5" soccer leagues.
- FBref aggregates and cleans data from from several data-collection providers.
- We collect data from 4 seasons: 2017-2018 to 2020-2021.
- Our raw dataset spans roughly 10,000 rows and 175 columns, and when combined with our output (including interim) data, uses about 40 MB of storage.
- Much of the raw data is scaled to a per-90 minutes (a proxy for scaling to per-game) basis.
- In our visualizations, we also scaled the data from 0-100 to make comparisons easier.

Data Prep

- Preprocess the data: numeric features are power-transformed and have interaction terms added, while categorical variables are one-hot-encoded.
- Narrow data set for accuracy, e.g. when insufficient time played over a season, or playing outside of the "Big 5" leagues for the previous season.
- Test model by predicting the 2020-2021 season to test our model via mean squared error (MSE).
- **Select 10 player metrics on which to make predictions for the 2021-2022 season.**

Modeling

Our novel approach builds upon existing public work by adding two categorical variables-- team play style and league-- in order to add predictive power to our models and better understand how these features contribute to overall player performance. To do this we use two machine-learning algorithms:

- K-means Clustering: accounts for team play style by grouping like-teams together based on past performance.
- Group Lasso Regression:
  - Assigns weights to each group of features indicating their respective importance as predictors, while unimportant features are regularized to zero.
  - The grouping element allows combining of certain correlated features.
- After classifying teams and leagues, a transfer matrix was created to factor in relative weights for the move, e.g. "teamA-to-teamB" or "leagueA-to-leagueB" values
- Predictions were then made for players upcoming season based on historical stats AND team play style and league effect.

Evaluation & Results

Evaluation: We evaluated our approach using mean-squared error (MSE) for each of ten player-level metrics to determine accuracy. The MSEs had to be evaluated individually, because each metric has a unique underlying data distribution.

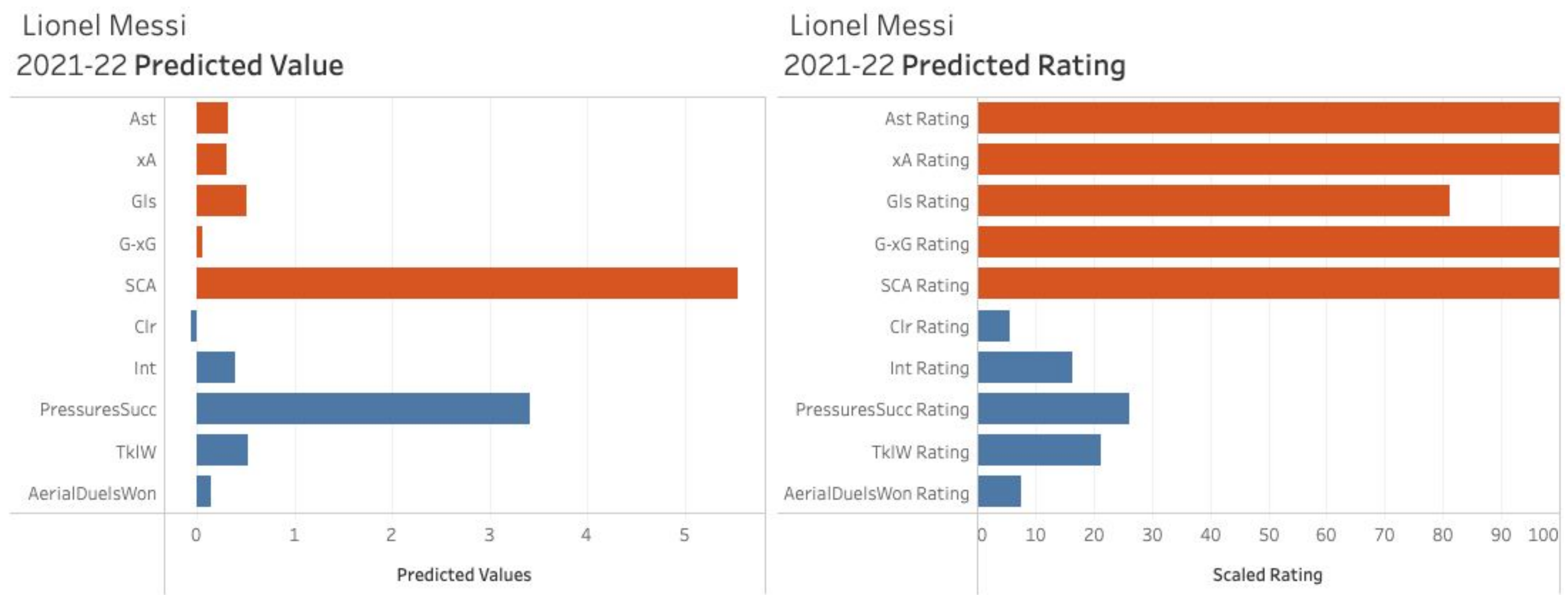
Results:

- **Compared to models without team play style and league effects, the test MSE improved in seven out of 10 models** that included these effects and was equal in the remaining three models.
- The results complement conventional descriptions of play styles and leagues (e.g., the Premier League, known for being a physical league, is a factor in the aerial duels model).

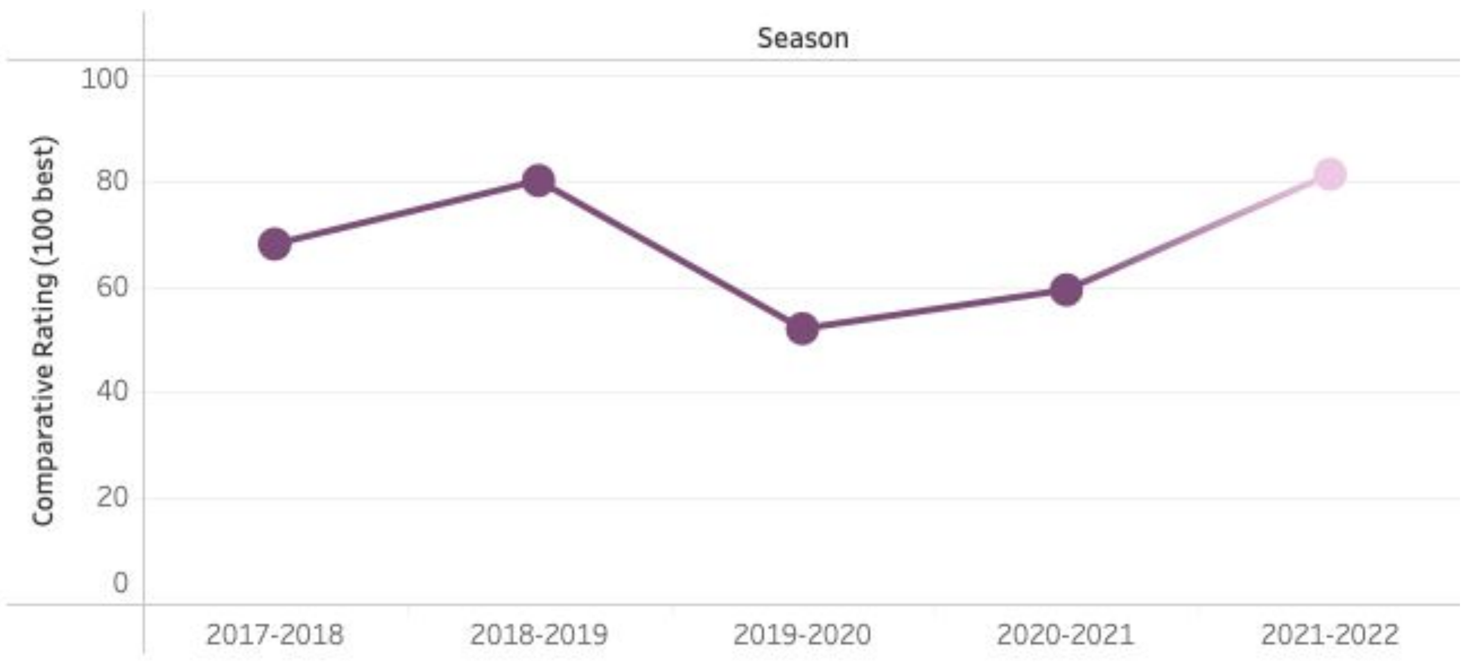
Example

Predictions: Once the models for each of the 10 statistics were trained and tested, we used them to predict each player's performance for the next season (2021-22).

- Here we see Lionel Messi's predicted values, and rating, for all metrics.
- e.g. see Goals ("Gls"): {.5 goals/game} & comparative rating {81 percentile):



A historical line chart shows the progression from the past four years and into our predicted **Gls** rating for 2021-22. Note that by model design, his predicted rating was affected by the new team and league he transferred to.



Visualizations

- We used Tableau Desktop to create a visualization which is published to Tableau Public. Its 3 main components are:
  - A **highly interactive Player Viewer** where clicking on a player in the scatter plot immediately shows predicted values, ratings, and historical context
  - A **Skill Finder** where club managers can filter by each of 10 predicted ratings to find the best player suited to their needs
  - A simple **Player Sort** table to quickly find the top player for any category
- There is much more to the visualization this not shown here! **Please visit the AFC Richmond Viewer and explore for yourself.**