The Quality and Progression of Europe’s Soccer Players

Presentation by: Daron Assadourian, John Pesanello, Michael Damsky

Introduction

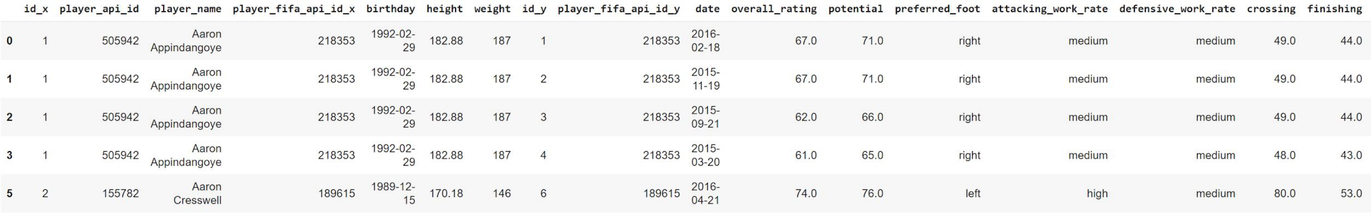
- We wanted to discover if we could predict how good a soccer player will become, given a multitude of attributes assigned to him at various points during his career.
- There is one attribute, Overall Rating that is assigned to a player on a scale of 1 to 100, so we want to predict any given player’s ‘peak rating’.
- This project would be very important to a general manager of a soccer team to evaluate young players that they are interested in signing or trading.

Dataset

- The data we analyzed came from the “European Soccer Database” dataset posted on Kaggle by Hugo Mathien consisting of seasons 2008-2016
- We focused on the “Player” and “Player_Attributes” tables
- “Player” consists of about 11000 rows, one for each player, and 7 features which include height, weight, birthday and player name
- “Player_Attributes” consists of about 183000 rows, one for each player rating update, sourced from EA Sports' FIFA video game series
    - This table had 42 features, corresponding to different in game attribute ratings from the video game series, including other features such as date of the update, and player ID

Dataset

- To get all the necessary data into one table, we merged Player and Player_Attributes on player_api_id
- Doing so allowed us to have access to all player data, from which we created an age feature from their birthday and when an update was released
- The combined table has about 153000 rows and 49 features, as a result of restricting the age of players to 21+
  - We deemed data to be too inconsistent for players under this age



	id_x	player_api_id	player_name	player_fifa_api_id_x	birthday	height	weight	id_y	player_fifa_api_id_y	date	overall_rating	potential	preferred_foot	attacking_work_rate	defensive_work_rate	crossing	finishing
0	1	505942	Aaron Appindangoye	218353	1992-02-29	182.88	187	1	218353	2016-02-18	67.0	71.0	right	medium	medium	49.0	44.0
1	1	505942	Aaron Appindangoye	218353	1992-02-29	182.88	187	2	218353	2015-11-19	67.0	71.0	right	medium	medium	49.0	44.0
2	1	505942	Aaron Appindangoye	218353	1992-02-29	182.88	187	3	218353	2015-09-21	62.0	66.0	right	medium	medium	49.0	44.0
3	1	505942	Aaron Appindangoye	218353	1992-02-29	182.88	187	4	218353	2015-03-20	61.0	65.0	right	medium	medium	48.0	43.0
5	2	155782	Aaron Cresswell	189615	1989-12-15	170.18	146	6	189615	2016-04-21	74.0	76.0	left	high	medium	80.0	53.0

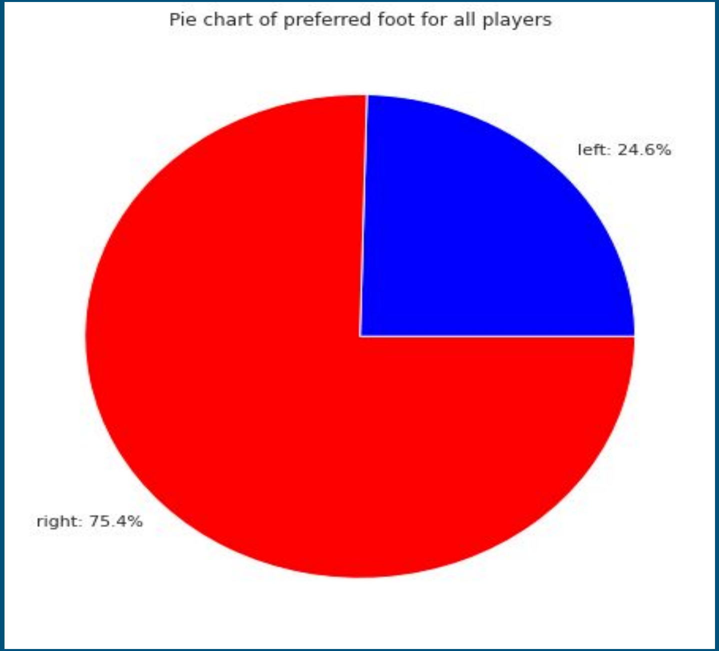
Approach to The Problem

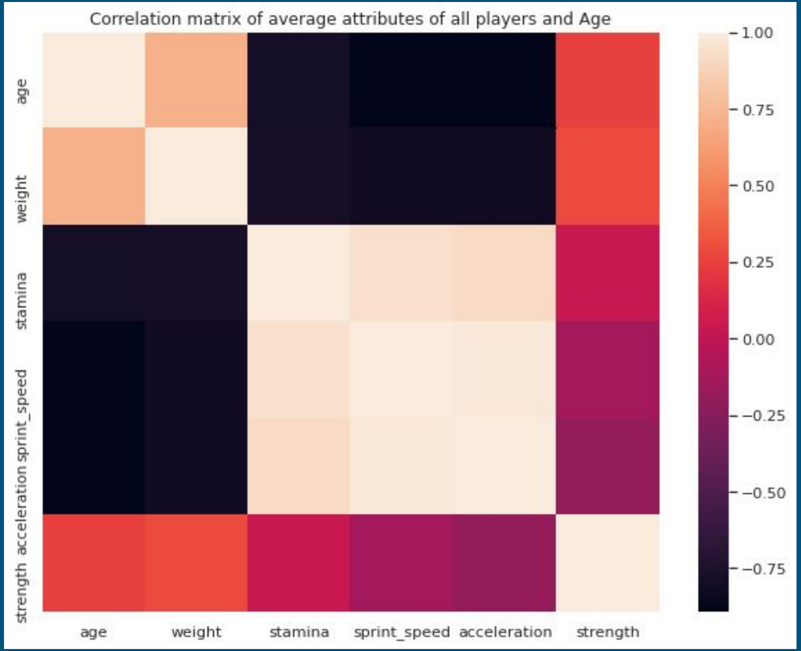
- We went through several steps in our process before we were comfortable making our predictions, as we had to intensely analyze our data.
- We used many techniques from class, including data preprocessing, data visualizations, dimensionality reduction, clustering, decision trees, and linear regression.

Learning About the Data

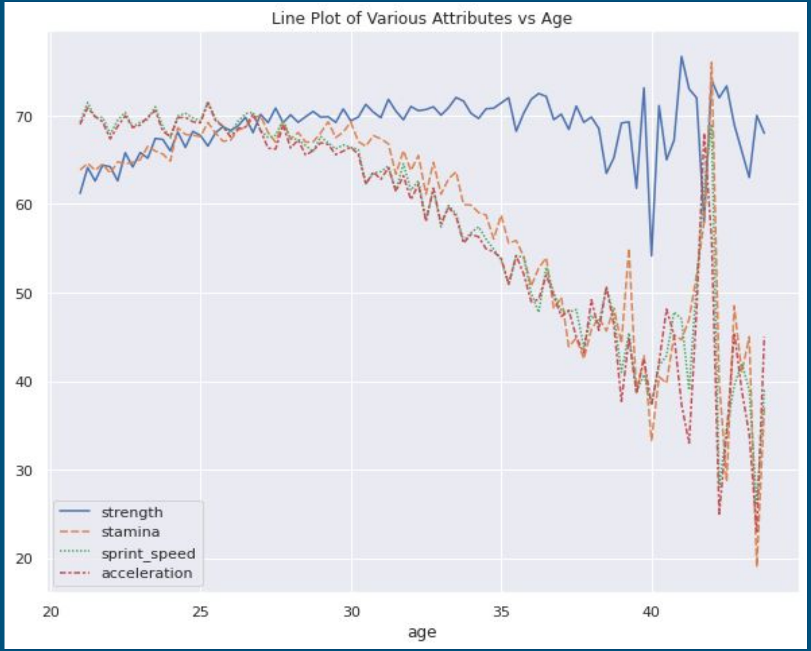
- We began our project by using data preprocessing techniques to manually create some important features that were missing from the dataset, features such as age and peak rating.
- We then used data visualizations to learn more about the data that we had, to learn which features worked together.

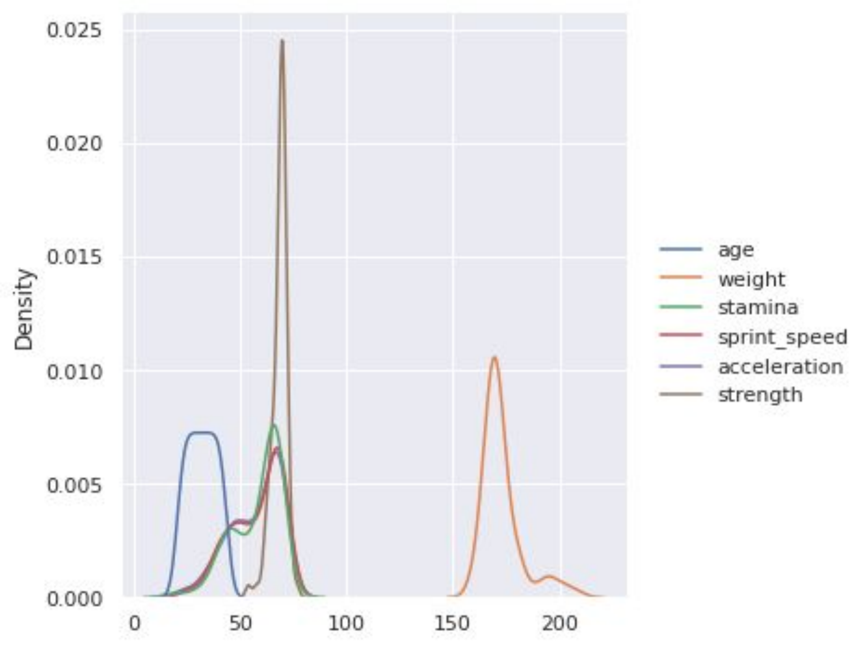
Analyzing Attributes



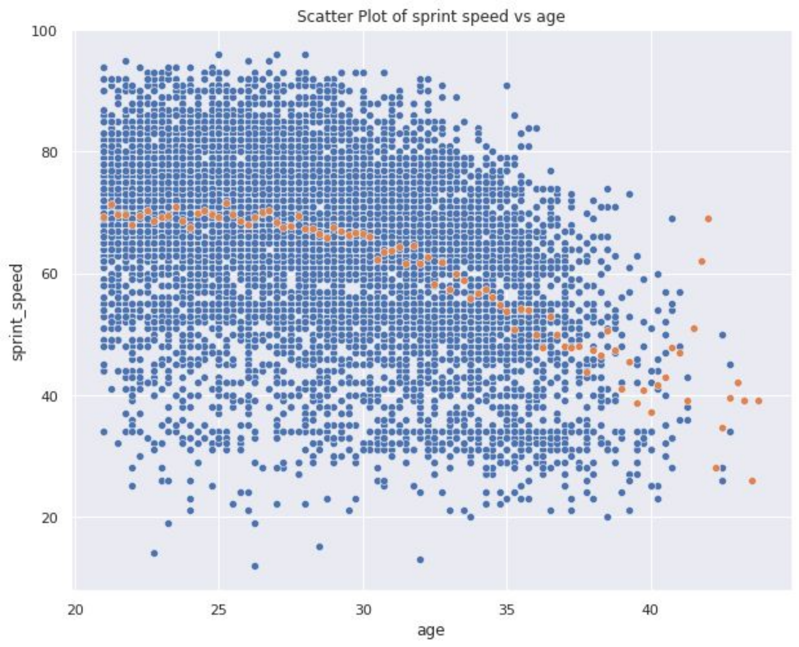


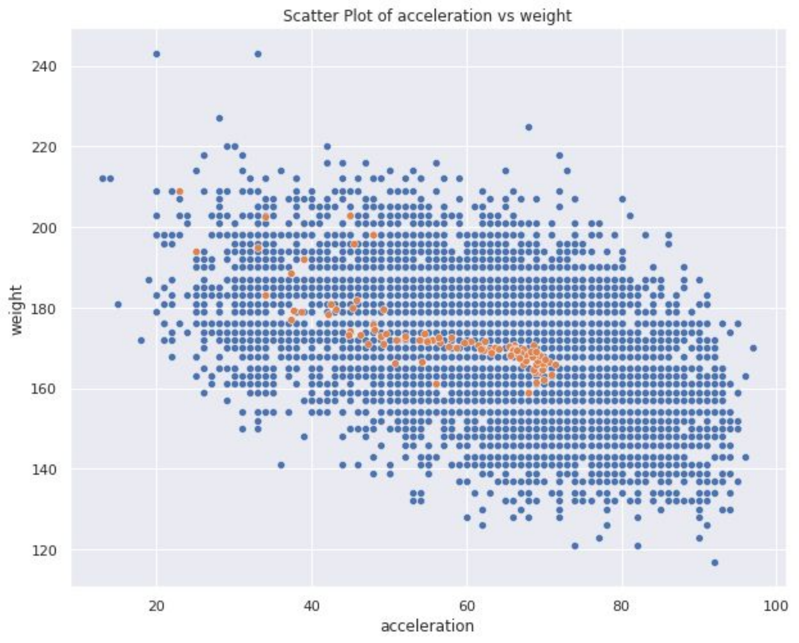
Progression Over Time and Distribution





Scatter Plots and Correlation

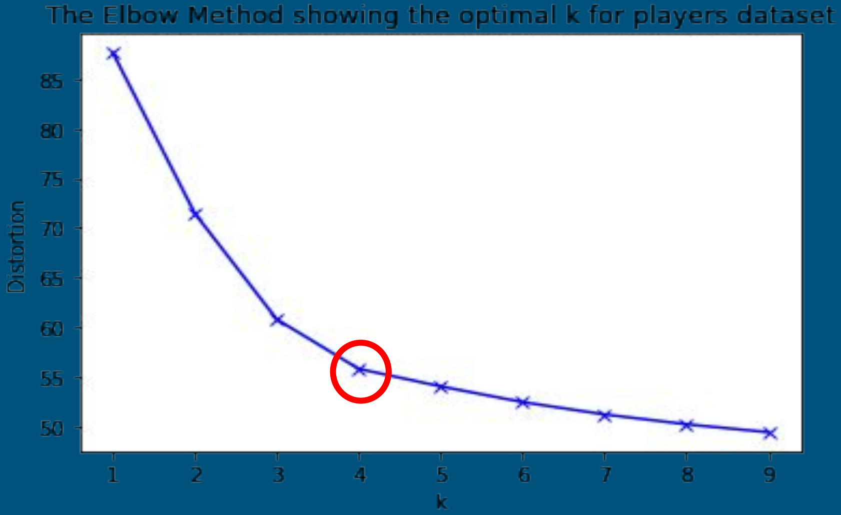


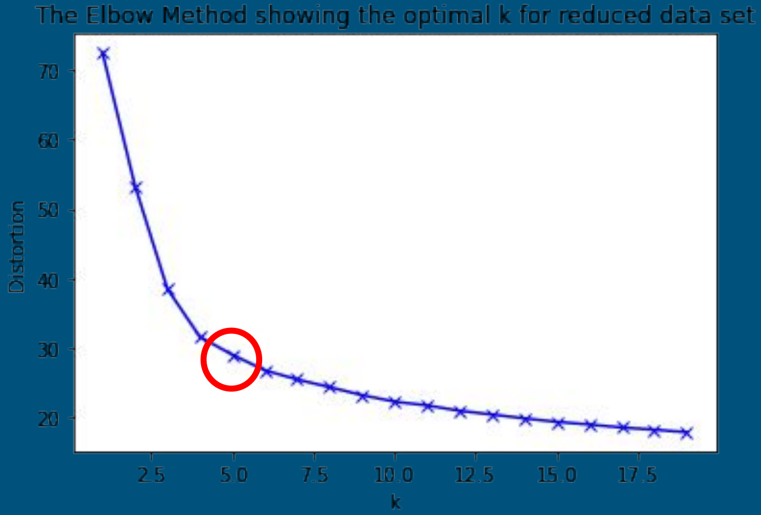


Reducing and Clustering The Data

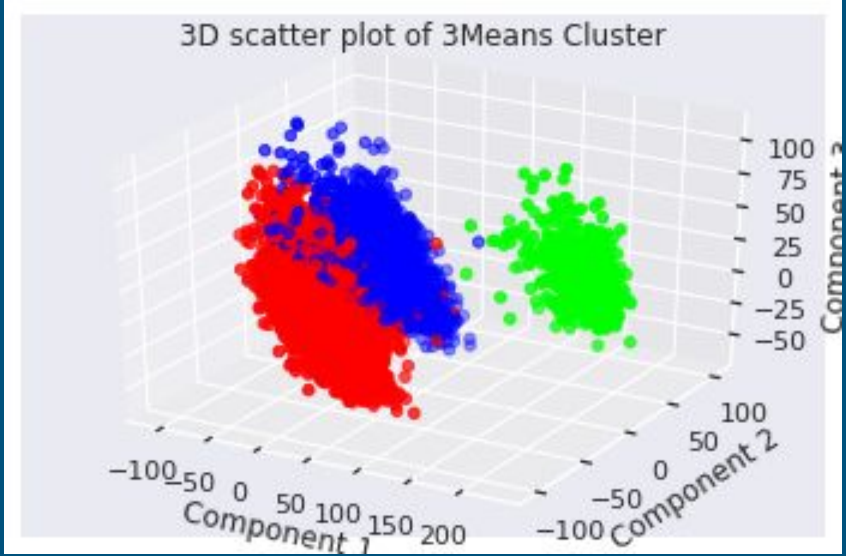
- Since the dataset had over 40 player attributes, we used Principal Component Analysis to reduce all of the dimensions down to three, so that we could visualize the similarities between players.
- We used k-means clustering to group players together based on their attributes to see if we could learn anything about the features and which players possessed similar features.

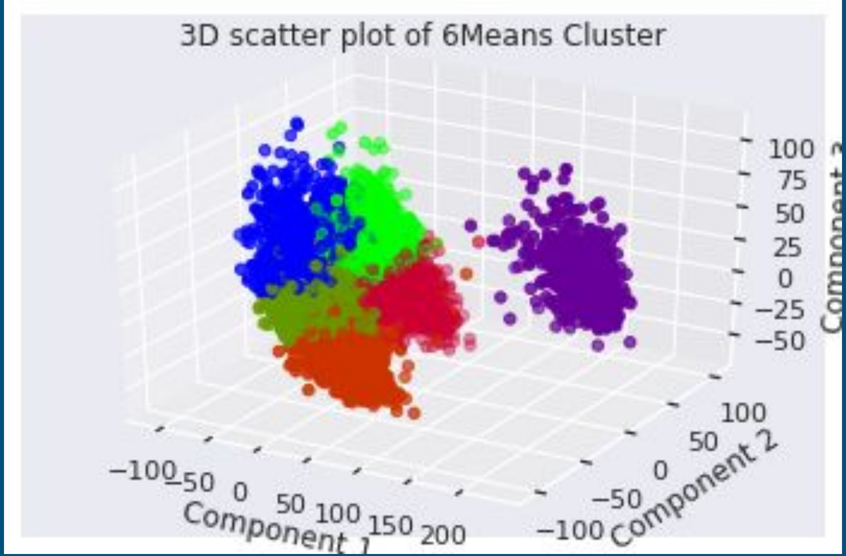
Determining Optimal K for Two Datasets





3Means & 6Means Clustering Interpretation





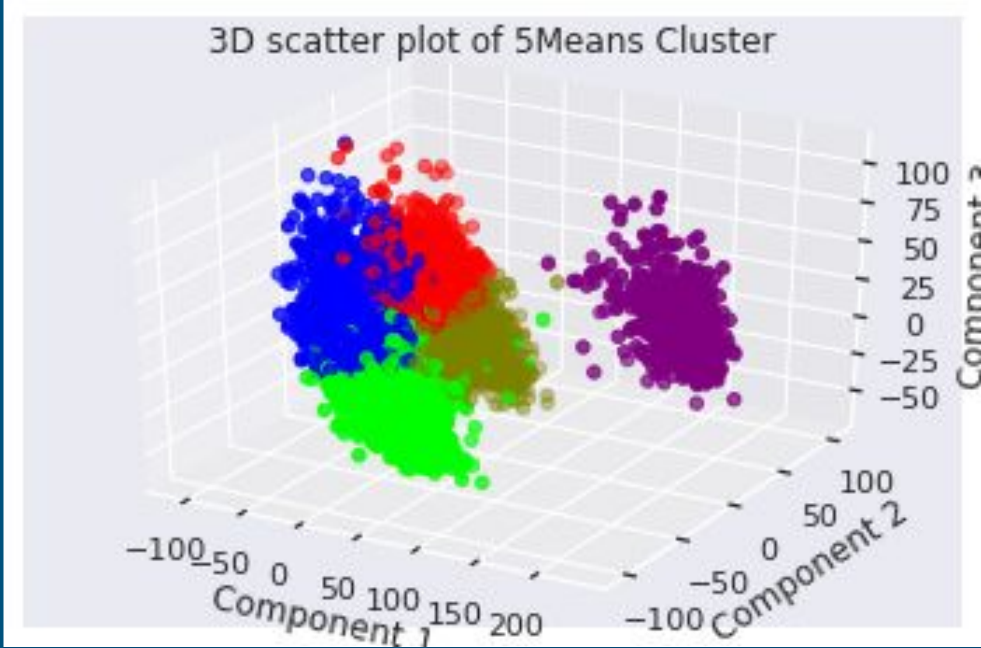
5 Means and the Interpretation

We noticed clear separation of

Clusters containing the goalkeepers,

the good attackers, the bad attackers,

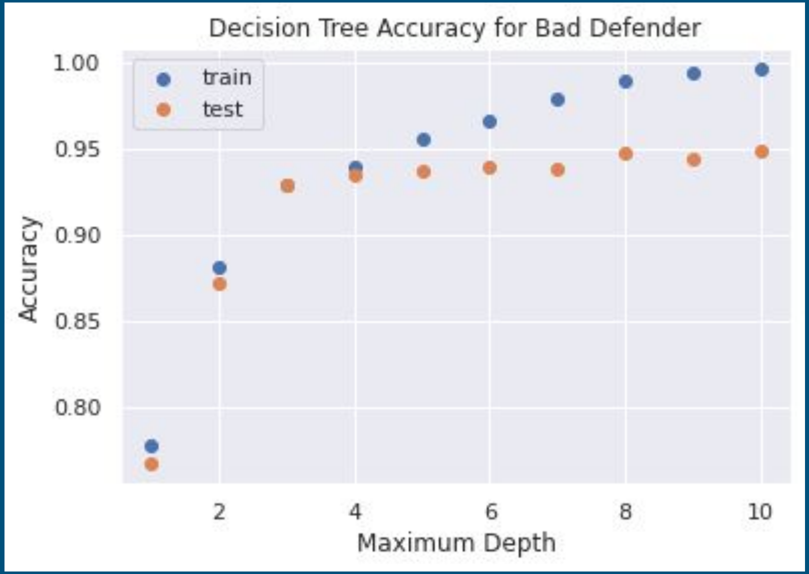
the good defenders, the bad defenders

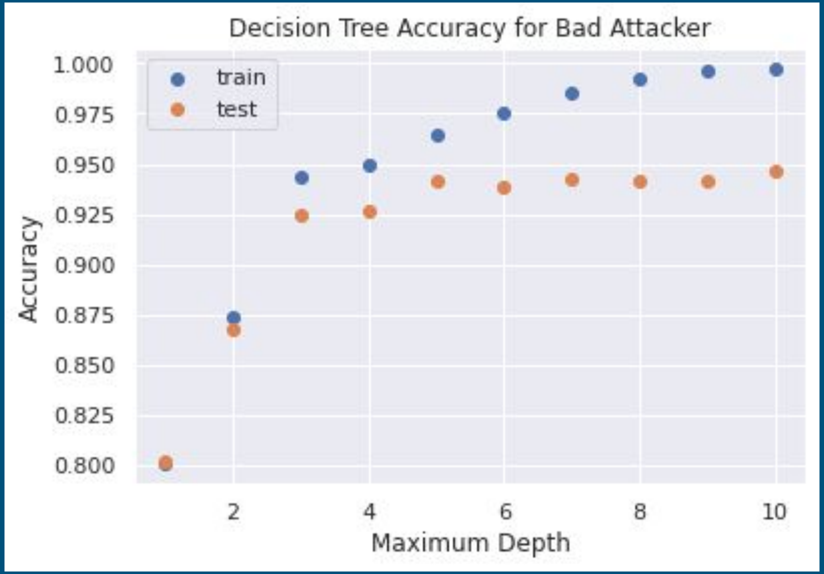


Classifying The Players

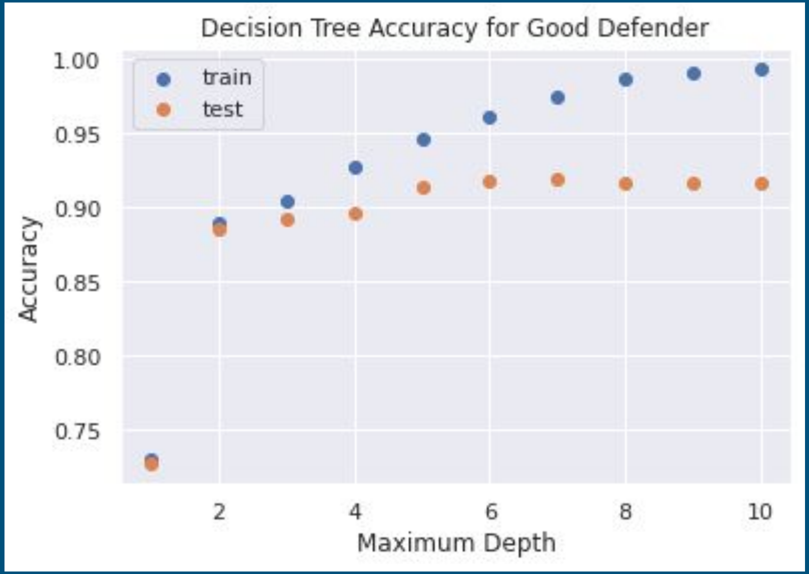
- After we noticed separations in the clusters, we also used decision trees to classify which position a player would play
- Overall, we wanted to see if, by using attributes, we could classify whether the player is an attacking or defensive player, and whether they are a “good” or “bad” player, with the target based on previous clusters
- The data was split 67% going into a training set and 33% going into a test set
- The features used for these 4 decision trees were height and weight, along with numerical attributes

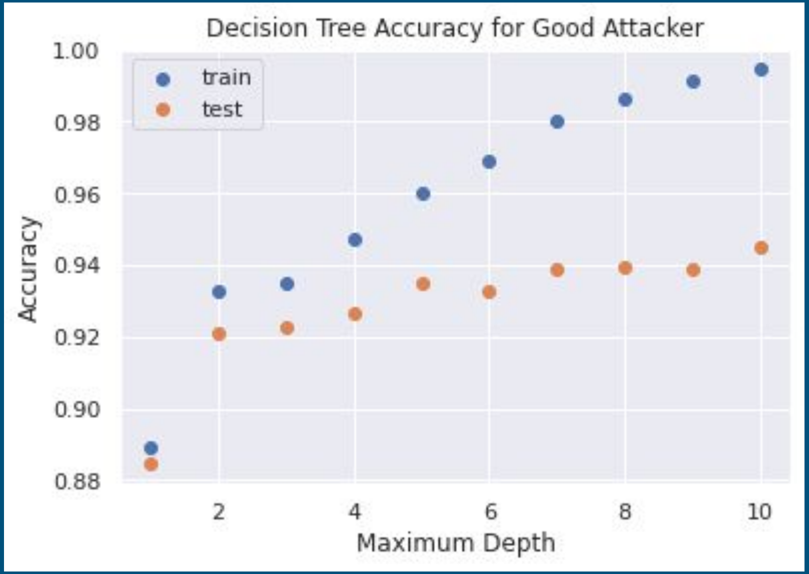
Decision Tree Accuracy for “Bad” Players





Decision Tree Accuracy for “Good Players”





Making Our Predictions

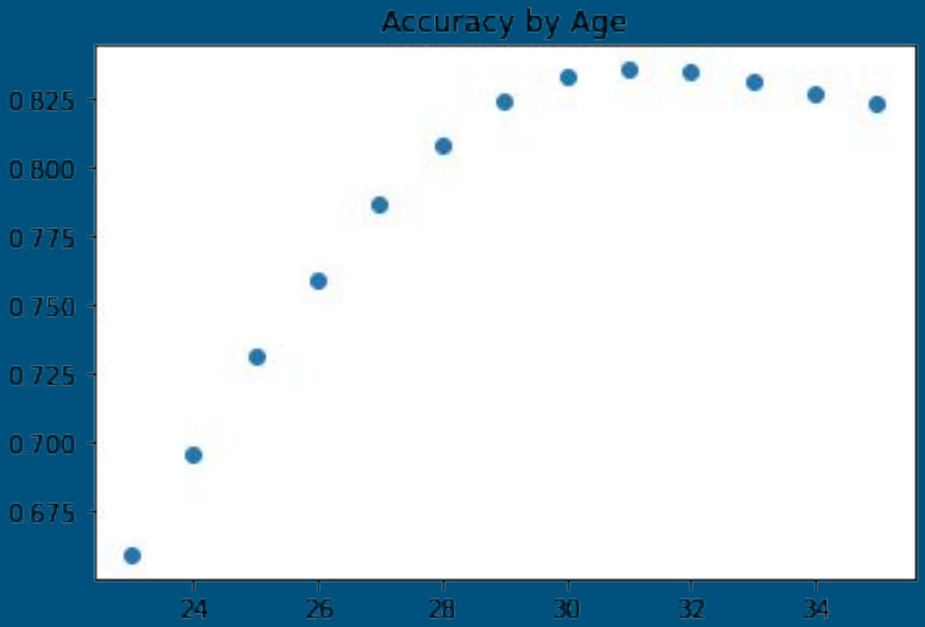
- After diving deep into the data and learning about the features and which players were similar, we used a linear regression model using the player attributes to predict the peak rating of every player.
- In order to get the best group of players to train on, we used players that had long careers to train the model, as we believed these players would have already passed their peak.
- We then tested on the younger versions of these older players to get our accuracy measures.

Changing Accuracy

The accuracy, within 4 points, of the model depended on which test set we used.

The average age players peak is around 29 years old, so the best accuracy is when we are making predictions for those players.

Prediction accuracy declines again as players pass their peak.



Real World Implications

- With the increasing use of data in the world of sports, teams will want to use tools like this one to evaluate players and design the best team possible.
- This tool is not quite ready to be used by teams, it needs more data and testing before it can be trusted by a billion dollar soccer team.
- When tools like this are used, they can sometimes make surprising predictions that can upset fans or cause considerable financial misfortune to teams, so making an accurate prediction is very important.

Conclusion and Future Work

- We found that the datasets contained informative data that was easily interpretable in various ways, hence our models and techniques yielded desired results.
- We would like to utilize more features to find more complicated answers
- Example of the above: How many points in various attributes is a player projected to change over a few years?
- Another Example: Scraping price values and goals of each player from the internet, combining it with current data and predicting price values of each player for future seasons