Good evening everyone, we are the Soccer Scientists. The team is comprised of myself Brian Clark, the incredible Eric Seitz, the unequaled Pallavi Pai, and the indelible Saharsh Gupta

History

- The first game of soccer is believed to have occurred in Mesoamerican cultures more than 3,000 years ago
- Versions of the sport have been played on many continents by many different cultures
- The game of soccer as we now know it, originated in the 19th century in England
- Known as 'The Beautiful Game', soccer has and continues to bring people together all over the world

Since soccer does not enjoy the popularity in the United States that it has in many other countries, we felt it was appropriate to give a little background and description of the sport that has come to be known as "The Beautiful Game". Soccer has been around for such a long time that it is believed to predate the Greeks and Romans, and it is widely believed that both the Greeks and Romans played a version of the game as way of training and preparing their soldiers for battle.

Rules

- Soccer has had a turbulent history, and at one point it was considered such a violent sport that it was banned in London by the mayor!
- In 1862, a semi-professional soccer player named Cobb Morley wrote the 'Laws of the Game'
- The 'Laws of the Game' contains all the rules for the sport
- While the rules have been modified, they've remained largely unchanged since they were originally written

Despite now being known as 'The Beautiful Game', soccer has had a rather turbulent history. In the 14th century it was considered such a violent and sinful sport that it was banned from being played by the mayor of London. It would be centuries before the official rules for the game of soccer were actually established by a semi-professional soccer player named Cobb Morley.

Rules Cont'd.

- Each team in soccer is comprised of 11 players
- The only player that can use their hands is the goalkeeper
- Players wear limited protective equipment, commonly only a shin guard to protect themselves from injuries
- In order to win the game, the teams must get the ball into the opposing team's goal more times than their opponent

Soccer is a rather complex sport that requires a lot of skill. The rules of the game are equally as complex, and for that reason we will only touch on a few of the most important. Each team has a total of 11 players, and the ball can only be advanced down the field by kicking or 'heading' the ball by every player except for the goalkeeper who is also allowed to use their hands. The primary goal of the game is to score more goals than your opponent in order to win, but at the end of the game if the number of goals for each side are equal, the game is considered a draw

Types of penalties

- There are two penalties that can be received during a soccer match
- The first is a ‘Yellow Card’
- If a player received two yellow cards, they are then given a ‘Red Card’
- A red card means the player must immediately leave the field and is not eligible to play in the next game

Contrary to American football, which can have dozens of different penalties, soccer has a much more simplified penalty system. There are two kinds of penalty that a player can receive, a yellow card and a red card. Two yellow cards leads to a red card penalty, and the player is ejected from the game and suspended from the teams next game

Soccer Field Dimensions

A soccer field is 125 meters by 85 meters. This translates into 136 yards long and 93 yards wide, making it much larger than an NFL playing field! Clearly shown are the penalty areas, the goal and surrounding features, and the center spot where the game begins

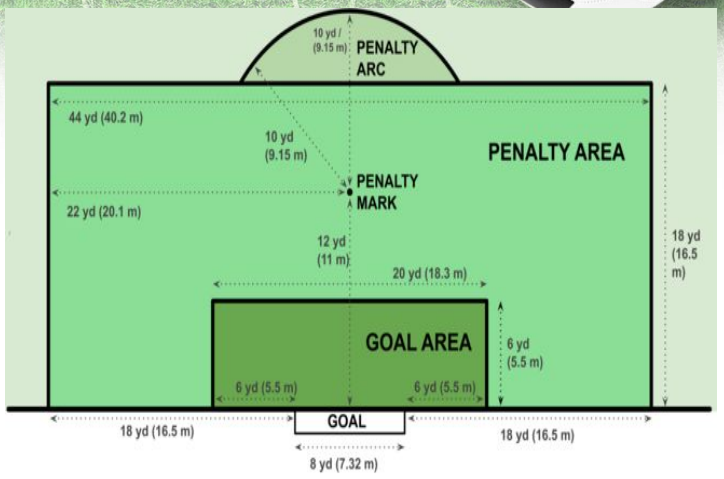
Player's Positions

A. Goalkeeper
B. Left fullback
C. Right fullback
D. Center back
E. Left back
F. Right back
G. Left midfield
H. Right midfield
I. Left forward
J. Center forward
K. Right forward

Here we can see a depiction of approximately where each positional player on the soccer field is located. The positions are typically simplified into 4 categories, 'Attackers', 'Midfielders', 'Defenders', and 'Goalkeepers'. The 'Attackers' play closest to the center line, while the 'Defenders' play closest to the 'Goalkeepers'. We have applied these categories to our data set to simplify the description of each players position

Different types of kicks

- A penalty kick is one of the most exciting moments in any soccer match
- A player, having been fouled by the opposing team in the penalty area, is given a single shot against only the goalkeeper
- All other players must be at least 10 yards from the goalkeeper



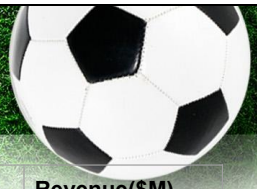
There are a couple of predominant ways in which scoring is done in a typical soccer match. Of these, the penalty kick and the corner kick are the most well-known. The penalty kick is one of the most exciting plays in soccer and tests the skills of a player one on one with the goalkeeper.

Different types of kicks

- A corner kick is awarded when the whole of the ball passes over the goal line, on the ground or in the air, having last touched a player of the defending team, and a goal is not scored
- All other players must be at least 10 yards from the corner arc during the kick

The corner kick is typically more of a team implemented means of scoring. The kicking player is within the corner arc and kicks the ball into the 'Goal Area'. The 'curve' attribute within our data set is directly linked to corner kicks, some of you may have heard it called 'bending the ball'. The kicking player can choose to try to curve the ball directly into the goal or kick towards other teammates giving them a possible chance to score.

Most Valuable Teams



Rank	Team	Country	Value(in millions)	Revenue($M)
1.	Real Madrid	Spain	4,231	796
2.	Barcelona	Spain	4,205	815
3.	Manchester United	England	3,983	765
4.	Bayern Munich	Germany	3,024	751
5.	Manchester City	England	2,688	678

From this image we can see the overall value of the top 5 soccer clubs in the world. Using this, we can make a comparison between the values of soccer clubs and American football teams. According to the Forbes list of the most valuable football clubs, Real Madrid and Barcelona, two Spanish teams, have the highest overall value. The values are in the form of U.S dollars, and both of the Spanish teams have a value greater than $4 billion dollars. The value of these soccer clubs is comparable to the value of NFL teams here in the United States.

So... Why Soccer?

Being able to extrapolate and predict a player's performance in matches would be absolutely vital for managers looking to create better and better teams.

While “Fantasy Soccer” might not be as popular in the states, building teams and making bets recreationally can be fun for friends and office pools.

Getting an upper edge on the “Best” or “Rising” players can give a huge boost in betting and team crafting.

We initially began looking at soccer teams rather than individual players, but we chose to evaluate data sets containing only soccer players as it gave us a larger data frame to work with and to be honest we wanted a topic that was a bit different from the data sets that other teams were interested in. Additionally, sports have a huge market for data science in predictions and evaluations. Soccer may not be as popular in the United States, but with the popularity it carries worldwide we were sure that we would be able to find multiple datasets containing the type of data we were looking for.

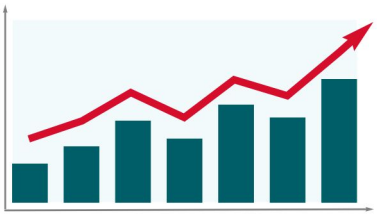
What were our first ideas?

We definitely wanted to do some sort of ranking or predictions with players.

Initially we thought of predicting a player's overall performance from past performance.

A year-over-year change.

We also thought of splitting the data up into player nationality, possibly players from certain countries had a higher chance of being star players



Our original ideas involved using data sets over a number of years, and based on those data sets we wanted to be able to predict the players ratings in additional seasons, a 'year-over-year' change. Our belief was that we would be able to demonstrate whether the players had progressed, regressed, or if their rating at each attribute remained the same. We felt that analysis such as this would be useful for predicting a winning fantasy soccer team, but found the undertaking a little more difficult than expected because of subtle differences between the data sets from year to year. Namely one data set would include players that another did not, or would list a player at a different position than other data had the same player listed at

Choosing our Data

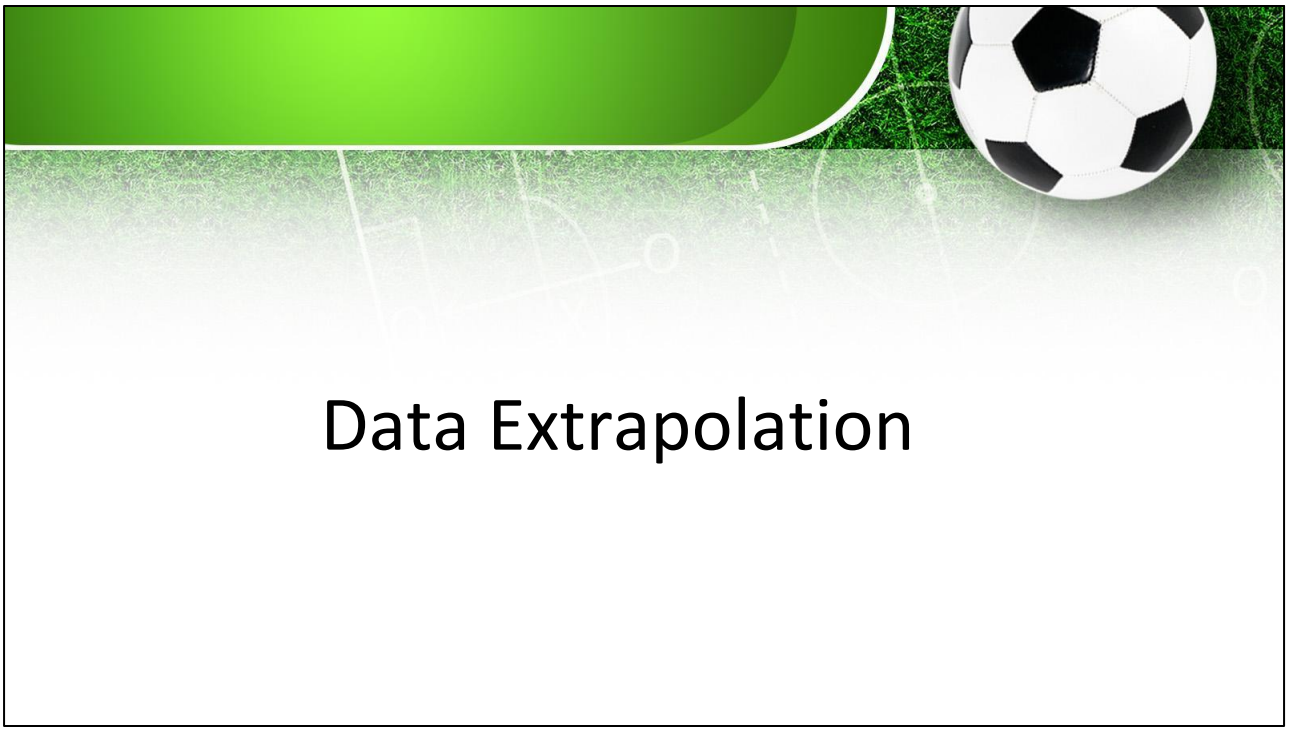
A majority of our data came from the "**FIFA 20 complete player dataset**" on Kaggle

https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset

We chose this dataset as it contained clear, detailed data on soccer player's performance from 2015 to 2020 with matching columns.

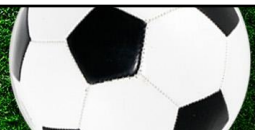
This turned out ideal for us as it meant we can easily compare different years, players, player nationalities, etc easier than trying to combine multiple sources.

As previously discussed, we wanted to choose a data set that was large enough to incorporate the amount of data exploration that we intended. Soccer has a number of years worth of data sets readily available with a surprising amount of data involved. Every attribute you could imagine is listed, from the players age and wage, to their international reputation and even their body type. The majority of the data sets that we would utilize would come from Kaggle.com



Eric's section

Dataset Column description



```
Index(['Unnamed: 0', 'ID', 'Name', 'Age', 'Photo', 'Nationality',
 'Flag', 'Overall', 'Potential', 'Club', 'Club Logo', 'Value', 'Wage',
 'Special', 'Preferred Foot', 'International Reputation', 'Weak Foot',
 'Skill Moves', 'Work Rate', 'Body Type', 'Real Face', 'Position',
'Jersey Number', 'Joined', 'Loaned From', 'Contract Valid Until',
'Height', 'Weight', 'LS', 'ST', 'RS', 'LW', 'LF', 'CF', 'RF', 'RW',
'LAM', 'CAM', 'RAM', 'LM', 'LCM', 'CM', 'RCM', 'RM', 'LWB', 'LDM',
'CDM', 'RDM', 'RWB', 'LB', 'LCB', 'CB', 'RCB', 'RB', 'Crossing',
'Finishing', 'HeadingAccuracy', 'ShortPassing', 'Volleys', 'Dribbling',
'Curve', 'FKAccuracy', 'LongPassing', 'BallControl', 'Acceleration',
'SprintSpeed', 'Agility', 'Reactions', 'Balance', 'ShotPower', 'Jumping',
'Stamina', 'Strength', 'LongShots', 'Aggression', 'Interceptions',
'Positioning', 'Vision', 'Penalties', 'Composure', 'Marking', 'StandingTackle',
'SlidingTackle', 'GKDiving', 'GKHandling', 'GKKicking', 'GKPositioning', 'GKReflexes',
'Release Clause'], dtype='object')
```

Eric

Here we can clearly see the 89 different columns worth of data that would need to be cleaned up and scrutinized. Some of the terms such as ‘ST’, which stands for striker, or ‘RDM’, which is right-defensive-midfielder, are the player’s rating at each of the different positions. Many of the columns held no value for our exploration of the data, such as ‘Club Logo’, ‘Photo’, or ‘ID’.

“Here we see 89 different columns from our dataset. These columns will be the data we clean and scrutinize for our project. Many columns have easily identifiable labels but some may not be recognizable at a glance. For example, column “ST” stands for “striker” and “RDM” stands for right-defensive-midfielder.
These are positions on the soccer field a player may be assigned to. There are also attributes such as “Strength”, “Stamina”, and “Aggression” that define how a player’s play style. Some columns, such as “Club Logo” and “Photo” have little to no value to us, but many others will come together and be useful in predictions.”

Eric
Here is a graphical depiction of the highest paid players according to the FIFA data set. Lionel Messi leads the way by a large margin over the next highest paid player, Luis Suarez.

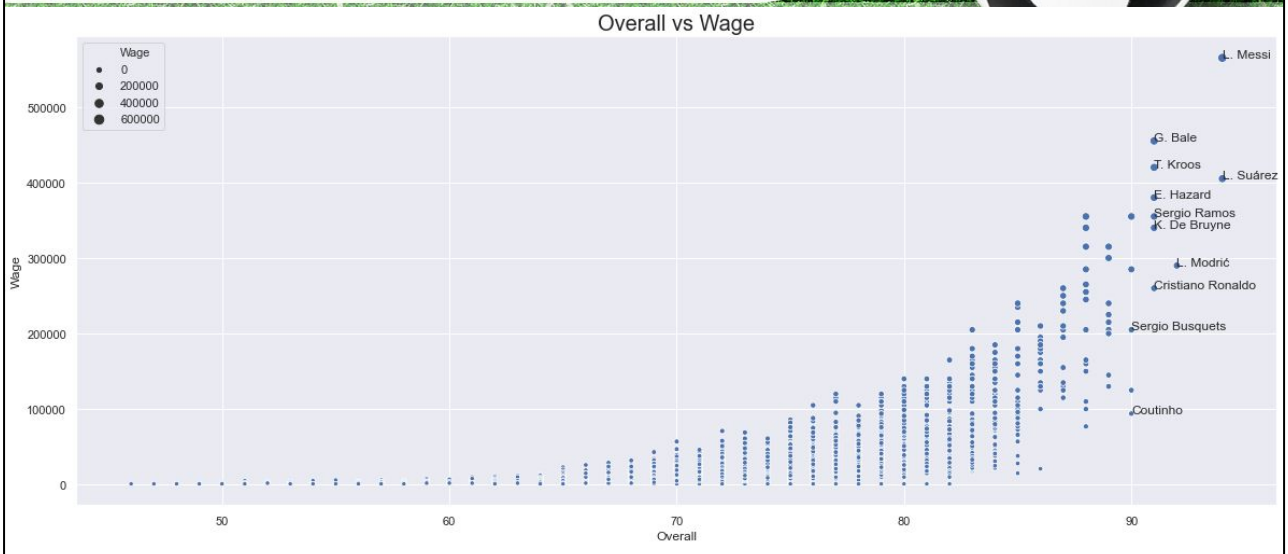
“We extrapolated data from out dataset we found interesting, mostly related to value and wages of players. Here we see the 10 highest paid players from our dataset. Lionel Messi leads by a considerable margin and shows that there may be some outliers or extreme values in our data.”

Value vs Wage

Eric
This is an image depicting the players value vs the amount that they are paid. We have marked the top 10 highest players. On this graph there is a strange anomaly on the far right, we are unable to explain why that mark is present. It does not correspond to any player from our data set, so we feel it may be a glitch within seaborn.

“Here we see a graph of all player wages mapped by value. We’ve marked the 10 highest paid players by their names. While there is a general trend upwards, notice some anomalies on the far right. We cannot explain why that dot on 1e8 exists as it does not correspond to any player in our records. This may be a glitch with seaborn. Barring that we see a general pattern, many players clustering around the same pay spot on the left, growing apart considerably as their value increases. Some players have considerable value but lower wages, this may be due to currency conversion rates inflating the “wage” of players in certain countries such as Germany or England over players from lower-cost countries.”

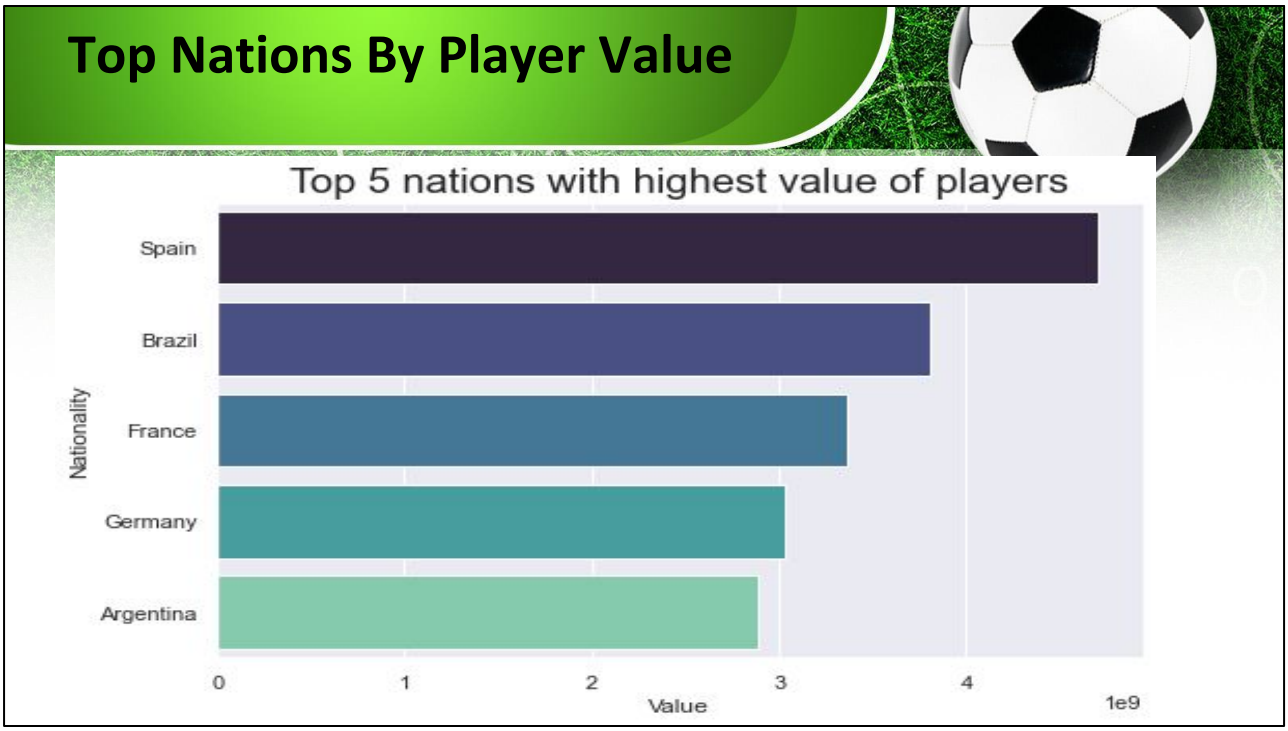
Overall Rating vs Wage



Eric
We also choose to show how players are plotted when we compare their overall rating vs the amount that they are paid. On each of these graphs it is clear to see that Lionel Messi stands out amongst the rest.

“Here we see a graph of all player wages mapped by Overall rating. Again, the top 10 players are listed by name and many familiar names are seen. We see a general trend upward at a rate that looks to be exponentially increasing, unlike wages which seemed more linear. Notice how many players seem clustered in the 75-85 Overall rating range, this makes sense as these players would be considered “above average” but much fewer being considerably better than all the rest. Think of this like class grades, where 70% is considered “Average””

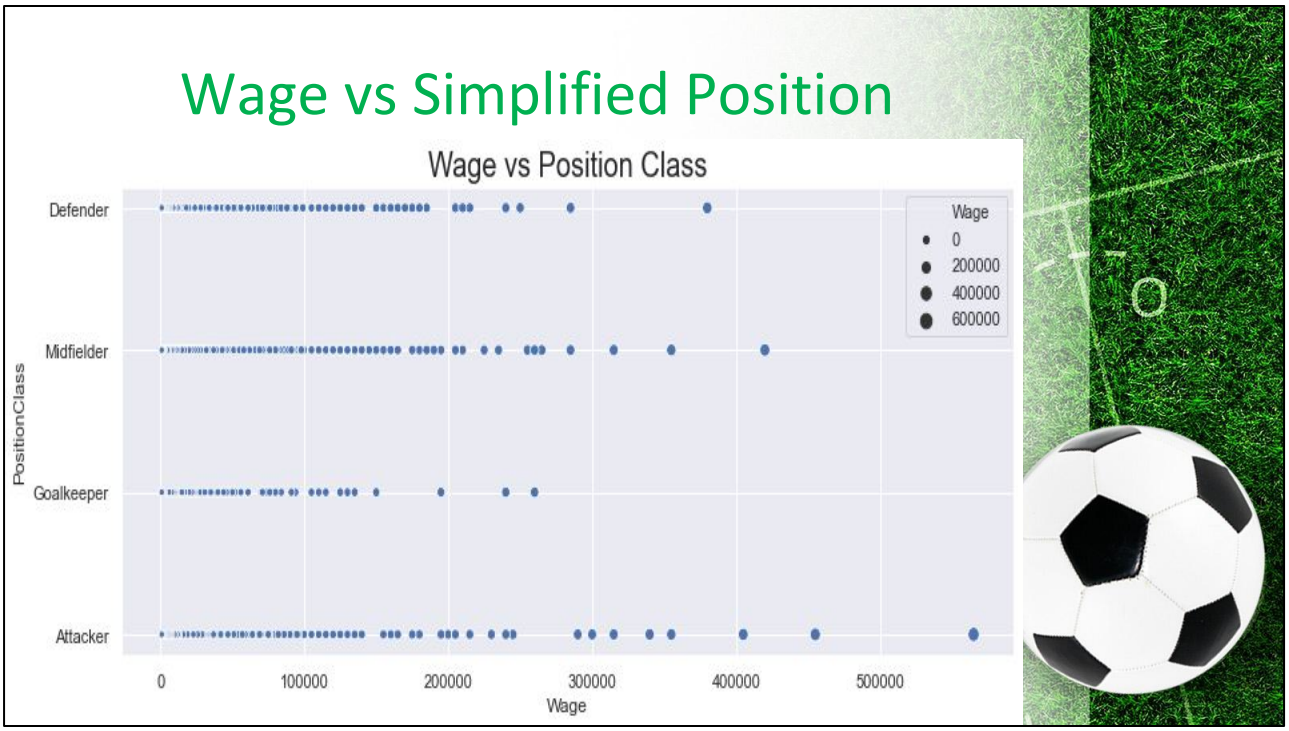
Top Nations By Player Value



Eric
Ironically, before beginning our analysis of the data set, as a team we had decided that we would like to look at 5 individual nations. We all agreed upon Argentina, Brazil, Italy, Germany, and Spain. After we decided on those nations, we found that 4 of our choices happen to be amongst the nations with the highest value of players.

“Here we see countries with the highest VALUE of players. Originally we wanted to look at 5 specific countries and their players, we chose Argentina, Brazil, Italy, Germany, and Spain. After choosing these countries based on instinct, we found 4 of our choices happened to be among the nations with the highest valued players.”

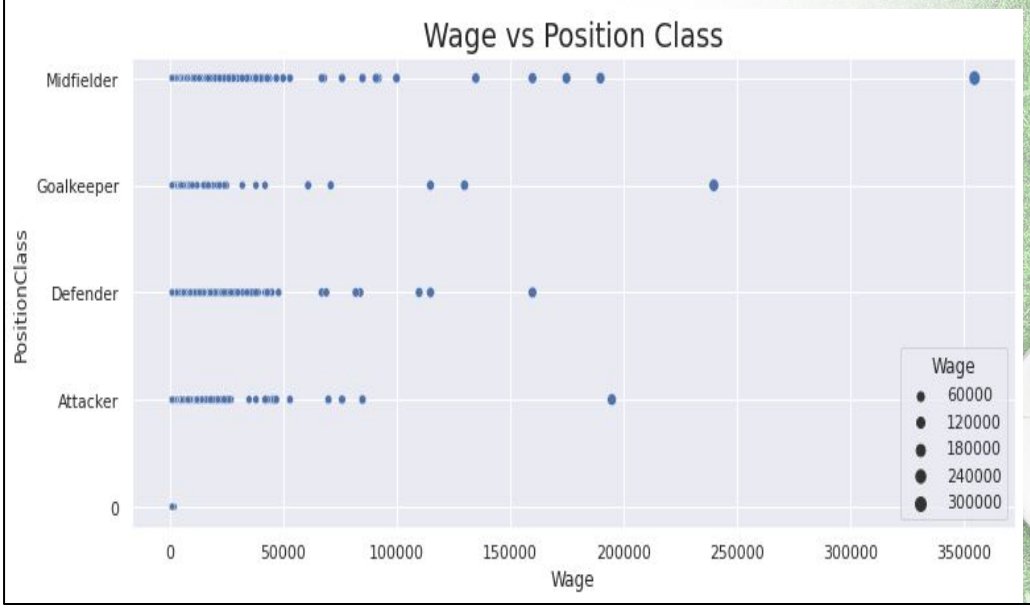
Wage vs Simplified Position



Eric
From this analysis on the entire dataset we can see that an Attacker position has the highest wage, followed by Midfielders, then defenders and lastly are goalkeepers when it comes to how much players are paid. However, these statistics change when we look at the dataset for each individual nation.

“Here we see a graph of all player positions mapped by wage. Notice how attacker has several outliers considerably higher than the others. This data combines all countries however, this will change once we look at the dataset of each individual nation.”

Wage vs Simplified Position - Germany



Eric
This is an example of the wage vs position breakdown for players only from the country of Germany as compared to the previous slide which was applied to the entire data set. Rather than ‘Attackers’ being the highest paid players, ‘Midfielders’ from Germany tend to make more, followed by ‘Goalkeepers’. The anomalies that appear at the bottom of the graph are depicting the players that could not be categorized into one of the 4 simplified position categories. This is because they are known as a ‘sub’ within the data set.

“Here we see the data from Germany alone of all player positions mapped by wage. Now Midfielder has an outlier considerably higher than the rest. Anomalies shown on the bottom left of the graph are players that could not be categorized into one of the 4 simplified position. These are “Sub” players within the dataset, or players can be moved into several different positions and not just assigned to one typically.“

Best of the Best

- **Best Crossing :** K. De Bruyne
- **Best Finishing :** L. Messi
- **Best Heading Accuracy :** Naldo
- **Best Short Passing :** L. Modrić
- **Best Volleys :** E. Cavani
- **Best Skill Moves :** Cristiano Ronaldo
- **Best Dribbling :** L. Messi
- **Best Sprint Speed :** K. Mbappé
- **Best Agility :** Neymar Jr
- **Best Reactions :** Cristiano Ronaldo
- **Best Balance :** Bernard
- **Best Shot Power :** Cristiano Ronaldo
- **Best Jumping :** Cristiano Ronaldo
- **Best Stamina :** N. Kanté

Eric
This is a description of just some of the best players according to their ratings in different categories. Lionel Messi and Christiano Ronaldo are two of the most skilled and well known players in the world, so there is no surprise that they are shown as being the best in multiple categories.

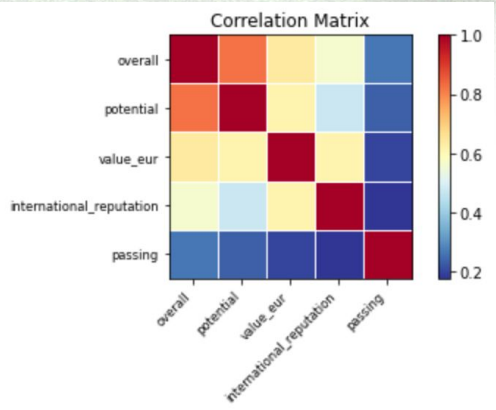
“Here we see the ‘Best of the Best’ according to several different categories. Lionel Messi once again pops up along with Christiano Ronaldo as the two most skilled and well known players in the world. It’s no surprise they appear in several categories. Highly valued and paid players would have multiple skills they excel at, or at least that’s the assumption. We can see evidence of that assumption here”

Correlation Coefficient - Overall

From a Linear Regression run on the Yearly_Brazil we were able to determine the correlation of variables to overall scores.

sofifa_id	-0.369888
physic	0.170639
pace	0.124157
shooting	0.215751
dribbling	0.226966
passing	0.265954
defending	0.213632
international_reputation	0.563724
height_cm	0.010418
weight_kg	0.041050
age	0.174653
value_eur	0.641934
potential	0.827021
overall	1.000000

Name: overall, dtype: float64



By this correlation coefficient run we see a clear correlation with 'potential', 'value_eur', and 'international_reputation' as the strongest variables.

Weight and height seem to have the least impact on overall scores

Eric
Might be good to talk about what we found correlated the most (if that would be interesting).
This linear regression was run on the "Yearly-Brazil" data set, data made from 2015-20 data. This is NOT the same data set as the "Regression models" run.
Notice how "Potential" and "Value" make up the largest factors for "Overall" scores. This goes against my prediction of physical attributes (height, weight, shooting, passing, etc) being more substantial.

"Here we start to see some data extrapolation and predictions. Using a dataset compiled from Brazilian players over the year 2015 to 2020 we ran a Linear Regression model to determine the correlation of variables to the Overall rating. Our assumption was that height, weight, and physical attributes would have a decent impact on score but instead we see something interesting. Based on the Correlation Matrix we see that "potential" and "Value" have the highest impact on overall rating, with physical attributes having a relatively small impact.

Correlation Coefficient - Overall, method

The ‘overall’ correlation coefficient listed before was gathered from a dataset containing data of Brazilian players only from 2015-2020.

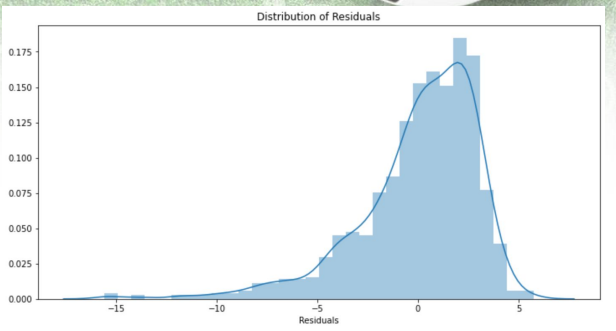
A simple Linear regression model was used, providing us an accuracy (in this instance) of 74% and a Root Mean Square Error of 3%.

Here we see predicted results

	overall	Predicted	Residuals
301	74.0	69.855133	4.144867
495	80.0	79.087082	0.912918
3033	70.0	68.250325	1.749675
2854	67.0	65.398135	1.601865
836	67.0	66.352302	0.647698
...	...	...	...
3075	65.0	64.853865	0.146135
3947	66.0	65.790322	0.209678
1246	82.0	84.840553	-2.840553
3781	67.0	66.279184	0.720816
2473	72.0	69.512443	2.487557

[1071 rows x 3 columns]

This linear regression model seems to be less accurate than others as little fitting or fine tuning was done to the model itself.



As seen by the distribution of residuals, a majority fall within +5, -5 which is expected. Extreme values past this are not ideal.

Eric
This linear regression model is NOT the same one listed under the “Regression Model” slide. This was a seperate run on different data to find coefficients and such as a side project.
This is basically just showing some data from the run “Correlation Coefficient”, how accurate it is and the distribution of residuals. Why does the distribution look good? Because it hovers around +, - 5

“Here we see some data related to the Correlation Coefficient seen previously. We measure how accurate the simple Linear Regression model was at 74%, with a Root Mean Square error of 3%.
By the distribution of residuals we see a majority fall within the +-5 range which is what we expect. The closer to 0 the better.
This model was less accurate as our upcoming main model as little fitting or tuning was done to the model itself. It was completed separately.

Correlation Coefficient - Player Skills

Here we see the correlation between player's skills.

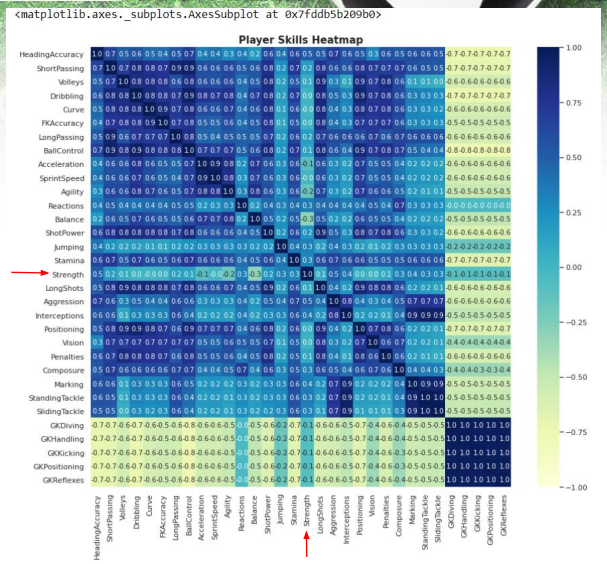
While many skills appear to be correlated on some level notice some anomalies:

- Strength and Balance yields -0.3
- Strength and Acceleration yields 0.0
- Strength and Agility yields 0.2

In fact, Strength seems to have little correlation with many other "physical" skills like accuracy, dribbling, etc.

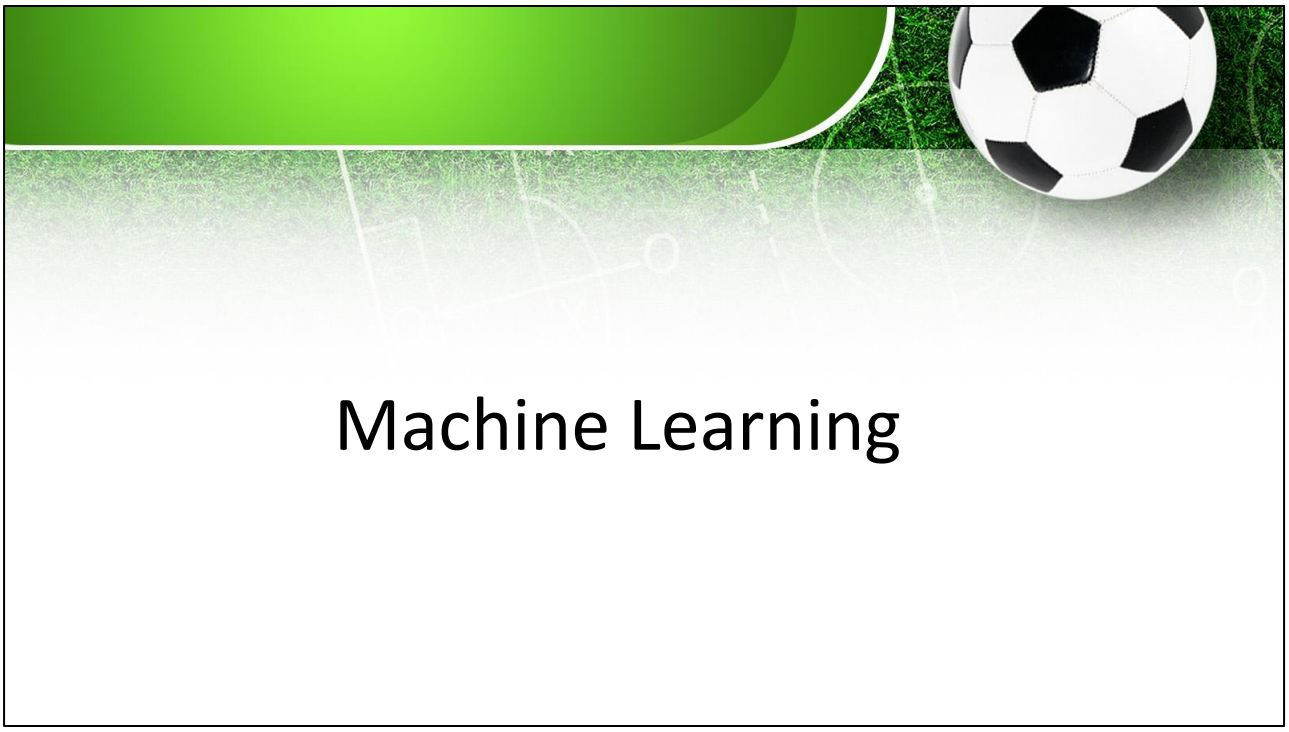
This seems unusual at a first glance and could be useful to determining how valued a player is.

A player being "strong" doesn't necessarily mean they're "good" in other areas.



Eric
This image is a comparison of player's skills against other skills within the data set to depict which ones have the biggest impact upon one another.
The chart might be a bit large and difficult to decipher, but here we focus on "unusual" patterns, as can be seen with Strength.
It would seem that strength actually has the least amount of correlation with other attributes. For instance higher strength tends to lead towards a lower level of balance and agility.

"Here we see a correlation coefficient of player skills. The graph may be intimidating at first but it yields some interesting data.
Notice the anomalies pointed at with red arrows, this is the "Strength" column, we see that contrary to popular belief strength has little impact on attributes like Agility, Acceleration, or Balance.
This tells us that our assumption that "strong" players are "good" players may be incorrect, an interesting observation."


Machine Learning

Preparing the data

- Many of the columns needed to be modified in some way so that the values could be used
- For example, the value and wage columns were in the form '€123.4M', and this would need to be turned into a useable value
- Other columns such as each positional rating like 'ST' and 'RDM' would have expressions rather than values
- These would be in the form of 87+3 or 84 – 7, and these numbers simply indicated how that players rating at that position had changed since the previous year

When we began looking at the data sets that we intended to use, we found that there were a number of attributes that contained values that could not be used for data exploration. For example, any of the columns that contained a value referring to an amount of currency, the value would be accompanied by the Euro currency symbol and a letter signifying thousands or millions. The ratings that were given to each player based on position was actually more of a mathematical expression. The additional + 3 or - 7 was there to indicate how the player had progressed or regressed based on the data from the prior season.

Preparing the data

- In total we have 85 features in the dataset
- Of these 85 features, 40 are already in a numeric format
- We applied some regex to strip the Euro symbol and convert the ‘K’,’M’ values in Wage and Value columns into comparable values
- Converted the Positional Rating features from expressions to float, stripped ‘lbs’ from weight and converted the height column to a representable number to be used.
- After the preparation we ended up with 72 features which can be used as ML inputs

With initially having 85 features in our dataset there were approximately 40 which were already in a numeric format and had no problems. A few had to be converted from string to float format, for example the value and wage were converted from Million or Thousands notation and required the correct number of 0’s to be added. The Positional rating columns were cleaned of the + and - signs so that they could be used as an actual value. After all of the cleaning that was necessary, we found that we had 72 features which could be used of the original 85

Principal Component Analysis

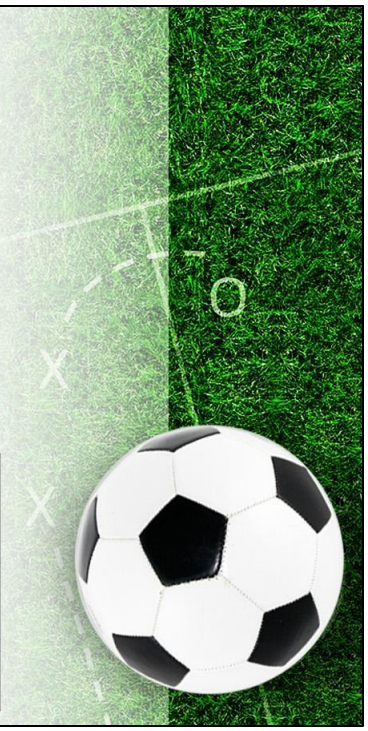
- Since we now had 72 features which can be used as machine learning inputs, we used PCA to reduce the feature space.
- Using sklearn PCA library, we transformed our scaled data into principal components.
- Using the n_components parameter we can transform our data into the given number of components, a float value for this indicates a threshold.
- To retain most of the information we used PCA with n_components equal to 0.99, and hence retaining 99% of the original information after being transformed.
- The 99% of the information of the 72 features was contained in around 27 transformed features.

By utilizing principal component analysis we were able to select which features would retain the most amount of information. We began with 89 columns of features, after data preparation we were able to decrease the number of features down to 72. We were further able to reduce the number of features down to just around 27 features that would retain 99% of the information needed for our analysis depending on the dataset.

Regression Models

- To begin, we decided to look at regression models on the entire data set
- A Mean-Absolute-Error score of 0.0 is the best achievable
- The highest “R^2 score”, or coefficient of confidence, possible is 1.0

	MLP Regression	Linear Regression	Random Forest Regression
Mean-Absolute-Error	**.199**	**.287**	**.134**
R^2 Score	**.938**	**.853**	**.967**

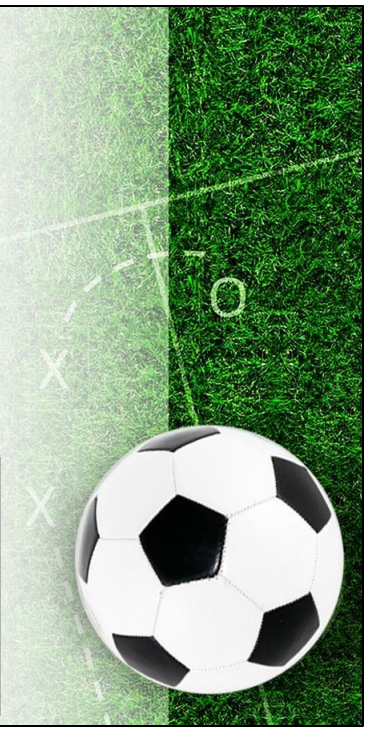


As we can see from the numbers within the table, we were able to attain high scores with our original machine learning models. Random Forest Regression was the best model with a coefficient of confidence of .967, and a mean-absolute-error value of .134. Both numbers are very close to the optimal scores that can be achieved. Linear regression did not fare as well because there is not a clear linear relation within this data set.

Random Forest Fine Tuning

- As Random forest had better results, we applied some hyper-parameter tuning with 3 fold cross validation.
- Used RandomizedSearchCV for a random search within the whole hyper-parameter space.
- Used GridSearchCV for an exhaustive search using the best parameters found from Random Search as a baseline.

	Base Model	Random Model	Best Random Model
Mean-Absolute-Error	**.130**	**.149**	**.120**
R^2 Score	**.967**	**.959**	**.967**

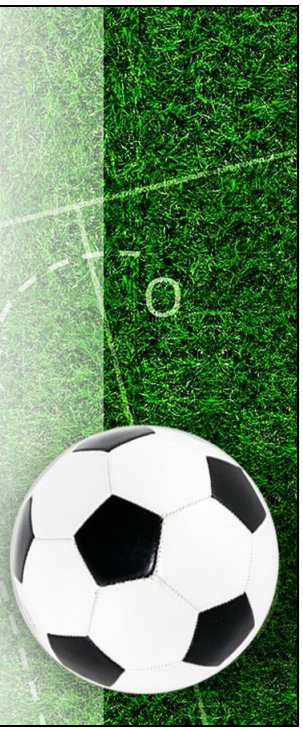


As we can see from the table the tuning did not result in much improvement overall, the best random model after tuning does have the lowest mean absolute error but same coefficient of confidence. This tuning was applied to the regression models used on the Germany data set.

Random Forest Fine Tuning

```
Fitting 3 folds for each of 300 candidates, totalling 900 fits
[Parallel(n_jobs=-1)]: Using backend LokyBackend with 2 concurrent workers.
[Parallel(n_jobs=-1)]: Done  37 tasks      | elapsed:  1.3min
[Parallel(n_jobs=-1)]: Done 158 tasks      | elapsed:  5.1min
[Parallel(n_jobs=-1)]: Done 361 tasks      | elapsed: 11.3min
[Parallel(n_jobs=-1)]: Done 644 tasks      | elapsed: 20.5min
[Parallel(n_jobs=-1)]: Done 900 out of 900 | elapsed: 28.7min finished

Fitting 3 folds for each of 1875 candidates, totalling 5625 fits
[[Parallel(n_jobs=-1)]: Done 1993 tasks      | elapsed: 24.4min
[Parallel(n_jobs=-1)]: Done 2600 tasks      | elapsed: 31.8min
[Parallel(n_jobs=-1)]: Done 3289 tasks      | elapsed: 39.7min
[Parallel(n_jobs=-1)]: Done 4058 tasks      | elapsed: 49.3min
[Parallel(n_jobs=-1)]: Done 4909 tasks      | elapsed: 59.6min
[Parallel(n_jobs=-1)]: Done 5625 out of 5625 | elapsed: 68.0min finished
```



When we began fine tuning the random forest regressor and again applying it to our data set containing only German players, we found that it would take a very long time to come to congregation. The top part of the slide shows that it took nearly a half an hour for random searching for hyper parameters, while the bottom section shows that it took more than an hour to run GridSearchCV, an exhaustive search, based on the results from the random search.

Regression Models - Nations

- These are the scores we were able to achieve when we ran the same regression algorithms on the individual nations

	Brazil	Italy	Germany	Spain
MLP Regressor	**.927**	**.944**	**.956**	**.959**
Mean-Absolute-Error	**.195**	**.172**	**.154**	**.160**
Linear Regression	**.696**	**.825**	**.830**	**.881**
Mean-Absolute-Error	**.355**	**.294**	**.297**	**.274**
Random Forest	**.896**	**.925**	**.968**	**.959**
Mean-Absolute-Error	**.205**	**.166**	**.128**	**.141**

This table represents the data that was collected when we ran our original machine learning models on the chosen nations. MLP regressor, and Random Forest regressor were the best performing machine learning models, but we found that Argentina was the only nation that performed in a different manner.

Regression Models - Argentina

Here are the values that we found to be contradictory to our original machine learning scores

	**MLP Regression**	**Linear Regression**	**Random Forest Regression**
Mean-Absolute-Error	**.191**	**.322**	**.167**
R^2 Score	**.752**	**.669**	**.948**

When the same regression models were applied to the data set containing only players from Argentina, we found that we had a much lower score in MLP regression, and in linear regression when compared with the all of the other results. Random forest regression remained fairly constant even on this data set. We believe the lower scores are in part due to the statistics that come from Lionel Messi being an Argentinian, and the elevated values he introduces into the data set, making it more difficult for linear regression, and MLP regressor to perform in the manner they had on the other data sets.

Classification Approach

- The classification approach was used to see if given a player statistics can we classify which position the player is best suited for
- We used MLP Classifier with a few extra hidden layers to extend the depth, and a random forest classifier as random forest regression was the best performing machine learning model based on the statistics
- Both the MLP Classifier and Random Forest Classifier performed similarly giving an accuracy score of under 50% when classifying among all the possible positions
- Another approach was used to see if we can classify players into a general playing position utilizing random forest classifier which had an accuracy of 68%

For classification we used the Position column from our dataset and using sklearn label encoder converted them to a number between 0 to (n-1) representing unique playing positions. Using the encoded position as our target, and the transformed features from PCA as our predictors for our classifying models we achieved an accuracy score of 50% with both MLP and Random Forest Classifier. Another approach we took was to encode the general playing positions and using that as our target we were able to classify a player using random forest into a general position with an accuracy of 68%.

Predicting Player Positions

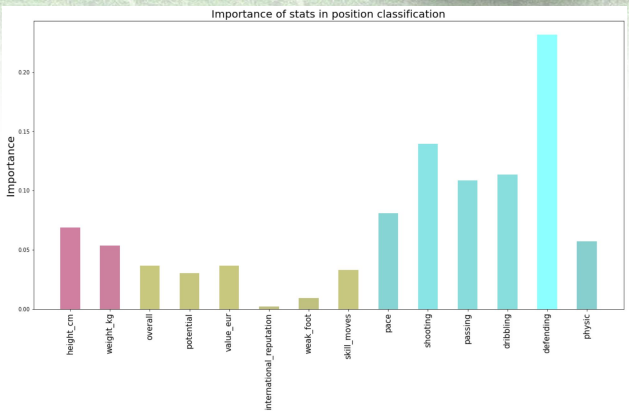
We wanted to see if we could predict the “best” position for a player based off of their listed stats.

Using a Random Forest classifier we evaluated the impact various variables had on the selection of player position and if we could accurately predict position.

0.6821345707656613

	precision	recall	f1-score	support
cb	0.77	0.91	0.84	67
wb	0.66	0.53	0.58	72
mid	0.63	0.73	0.67	131
wing	0.53	0.51	0.52	65
st	0.83	0.70	0.76	96
accuracy			0.68	431
macro avg	0.68	0.67	0.67	431
weighted avg	0.69	0.68	0.68	431

This Random Forest classifier returned an overall accuracy score of 68%, we can see that positions ‘cb’ and ‘st’ give the best results.



Pace, shooting, passing, dribbling, defending, and physical appear to have the highest impact.

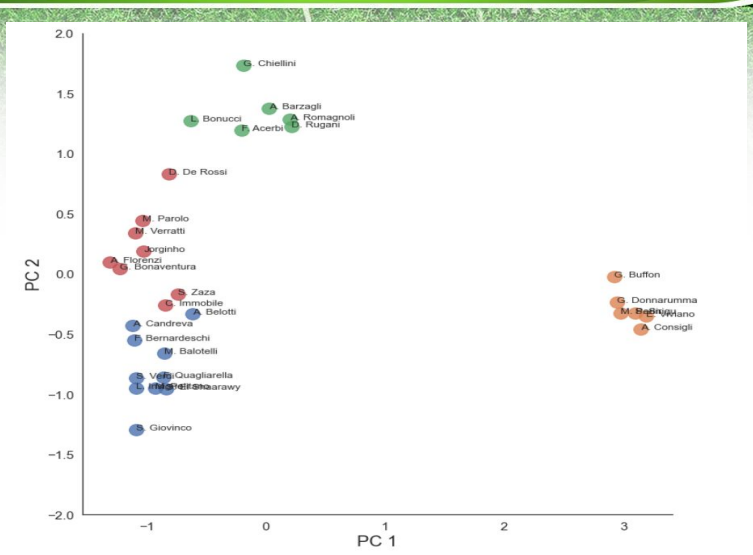
The random forest classifier has an accuracy of 68% for classifying a general playing position for a player. We can see the precision and recall for the various classes. We can also see various statistics and how much they impact the classification of players playing position.

Clustering Models

- We considered two approaches for performing clustering: K-means and Hierarchical with an aim of grouping players with similar characteristics into one group
- Upon testing the models, we found that K-means provides us with a better performance result as compared to Hierarchical clustering
- We selected all the players with an overall rating of above 80 and used principal component analysis to reduce the number of features to 2

For clustering, we aimed at grouping players such that observation within a subgroup were similar to each other and different in different subgroups. We used two approaches- K-means and Hierarchical clustering for forming clusters. K-means uses Euclidean distance to find centroid of points and assigns each player according to the cluster where he belongs. Hierarchical clustering on the other hand plots data in agglomerative manner i.e. the bottom up manner in a dendrogram. It starts with individual observations merges the observation into branches. We specified 4 clusters for grouping of players using K means. We found that k-means gave us better performance results as compared to hierarchical clustering.

Clustering - Italian Soccer Players



Legend

- Goal Keepers
- Forward
- Defenders
- Midfielders

This is a plot of players belonging to Italy soccer team. On running the principle component analysis we noted that about 50% of information from all the features in the dataset was retained in 2 components. However on plotting the information in clusters we saw that information was sufficient to categorize players into goalkeepers, forward, defenders and midfielders. The orange cluster on the right are the goalkeepers, the blue ones at the bottom are forward players, the red ones in between represent the midfielders and green ones at the top are defenders.

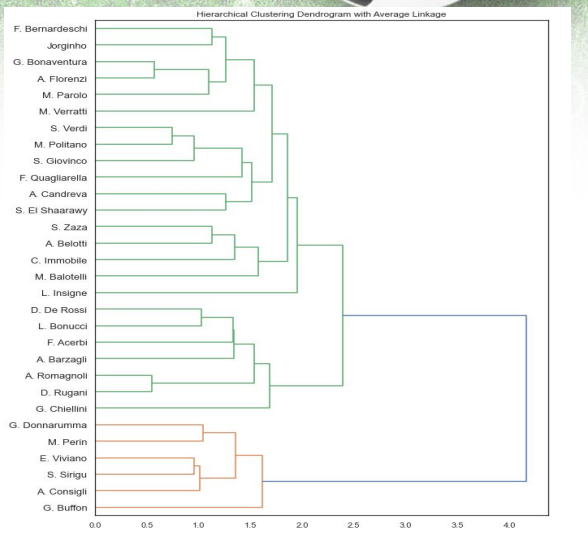
Linkage of Italian Soccer Players

Goal Keepers
L. Viviano
G. Donnarumma

Forward
A. Belotti
S. Giovinco

Defenders
L. Bonucci
D. Rugani

Midfielders
M. Verratti
A. Florenzi



Based on our clusters, we noted that Viviano and Donnarumma were clustered as goal keepers, Belotti and Giovinco clustered as forward player, Bonucci and Rugani as Defenders and, Verratti and Florenzi were clustered as midfielders. On the right side is the dendrogram representing hierarchical clustering of Italian players. The players at the bottom in red are the mostly the goalies and the ones at the top are all other players. So we got similar results like we got using K-means , however, k-means gave us better clustering.

Regression Results

- We were able to regress to the Overall Rating with a very high coefficient of confidence using the various models, and hence for a new player we can find an Overall/Potential Rating.
- While we implemented three different models, we had the most amount of success and the least amount of error using MLP and Random Forest regressors
- Depending upon which data set it was applied to, either MLP or Random Forest could provide the highest coefficient of confidence
- As discussed in the 'Regression' section of the presentation, Linear regression had a difficult time accurately testing the different data sets because there is not a clear linear correlation within the data

We used multiple machine learning models in an attempt to find which one was the most accurate for our data set. We had mixed results, but the model that seemed to return the highest scores on the most consistent basis was Random Forest regression. We would go on to use this model in our attempt to predict players positions. Linear regression demonstrated the worst behavior on every set that we applied it to, and had the highest mean-absolute-error.

Clustering Results

- Using the Clustering approach, for a new player with given statistics we can cluster the player into a general position class
- We were able to cluster the players into simplified position classes within margin of error
- Despite retaining about 50% of the feature characteristics in two components, we were able to achieve at least 80% accuracy in forming clusters

Classification Results

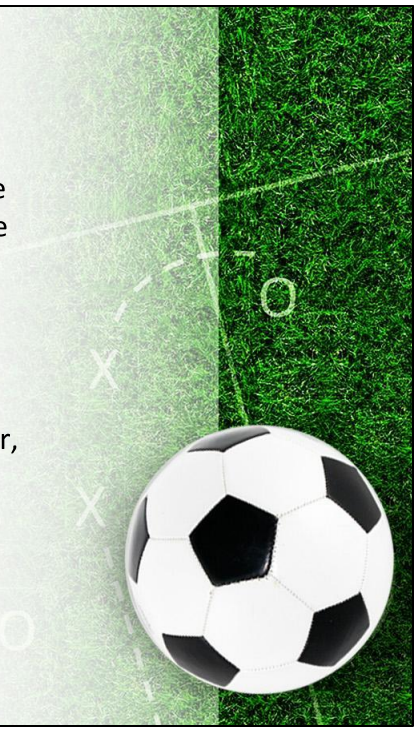
- The classification approach did not give us good results, we can only predict a players playing position with an accuracy of around 50%. While this can still be used as a starting point for further analysis.
- We could have got better results if the clustering approach could have been used with the classification approach as after clustering into a general position we are limited to a smaller number of classifying targets and hence could have increased the accuracy but we were limited by the time constraint

We used MLP Classifier and Random forest classifier for our various classification approaches. In general the random forest classifier performed better than the MLP classifier when applied to our datasets. Of the two approaches, we were looking to classify players into a specific position, or into a general position, and we achieved a better accuracy for general positioning because there were a smaller number of targets to predict. Similarly if clustering was used alongside of classification we believe we could have had an even better accuracy score for predicting a specific playing position.

Future Work

- Using more feature engineering we would most likely be able to find a better set of features to train our machine learning models. Genetic algorithm was considered, but PCA was faster and easier to implement given the time constraints

- Using the Clustering and Classifying approaches together, for a new player given statistics from a game we can cluster the player into a general position class, and then classify them to a playing position




Questions?