# CS 590V FINAL PROJECT

KAILASH NATHAN SRINIVASAN

**Metadata**

## European Soccer Database

- +25,000 matches
- +10,000 players
- 11 European Countries with their lead championship
- Seasons 2008 to 2016
- Players and Teams' attributes* sourced from EA Sports' FIFA video game series, including the weekly updates
- Team line up with squad formation (X, Y coordinates)
- Betting odds from up to 10 providers
- Detailed match events (goal types, possession, corner, cross, fouls, cards etc...) for +10,000 matches
- New table containing teams' attributes from FIFA.
- Consists of 7 Tables.

| | |
|---|---|
| **1** | Player_Attributes |
| **2** | Player |
| **3** | Match |

| | |
|---|---|
| **4** | League |
| **5** | Country |
| **6** | Team |
| **7** | Team_Attributes |

The data was sourced from:

- [http://football-data.mx-api.enetscores.com/][1] : scores, lineup, team formation and events
- [http://www.football-data.co.uk/][2] : betting odds. [Click here to understand the column naming system for betting odds:][3]
- [http://sofifa.com/][4] : players and teams attributes from EA Sports FIFA games. *FIFA series and all FIFA assets property of EA Sports.*

Some issues with the data include some players are missing from the lineup (null values).

**Interest** :

Being an avid follower of football(soccer) for over ten years now ,it was a curiosity to explore what patterns would emerge from this dataset with the right questions. The statistics from FIFA

Games as well provide an interesting aspect as to how the rating system works as I am a huge fan of the game . I was hoping these insights would help me next year in my fantasy football league.

The Questions that I had planned to ask:

- Identify the top 20 players across various positions and try to find out what makes them special.
- Explore the player attributes provided by EA SPORTS to check distributions of overall ratings using standardised stats.
- Predict the probability of total number of goals scored in a game.

The questions I ended up asking :

- Comparing the top 100 players in World Football as of 2013 using the statistics EA Sports uses in its games to find some pattern and build some mental models as to what makes these players standout.
- The age old comparison of who is the better player among Cristiano Ronaldo and Lionel Messi.[2]
- Comparison of the style of play among the top 15 teams in World Football and 15 random teams.
- Why English Clubs struggle in UEFA Champions League while Spanish Clubs seem to have a good fortune of late?

**Analytics:**

- K means Clustering on the top 100 players and their attributes according to FIFA.
- Parallel Coordinate Comparison of style of play between the clubs.

- Z test to determine whether English league Is harder when compared to the Spanish League based on Odds from Betting Sites.
- Calculation of the value of a goal scored based on final result, home or away ,odds and minute when a player scored to compare goal scoring records of Lionel Messi and Cristiano Ronaldo.
- Figuring out the position each player plays based on his other attributes as the data was not available.

**Visualizations**

- Radial chart between the top 2 players
- Line chart
- Heat map
- Histogram
- Pie Chart
- Parallel Coordinate Chart
- K means Clustering Chart

**References:**


1. https://www.kaggle.com/hugomathien/soccer
2. http://www.economist.com/blogs/gametheory/2015/03/statistical-analysis-football
   (Which intrigued me to ask the second question)

3. https://fivethirtyeight.com/features/lionel-messi-is-impossible/
4. https://www.reddit.com/r/FifaCareers/comments/2kei1c/attributes_explanation_guide/
5. https://www.reddit.com/r/FIFA/comments/3u85gs/a_indepth_guide_to_custom_tactics/

Story and Process:

Question 1:

Comparing the top 100 players in World Football as of 2013 using the statistics EA Sports uses in its games to find some pattern and build some mental models as to what makes these players standout.

- After merging two tables to get the Player Name attribute from Player table using the unique id with the Player Attributes column ,I made a list of top 100 players across various position after cleaning the data and removing the null values.

The various attributes that are involved in the dataset are :

**Physical**

Acceleration – how quickly a player can reach his top speed. Some players have high acceleration but only average sprint speed so, in this case acceleration become almost irrelevant – it is no good getting to your top speed really quickly if that top speed is not very quick.

Agility – the ability to change direction at pace. Players with high agility can perform acrobatic shots or clearances.

Balance – the likelihood of a player staying on his feet during/after a physical challenge. Also affects how well a player can perform a difficult shot whilst facing away from goal.

Agility and balance influence a player's dribbling skill and, more generally, how responsive the player you are controlling feels. If a player has high stats for agility and balance then they will move fluidly and be able to get in or out of tight spots. Also, high agility and balance are needed to get the most out of fast players; your fast player could feel sluggish and a little unresponsive if they lack agility and balance.

Jumping – how high a player can jump/how well he reaches the ball through the air. However, for a player to be really good in the air they need to have high jumping, strength, aggression and heading accuracy attributes. Remember to compare the jumping attribute of the player with his height. A 6'5" player does not need need to have a high jumping attribute to connect with the ball.

Reactions – how well and the amount of time it takes a player to react to certain events (e.g. spotting opposition runs and attacking loose balls after tackles/deflections). An attribute that is vital for effective dribbling.

Sprint speed – how fast the player runs while at top speed. It needs to be combined with acceleration, agility, balance and reactions in order to be fully effective.

Stamina – affects the amount of time a player can sprint before slowing down. Also dictates the loss of energy during a match and how fast it recovers after a match.

Strength – the likelihood of a player succeeding in a physical challenge/jostle.

**Technical**

Ball control – how well a player initially controls the ball, then how good they are at keeping it under control. Desirable for dribblers and also for strikers –

you could have the best striker in the world up front but if he cannot trap a bag of cement then he is not going to get goals.

Crossing – the likeliness of a player finding his teammate in the box when performing a long ball from the wing. Also dictates the power in the cross (as in the pace of the ball, not distance) and the amount of curve that he can apply at such pace (coupled with the curve attribute). Also determines how easy it is for a player to get the ball in the box in the first place. If your winger keeps getting blocked when you are expecting him to whip the ball in , there is a good chance he has a low crossing attribute.

Curve – how much bend a player can get on the ball. Matters for shots, crossing, short passing and through balls if an opposing player is obstructing the straight ball to a teammate.

Dribbling – how well the player controls the ball while moving. A player with a high dribbling stat will keep it close to his feet and knock it through tight gaps to beat his opponents.

Finishing – the accuracy of shots (using foot), inside the penalty area.

Free kick accuracy – the frequency of which a player will hit the target when taking a free kick. However, some of this is up to you.

Heading accuracy – affects your player's ability to get their head on the ball, then how accurate that header is going to be. It applies to headed passes as well as headers at goal.

Long passing – the ability to pick out a teammate that is far away. Also determines how quickly the ball gets to them. This applies to x/□ no matter the length of time you hold it, switching the play and any kind of through ball.

Long shots – the accuracy of shots from outside the penalty area.

Marking - your player's ability to stay close to an opposing attacker and stop him getting to a cross/pass from a teammate. Also contributes to tracking runs.

AI players will do this off the ball, and when you are in control of a defender their marking stat will affect how good they are at containing.

Penalties – a player with a high penalty attribute will have an easier time scoring a penalty as the 'green zone' will be bigger and the marker will move slower.

Short passing – the player's accuracy and speed of passing over a short distance (does not include through balls).

Shot power – the amount of power a player can put into a shot while still keeping it accurate. This is not power alone. A player with a low shot power can still hit the ball hard into the back of the net but the longer you hold B/O the more likely he is to miss.

Sliding tackle – how well a player can win the ball and keep possession of the ball while making a sliding challenge without conceding a foul. If the stat is low, he is more likely to miss time the challenge or kick the ball away.

Standing tackle – how well a player can win the ball and keep possession of the ball while making a standing challenge without conceding a foul. If the stat is low, he is more likely to miss time the challenge or kick the ball away.

Volleys – technique and accuracy of shots taken when the ball is in the air (this tends to be coupled with the balance trait if he is not fully facing the goal).

**Goalkeeper**

GK Diving – in combination with height determines the goalkeeper's ability to reach and make a save whilst diving through the air.

Gk Handling – the frequency that the keeper catches the ball rather than parrying it and whether or not he holds onto it.

GK Kicking – the length and accuracy of goal kicks, from out of the hands or on the ground. The length and accuracy of throws are partially determined by the kicking attribute but mainly determined by the long throw trait.

GK Positioning – the goalkeeper's ability to position himself correctly when saving shots. Implies that you should rely solely on the AI to position your goalkeeper; henceforth even goalkeepers with high positioning can sometimes make bad decisions. Do not focus on this attribute too much, if the other attributes are high enough you can have a good goalkeeper.

GK Reflexes – how quickly the goalkeeper reacts to a shot on goal. If the stat is low he will make a move later or could miss it entirely.

**Mental**

Aggression - This works in tandem with strength; you'll usually find that a player has high scores for both, rather than one or the other. The frequency of jostling, stand tackling and slide tackling. Aggressive players generally win more tackles, but also risk giving away more free-kicks.

Attacking positioning – a player's ability to spot open space and move into good positions that offer an attacking advantage. The higher this attribute, the more likely a player is to make enough space to receive the ball in dangerous areas.

Interceptions - This applies more to AI controlled team-mates rather than the player you have selected, as it determines the ability to read the game and intercept passes. If you have seen players stick out legs or do something unexpected to prevent a pass, it is likely that they have a high score for Interceptions.

Vision – the player's awareness of the position of his teammates and opponents around him. You may be able to see an AI player's run but if the player you are controlling has low vision, he will not be able to pick out that pass and will most likely opt for something that he can see or just hit it however hard he wants in the direction you happen to be holding your analogue. Work in tandem with the long passing attribute.

- I had to figure out what position each player plays based on his attributes by setting attribute ranges for each position .For example , an attacker has good finishing skills and good sprint speed while a defender is good at intercepting . I set up some basic conditions to figure out the position of each player.

I visualised the data set using Bar chart ,Pie Chart ,Heat Map ,Line Chart and Row chart hoping to find some models .These are the some of the observations.

- Even Distributuon with regard to average Overall over Position.There seemes to be less defenders in the top 100 players when you compare the count.There are more attacking players with higher average overall which was expected more Right footed players in the top 100.

- Sprint Speed and Ball Control are attributes more in common among Midfielders and Attackers which explains the cluster on the lower region .On further probing ,it is observed that they are from the GK and DEF which makes sense.

- Vision is not a common trait and its found only in the top Playmakers in World Football which is also evident.

- Vision is not a common trait and its found only in the top Playmakers in World Football which is evident from the Histogram. Most of the top players have good ball control as seen from the left skewed histogram. Same goes with the short passing attribute.

- It is a straightforward comparison between the overall attribute of the player and various skills. There seems to be a linear relation with respect to the position the player plays.
- Most players are on the incline as there is a linear relation between current overall and potential. There seems to be a very few players in decline.

- There seems to be some outliers in the data when comparing GK rating and the overall player attribute

Question 2: The age old comparison between Cristiano Ronaldo and Lionel Messi.

The attributes from FIFA and the goal scoring records of both of them for the calendar years of 2013 and 2014 are used to make a radical chart comparison between them.
The goal scoring record also had info regarding the time they scored ,the odds of victory before and after they scored. I used a weighted average approach to calculate the value of each goal they scored .This is significant because
For example, take the two situations mentioned above. With the score tied in the 90th minute, a team playing at home has an 11% chance to win, 82% to draw and 8% to lose. Multiplying by three points in the standings for a win, one for a draw and zero for a loss, its expected points (EP) are more. With a one-goal lead, in contrast, its win probability shoots up to 95% and its draw odds fall to 5%, which equates to 2.89 EP. The gap between them of 1.76 points is the EPA associated with such a critical goal. In the alternate situation, with a two-goal

lead in the 90th minute, the home team already has a 99.7% chance to win and 0.3% to draw. Adding on a final goal as an exclamation point is worth less.

I took into account the style of play of both their teams as well.

Observations:

It was evident from the chart that both of them have very similar attributes.The change of position of Ronaldo in the recent years from a Winger to a Center Forward has given him a edge in the attributes such as Finishing ,Positioning,Sprint Speed and Shot Power.Messi position as a Playmaker suits his short passing and ball retention skills .He has better dribbling ,passing,chance creation and skills and acceleration and balance which makes him the best playmaker in World Football at the moment. The style of football played by the teams is significant.Real Madrid play more of a Counter attacking style of football at a rapid pace which suits Ronaldo. Barcelona are known for their "Tiki Taka" style of football of retaining position and passing the ball short.Ronaldo is a better Center Forward while Messi is the best playmaker in World Football.

**Question:**

Comparison of the style of play among the top 15 teams in World Football and 15 random teams

Explanation of the team style of play parameters.

**Build up play - speed**

First thing to clear up: on slow build up, your team mates will NOT make LESS runs, they will simply hold off the run a bit longer. That way you'll have time to control the ball before looking for a new pass. Slow build up is better for players that are less experienced, when you're adapting to a new formation or when your players are of lower quality.

Fast build up: exact opposite of slow build up (surprise surprise). Your players will position themselves directly after you get the ball and make runs in a fast succession. Good for when you have good players and know the formation.

*My advice*: start slow and gradually work your way up. The faster the better, but **make sure that you and your players can follow.**

**Build up play - passing**

Pretty straightforward. Your players will come shorter and start their runs closer to you on short passing and stay further away and start their runs further away on long passing.

This also has an effect on your defensive positioning. If your players stay further away from each other when you have the ball, it will create space for the opposition to dive in to on counters.

*My advice:* with the improved interceptions on FIFA16 it's better to play on short passing. **You can manually trigger runs to send your players further up the field but you can't call them to come short.**

**Build up play - positioning**

Positioning in build up play ONLY reflects on what's happening on your half of the pitch, but it can have an effect on the way you attack.

Organised is the safer option. Your central defenders keep their position better, your fullbacks don't make forward runs as much and stop their run earlier, your midfielders tend to wander off less and are defensively better positioned, ...

Free form is the more creative option. Your fullbacks are more involved while attacking, your central midfielders are more dynamic and they make more runs. On the flip side your fullbacks and midfielders can get caught out of position more easy.

**Chance creation - passing**

You can play safe (<30), risky (>70) or something in between. On safe, your player tends to position himself to get the ball with a low chance of interception. On risky, he'll dive into a pocket behind the opponent much quicker if he sees a hole to play through.

**Chance creation - shooting**

With shooting on high, your players will position themselves on the edge of the box to shoot. Shooting on low means they'll run further into the box and tend to walk the ball in the net (think Barcelona). Shooting on high tends to create more long shots, shooting on low creates more close one-on-ones with the keepers.

*My advice:* With keepers having line reflexes of a hardcore cocaine addict, shooting from further out seems like a more viable option. If you like beautiful tiki-taka Barçagoals, go for low shooting.

**Chance creation - positioning**

This positioning reflects on what happens in the OPPONENTS half. It will have the same effect (more/less creativity, more/less position changes, more/less drifting around, ...) as the other positioning, but on your strikers and wingers.

The team width slider determines how your team positions itself on the pitch when **defending**. It has nothing to do with how your team plays in attack.

I took the top 15 teams and use Parallel Coordinates to measure their style of play and find what was common.
Then I choose 15 random teams and used the same technique to figure out what they were lacking.
The correlations cannot be inferred as there are other factors involved here as well.

**Question :**Why English Clubs struggle in UEFA Champions League while Spanish Clubs seem to have a good fortune of late?

There has been a decline in English dominance in the club level of European Competition and I wanted to know whether the reason was whether the league has become so much tougher and hence handicapping them .
I used hypothesis testing to find the difference in the level of competition in the top divison of English and Spanish leagues for the top teams.
I took the odds provided by Betting 365 for their home games as home teams are mostly favourites to win.
I took the propotion of wins based on these odds and used the Two Sample Propotion test to find the Z and P values. It showed a significant difference suggesting that English teams find it harder in the league.

Links:

DB:

https://www.kaggle.com/hugomathien/soccer

QUESTION 1:

http://edlab-www.cs.umass.edu/~kailashnatha/template.html

K Means Clustering :

http://edlab-www.cs.umass.edu/~kailashnatha/kmeans.html

QUESTION 2:

http://edlab-www.cs.umass.edu/~kailashnatha/mesron.html

QUESTION 3:

http://edlab-www.cs.umass.edu/~kailashnatha/parallel.html

http://edlab-www.cs.umass.edu/~kailashnatha/random.html