Application of Latent Dirichlet Allocation on FIFA 19 dataset to emulate facilitating scouting process

Chuk Uzoegwu
Princeton University
cuzoegwu@princeton.edu

Abstract

Could there possibly be a way to partly automate the effort of finding talent through the implementation of AI? In this project I use a complete Fifa 19 dataset (assuming that the reporting is accurate) which contains information about 18207 players worldwide including their nationalities, body types, preferred foot, and ratings for different soccer skills defined. I implement the Latent Dirichlet Allocation modeling method to analyze the latent structure between the different soccer skills and see if there is perhaps some relationship between key skill characteristics and other attributes such as preferred foot, the nationality of players, body types, value, and overall ability. I analyze the latent structure relationship to the average heights and weights of the players as well as their overall ability then perform an in-depth analysis on the relationship with overall and potential which includes fitting a Linear Regression Model to determine how well certain top attributes determine a players overall and potential. The goal of the project is to think like a scout would if he/she had this dataset.

1 Introduction

In the sports world, one of the primary indicators of a team that is or will be successful that the team possesses a good set of scouts. The main assignment of these scouts is to identify talent from wherever they may be hidden to help benefit the team. In the world of soccer, this task can be even more so arduous considering the fact that it is a worldwide sport. The game is played on a global stage, and great talent can be acquired from all over the globe. Consequently, scouts and coaches alike often traverse through hundreds to thousands of scouting reports in hopes of being able to identify a singular potentially impactful player. This is where the use of computer intelligence can be very helpful. This project aims to do a similar work on a smaller scale by looking at the "scouting reports" of players produced by the Fifa 19 gamemakers in order to go through a similar process of deciding certain attributes about the players. In this project we focus on analyzing what the latent structures of the skills tell us about height and weight of the players and their overall and potential. We then will perform a linear regression analysis to determine how well a certain set of attributes predict the potentials and overalls of the players. The goal of this is interpret the data in the context of how it would be of use to a scout with this dataset assuming that these were accurate measurements of ability.

2 Related Work

Serving as examples of the proliferation of the use of artificial intelligence in the sports world, two data-driven companies, JUST ADD AI and a Dutch Sports Company called SciSports have each made their own impact in the sports world via AI. They have used machine learning to characterize over 90,000 players internationally and have experienced success such as helping an unsigned Dutch free agent to find the right team which consequently has led to him being one of the top goal scorers in the Dutch professional futbol league. SciSports notably uses a 3-D camera to analyze the real time game that is happening called BallJames. This technology allows them to get a full picture of the game by keeping track of aspects of the game and the players that the stat sheets do not often accurately report if at all[1]. Similarly JUST ADD AI has created a new system known as the JAAI (Just add AI) Scout which utilizes different aspects of the IBM Watson in order to produce impactful player analytics as well as do other practical activities such as read player scouting reports. They too experienced some live success but this time it involved a player in the German Bundesliga Futbol League.[2]

3 Methods

3.1 Preprocessing

I first decided to split the dataset into training data and testing data using the train_test_split function of ScikitLearn.

I then took the training data and gathered the columns that I wanted to fit my LDA model with. The data that I included were the weak foot (the ability to use your less-dominant foot), skill moves (the ability to do soccer tricks in game), work rate, and all of the individual skill ratings from crossing to goal-keeping reflexes. When the columns are joined the dataframe that we are left with is one of dimensionality: 14565 rows x 37 columns that contains the ratings of the desired attributes. The matrix however contained values that did not match. The 'skill moves' and 'weak foot' columns contained categorical values from 1 to 5 where 1 is very poor and 5 is excellent. The 'work rate' column contained string categories of the form '[High, Medium, Low]/ [High, Medium, Low]' that represents how hard the player works on the offensive/defensive end. I considered the option of reforming the data so that all of them contained categorical values similar to that of 'skill moves' and 'weak foot' but I decided instead to convert the categorical data to be more so continuous since the continuous data would provide more accurate analysis. For the skill moves and weak foot I individually replaced all the occurrences of 5 with a random number between [90, 99] (excellent), occurrences of 4 with a random number between [80, 90) (very good), occurrences of 3 with a random number between [70, 80) (good), occurrences of 2 with a random number between [60, 70) (poor), occurrences of 1 with random number between [10, 60) (very poor). Similarly the replacement of the work rate values went as follows: 'High/High': rand [90, 99], 'High/Medium' or 'Medium/High': rand [80, 90], 'High/Low' or 'Low/High' or 'Medium/Medium': rand [75, 85], 'Medium/Low' or 'Low/Medium': rand [60, 75], 'Low/Low': rand [10, 60].

3.2 Feature Selection

Not much feature selection was done with exception to identifying players that did not have skill ratings. Within the training set there were 40 players who did not have any skill ratings assigned to them so consequently I decided to drop them from the training set. As a result I was left with a matrix of dimensionality 14525 rows x 37 columns. The other option that I considered was filling in the null values with a randomly generated integer however that would greatly skew the results considering the factors that are considered when assigning ratings.

3.3 One-Hot Encoding vs. Continuous

I considered one-hot encoding but chose continuous because they provide a more unique latent structure.

3.4 One method in detail

Spotlight Model: Latent Dirichlet Allocation

Latent Dirichlet Allocation is a topic model that can discover abstract topics from a collection of documents. It is also a generative probabilistic model that aims to map documents to a set of imaginary topics that can capture all the words that appear in the documents. It defines a joint probability distribution over both the observed and latent variables, the observed variables being the words in the document, and the latent variables being the topic structure:

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) = \prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n})$$

where $\beta_{1:K}$ are the topics, $\theta_{d,k}$ is the topic proportion for topic *k* in document *d*, and $z_{d,n}$ is the topic assignment for the *n*th word in document *d*. Latent quantities are non-negative and sum to one.

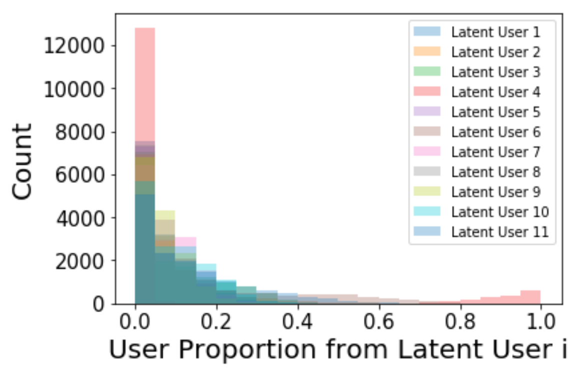
The idea behind LDA is that documents can be explained by a distribution over topics, while topics can be explained by a distribution over words. Since a document has words from many topics, it is referred to as an admixture, or mixed membership, model. All documents in the collection share the same set of topics, but each document exhibits topics in different proportion. Each word is assigned to a topic, which is in turn drawn from a document-specific distribution. The ”words” in this case are the ratings for the skills of each of the players while the ”documents” are the actual players and the ”topics” are the assignment of the features.

I implement LDA by importing the function from *sklearn* and fitting the model with the attribute statistics assigned by Fifa to the players in the training set. I fit 2 models, one where the number of components is equal to 4 to account for the 4 general aspects of the field (Goalkeeper, Defense, Midfield, Attackers) (it is not shown as the results were not too interesting) and 11 to get a more mixed look at the relationship of the statistics considering that there are 11 players on the field.

Geometrically, LDA can be interpreted as a dimensionality reduction. For a set of V words and K topics, LDA projects a point in the V-dimensional simplex onto the K-dimensional simplex. Topic proportions are a K-dimensional Dirichlet distribution, and topics are a V-dimensional Dirichlet distribution.

4 Results

4.1 11 Latent Users

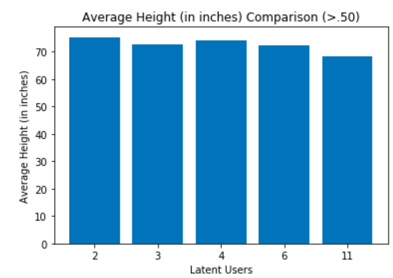


In the histogram for n_components = 11 we see a more skewed left graph which makes sense considering with more topics (latent users) there is a greater distribution of words (characteristics) among the latent users which results in lower proportions. One latent user that is particularly notable in this distribution is latent user 4 who is in fact ”goalkeeper” latent user which further emphasizes how strictly different the goalkeeper role is from the others. The score for the LDA with the training data is -6628.69695522525 while for the testing data it is -6631.598170223036. The minimal difference suggests a good fit for the LDA model.

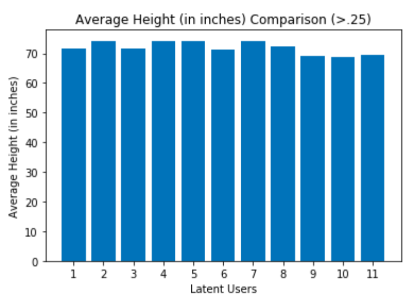
Additionally looking into each of the latent users individually and assigning them a role based off the top skills (goalkeeper, defender, midfielder, attacker), it is apparent that we have 1 goalkeeper latent user [4], 4 defender latent users [2, 3, 5, 7], 4 midfield latent users [8, 9, 10, 11], and 2 attacking latent users [1, 6] *(See Appendix)*. That distribution is actually representative of the distribution of positions for most teams' starting 11. What this suggests is that if a scout were to receive a report of just the skills of the sets of players in a league/camp/etc. he could make an adequate estimate of the distribution of attackers, midfielders, defenders, and goalkeepers to decide if it would be worth further exploring that league.

That being said let's turn our attention to discovering relationships between the skillsets and other player attributes. We will take a look at 4 of them: overall, potential, height, weight, then take a deeper look into 2 of them, overall and potential. For each category we take the average for each latent user with a proportion threshold of .25 and .50. .25 is a low threshold but given the skewness of the graph, it was necessary in order to get some value form each of the latent users and see if characteristics makes a key difference. Increasing the threshold to .50 gives a truer sense of how the characteristics of the latent users impact the given category.

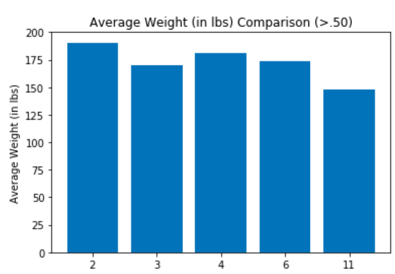
4.2 Height/Weight



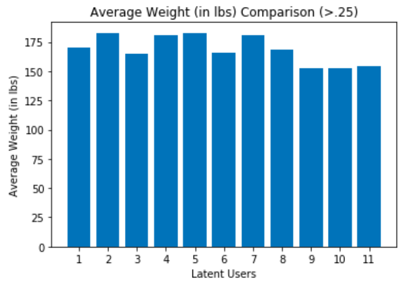
	2	3	4	6	11
0	75.310811	72.666667	74.094001	72.46823	68.278049



	1	2	3	4	5	6	7	8	9	10	11
0	71.779487	74.241143	71.540803	74.094001	74.178571	71.111079	74.07619	72.486034	69.291139	68.732759	69.361261



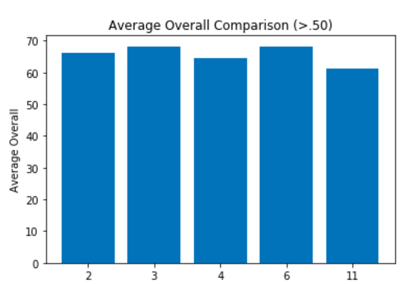
	2	3	4	6	11
0	190.459459	170.0	180.541744	173.213313	147.707317



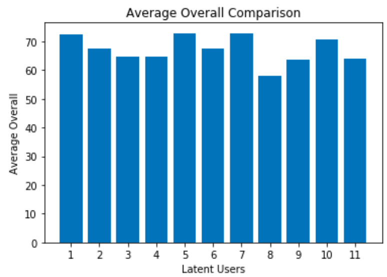
	1	2	3	4	5	6	7	8	9	10	11
0	170.625641	182.733714	165.216969	180.541744	182.892857	165.496135	180.819048	168.757914	152.253165	152.937931	154.535357

A key thing to notice is, given the positions we assigned we each latent user it is suggested that those with the skillset of midfielders and attackers tend to be smaller than those with skillsets of a defender and those with skillsets of a goalkeeper. Even more so we can see and make a good guess for which characteristics are most correlative (suggesting a causal relationship) with the height and weigt. For example, looking at latent user 2, a defending latent user, we see that it's top characteristic is strength and it is the top latent user in terms of size for both thresholds which suggests a strong relationship between body height and weight. To further emphasize that point we can take a look at latent user 8, whom we consider a midfielder but has a size that more so matches that of the defender. Although a majority of it's top 10 skills are midfielder characteristics, one of it's top 5 is strength. Perhaps user 8 would be what is considered a center defensive mid. We may already think this ”bigger body = more strength” relationship or the ”smaller body = more balance” relationship that seems to be reflected my the midfield latent users, to naturally be the case but it is interesting to see this model reflect it. Another less obvious note that may debunk a belief a young scout may have is the lack of relationship between shot power and body height and weight. If a scout were to be on the lookout for a player with a powerful shot or long range shooting capabilities, an inexperienced one might go searching for bigger players. However comparing our ”attacker” latent users [1, 2], we see that although ”shot power” and ”long shots” is more of a skill of latent user 2, it is still the case that latent user 1 is bigger. Now granted latent user 1 also has a top skill of strength which may have something to do with it, but there is reason to perhaps suspend that belief and look over a broader range of players. Similar to finding the distribution of player positions based on the sets of skill alone, it seems like a scout could also find the distribution of the height and weights of players based on the skillsets as well. There are many other interesting correlations one could find here but let's move on.

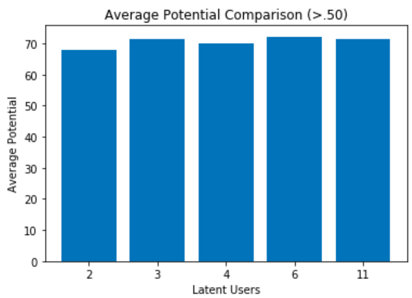
4.3 Overall/Potential



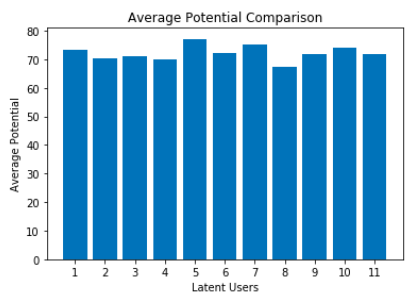
	2	3	4	6	11
0	66.27027	68.0	64.651206	68.249622	61.35122



	1	2	3	4	5	6	7	8	9	10	11
0	72.497436	67.410286	64.569948	64.651206	72.9375	67.474664	72.866667	58.184358	63.56962	70.698276	64.004612



	2	3	4	6	11
0	67.77027	71.5	69.881262	72.189107	71.341463



	1	2	3	4	5	6	7	8	9	10	11
0	73.389744	70.170286	70.994819	69.881262	77.133929	72.097338	74.980952	67.417132	71.708861	74.015517	71.694466

Comparing the latent users in the different positions could give us some insight into finding the key characteristics that a scout may want in each of those positions. Looking at the top 5 characteristics of latent users 1 and 6 a scout could perhaps suggest that it would be better to look for a forward who can do other things other than shoot the ball ("long passing" and "FKAccuracy") or that it is better to find a forward who has "composure" (the ability to not get rattled) than a player who can use his weaker foot. Looking at the midfielders [8, 9, 10, 11] a scout may perhaps see that it is best to look for a midfielder who is mobile ("agility") and has the ability to stay on his feet under pressure ("balance") first and foremost. Amongst the defenders [2, 3, 5, 7] a scout would suggest, perhaps least surprisingly, that the primary traits to look for in a good defender is 'Standing Tackle' (ability to tackle while on your feet), 'Sliding Tackle' (ability to tackle while sliding), and 'Interceptions' (ability to read a pass being made and get to it). Finally for a goalkeeper, the scout would recommend finding a goalkeeper that is quick to react to make a save ('GKReflexes') and can dive far ('GKDiving') but first and foremost that works hard. As an aside from a fan of the sport and occasional player, all of those recommendations would be considered to be very good.

An interesting fact to also point out here is the differences in the potential and the overalls of all the midfield latent users that seems to suggest that perhaps the ability to use one's weak foot is not indicative of how good a player is or can be (overall/potential). If we look at latent user 8 we see a player who has the best ability to use his weak foot yet has the lowest overall and potential. On the other hand we see in latent user 10 that 'weak foot' is a top 5 skill and it has the highest potential and average over all of the midfielder latent users. This can be interpreted to mean that although having the ability to use his weak foot will certainly help a player to be good, it is also possible that he could be good without it. Same goes with how hard the player works. A stronger correlation however, seems to be a midfielders ability to cross and their ability to use their weak foot. Looking at users 10 and 11, who both have 'Crossing' as a top 10 skill and are the only latent users who do, they also have 'weak foot' as a top 10 skill. It should also be mentioned that users 10 and 11 have skills that are linked to the "winger" position or "outside (left/right) midfielder" position so crossing is essential but also the ability for the player to cut inside and make a cross, make a pass, or take a shot with what may be their weaker foot.

4.4 Overall/Potential In-Depth

I took the top 5 latent user players from potential and overall with a .25 threshold and top 3 latent user players from potential and overall with .5 threshold. The top 5 latent user players for potential and overall with a .25 threshold were the same [5, 7, 10, 1, 6] while the top 3 potential for a .50 threshold were [6, 3, 11] and top 3 overall were [6, 3, 2]. I then found the top 10 skills from those groups of players, based on the pseudocounts, and used them to fit a linear regression model to determine how good of a predictor they were for overall and potential respectively:

4.4.1 Potential In-Depth

	Feature	Count
11	Weak Foot	366756.382030
16	LongShots	358619.389187
4	ShotPower	340680.385127
21	Volleys	281041.670969
19	BallControl	271517.655841
10	Skill Moves	263102.105237
8	Finishing	262249.436952
23	HeadingAccuracy	254399.090582
25	Positioning	236615.003288
17	Stamina	217714.935958

Top 10 (.25)

	Feature	Count
6	Weak Foot	426800.777563
14	Skill Moves	426537.739917
18	Positioning	304942.633131
5	Work Rate	295038.396370
8	SprintSpeed	276522.827667
0	ShotPower	199380.822082
2	Balance	190851.020544
3	Finishing	182783.755848
9	LongShots	181031.172941
11	Acceleration	179469.049468

Top 10 (.50)

To the left is the top 10 with threshold .25 and the right is top 10 with threshold .50. I then for each took the data of their respective top 10 skills from my original training set and took the data for potential from my original training set and fit them to a Linear Regression model with all parameters as default (when I altered the parameters I received worst scores). The R^2 score for each of them was as follows:

	.25 Potential R^2	.50 Potential R^2
0	0.207897	0.112101

This shows that the set of skills on the left are more predictive of the players potential than the set of skills on the right and perhaps one would argue in the order given. Interestingly enough 'weak foot' is at the top for both sets which perhaps counters the intuition stated previously that 'weak foot' does not affect a player's ability. This comes as a relief as it would make sense that a player's ability to use their weak foot is a good sign that they could become a good player. Given the types of skills are more similar to what you would see with a attacker, it could also be said that generally more offensively geared players are the best soccer players.

4.4.2 Overall In-Depth

	Feature	Count
11	Weak Foot	366756.382030
16	LongShots	358619.389187
4	ShotPower	340680.385127
21	Volleys	281041.670969
19	BallControl	271517.655841
10	Skill Moves	263102.105237
8	Finishing	262249.436952
23	HeadingAccuracy	254399.090582
25	Positioning	236615.003288
17	Stamina	217714.935958

Top 10 (.25)

	Feature	Count
6	Weak Foot	362992.260820
16	Skill Moves	356411.152417
14	HeadingAccuracy	261665.907167
4	StandingTackle	231045.448985
7	Interceptions	228348.060420
5	Work Rate	219284.921100
11	Aggression	218022.428613
9	Marking	204473.654214
0	ShotPower	199380.822082
3	Finishing	182783.755848

Top 10 (.50)

To the left is the top 10 with threshold .25 (the same as the top 10 for potential) and the right is top 10 with threshold .50. I then for each took the data of their respective top 10 skills from my original training set and took the data for potential from my original training set and fit them to a Linear Regression model with all parameters as default (when I altered the parameters I received worst scores). The R^2 score for each of them was as follows:

	.25 Overall R^2	.50 Overall R^2
0	0.469983	0.394656

Once again this shows that the sets of skills to the left are more predictive of a players overall ability. The appearance of 'weak foot' once again further solidifies that the previous intuition mentioned regarding the affect 'weak foot' has is false which once again does make sense. It could still be the case that a player can be great or have great potential without being able to use the weak foot, however this suggests that more likely than not, lack of being able to use one's weak foot is detrimental which does make sense.

5 Discussion and Conclusion

In this project I've demonstrated the ability for machine learning and AI to assist a scout in making decisions. We showed that a scout could reasonably determine the distribution of player positions as well as body heights and weights based off solely the data of skills the scout receives. We also analyzed based off of the averages of overalls and potentials, certain skills that may be critical to a player's overall or potential *within* their given positions. Finally, we determined the top 10 skills that impact a players potential and overall and found that the top 10 skills that did the best for both were the same skills based off fitting a Linear Regression model. Despite a previous intuition of the lack of importance of 'weak foot' ability we found out at the end that it is very impactful.
It should be noted that the fact the linear regression scores were not great does make sense since soccer is a sport where you tend to excel at a certain set of skills within your position so trying to determine the ability of any player over only a small subset of skills would not be fruitful. Given more time, what could be done instead is look at the top 10 skills over the given latent user 'positions' that we assigned. Then we could extract the data of all the players and put them within their respective positions then linearly regress them. That would likely lead to a better R^2 score for both overall and potential.
Additionally it makes sense that the linear regression model for overall did better than that of potential because potential is more of a trick thing to calculate and often requires a "special eye". That being said, as we showed, there are certain characteristics that a scout should watch out for that

could be found through machine learning.
Lastly, it should be noted that within both the top 10's for potential and overall, most of the skills are attacking based. This perhaps is a bias of FIFA gamemakers and goes back to what I was talking about regarding soccer being a very positional sport. Granted though the more offensive players are generally seen as the best/most important players and often get paid the highest but from an objective standpoint, one might argue that that shouldn't be the case. A lot more can be done from here as I mentioned throughout the report such as fitting more LDA models with different ranges of components to find a better fit, going more in-depth into finding the top characteristics for each position as well as finding the top areas of the worlds to search for when looking for certain talent but alas this is all for now.

References

[1] Finding the next Football Star with Artificial Intelligence. SAS, www.sas.com/en_us/customers/scisports.html.

[2] Becker, Roland. How to Find New Football Stars with AI Technology. Client Success Field Notes, 3 Jan. 2019, www.ibm.com/blogs/client-voices/how-find-new-football-stars-ai/

6 Appendix

	Feature	Latent User 1 Pseudocount
25	Composure	84797.449788
1	Finishing	79465.681105
8	LongPassing	78269.977045
19	LongShots	77372.497982
7	FKAccuracy	76033.741532
17	Stamina	68471.293360
15	ShotPower	64918.787856
18	Strength	64414.696484
22	Positioning	63343.076947
23	Vision	63282.502245

	Feature	Latent User 2 Pseudocount
18	Strength	134627.988910
36	Work Rate	113825.096530
21	Interceptions	108869.775411
27	StandingTackle	106959.820995
20	Aggression	104037.428329
2	HeadingAccuracy	91647.459828
35	Skill Moves	89207.134395
26	Marking	87304.162190
34	Weak Foot	86017.031047
17	Stamina	81098.010898

	Feature	Latent User 3 Pseudocount
28	SlidingTackle	124320.035461
27	StandingTackle	124085.627990
21	Interceptions	119478.285008
26	Marking	117169.492024
20	Aggression	113985.000285
16	Jumping	108849.950417
36	Work Rate	105459.824569
35	Skill Moves	97183.671450
34	Weak Foot	95695.124462
11	SprintSpeed	93559.805395

	Feature	Latent User 4 Pseudocount
36	Work Rate	131755.631614
33	GKReflexes	112963.380116
29	GKDiving	111597.194190
34	Weak Foot	108967.374521
32	GKPositioning	107486.693713
30	GKHandling	107239.915233
31	GKKicking	104732.677928
18	Strength	96637.109580
13	Reactions	95968.215487
16	Jumping	90415.054511

	Feature	Latent User 5 Pseudocount
26	Marking	96979.604307
20	Aggression	81301.045670
13	Reactions	75701.215468
36	Work Rate	71736.288584
17	Stamina	71287.567304
11	SprintSpeed	71278.384624
25	Composure	68322.147285
18	Strength	66780.843037
10	Acceleration	60625.884888
27	StandingTackle	59715.983284

	Feature	Latent User 6 Pseudocount
15	ShotPower	199380.822082
1	Finishing	182783.755848
34	Weak Foot	181280.105311
19	LongShots	181031.172941
4	Volleys	178195.477811
24	Penalties	177192.682439
22	Positioning	173271.926341
35	Skill Moves	170020.346572
2	HeadingAccuracy	170018.447339
9	BallControl	167587.500589

	Feature	Latent User 7 Pseudocount
28	SlidingTackle	113046.608309
27	StandingTackle	104950.971215
35	Skill Moves	93081.758665
21	Interceptions	91483.227892
3	ShortPassing	88591.320731
2	HeadingAccuracy	84380.643243
17	Stamina	77956.075295
34	Weak Foot	76625.239490
15	ShotPower	76380.775189
8	LongPassing	75801.684470

	Feature	Latent User 8 Pseudocount
34	Weak Foot	125846.334660
16	Jumping	121488.278605
36	Work Rate	117811.248911
18	Strength	98168.703764
10	Acceleration	88753.605672
11	SprintSpeed	87062.946352
14	Balance	83380.309065
35	Skill Moves	79736.233863
12	Agility	77009.624247
17	Stamina	73122.256989

	Feature	Latent User 9 Pseudocount
36	Work Rate	99463.561634
12	Agility	78601.187405
8	LongPassing	75783.068922
9	BallControl	74106.744631
6	Curve	72465.053779
35	Skill Moves	72362.759994
23	Vision	66319.981794
17	Stamina	62894.124341
3	ShortPassing	62633.999028
24	Penalties	61496.817543

	Feature	Latent User 10 Pseudocount
14	Balance	120444.268481
12	Agility	110684.688697
34	Weak Foot	108851.037229
0	Crossing	107219.213487
9	BallControl	103930.155251
4	Volleys	102846.193158
7	FKAccuracy	102200.352477
19	LongShots	100215.718265
11	SprintSpeed	98708.849547
23	Vision	96559.797387

	Feature	Latent User 11 Pseudocount
14	Balance	190851.020544
36	Work Rate	189578.571801
11	SprintSpeed	182963.022272
10	Acceleration	179469.049468
35	Skill Moves	159333.721895
12	Agility	159049.096056
5	Dribbling	151024.282554
34	Weak Foot	149825.547790
0	Crossing	138783.825291
22	Positioning	131670.706790