ABSTRACT

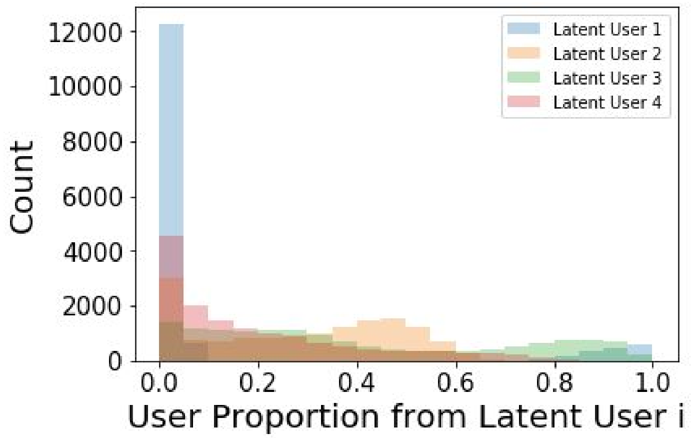
- Could machine learning and AI help to automate the process of finding the talent by taking and analyzing scouting report data?
- Scouts have a big responsibility in ensuring that the team they work for has the chance of obtaining the personnel that fits with the team and that will help take the team to the next level however that can involve a lot of on-site visits and paperwork.
- They would benefit greatly from having a means by which they can automate the learning process of the data and being able to identify key patterns that could be useful in deciding who they should complete a more detailed scout of.
- This project applies the Latent Dirichlet Model on a Fifa 19 dataset, which we treat as our preliminary “scouting reports” and attempts to find what the latent structure between the skill ratings of the players could tell us about other key characteristics of the players like nationality, body weight, market value, and potential.

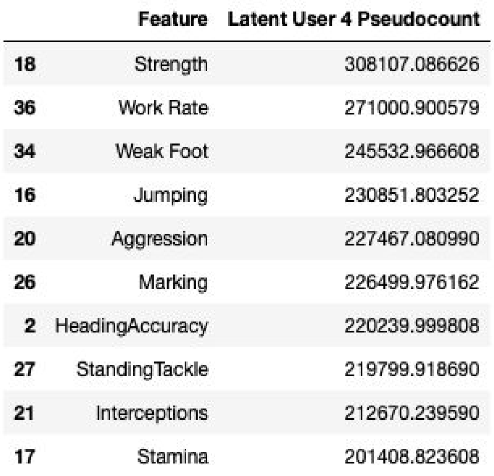
\

BACKGROUND & APPROACH

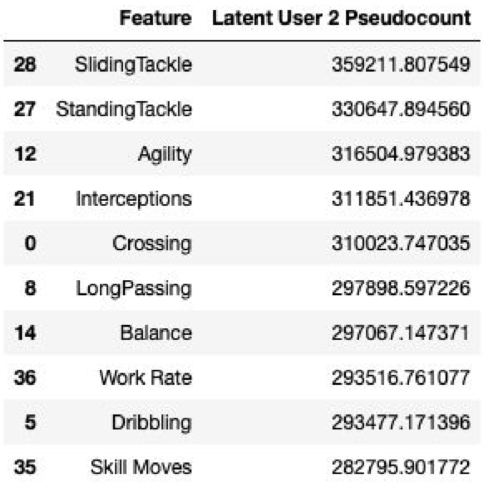
- **Preprocessing**
  - 80%/20% split of the data into a training set and a testing set.
  - Combined the columns with ‘Skill Moves’, ‘Weak Foot’, ‘Work Rate’, and all the defined skills from ‘Crossing’ to ‘Goalkeeping Reflexes’.
  - Converted the categorical values in the ‘Skill Moves’, ‘Weak Foot’, and ‘Work Rate’ columns into randomly generated continuous values to match the value type found in the defined skills columns
- **Feature Selection**
  - Dropped the rows of the matrices that contained null values in place for the defined skills
  - Following preprocessing I was left with a matrix of ***14525 rows x 37 columns***
- **One-Hot Encoded vs. Continuous Matrix**
  - Opted to use the matrix with continuous values to fit the LDA model rather than the one-hot encoded matrix as the continuous matrix will lead to more discrete latent relations as opposed to a binary or categorical matrix.
- **LDA: n_components = 4 vs. n_components = 11**
  - I fitted one LDA model where the number of components was equal to 4 and another where the number of components was equal to 11 however I opted to focus on the analysis on the model where the components was equal to 11
  - Reason: I expected the latent structure to be representative of the four different general positions on the soccer field: Goalkeeper, Defenders, Midfielders, Forwards. Consequently it won’t produce as interesting latent users as say a model where there were 11 components.
  - The reason for 11 components is that there are 11 men on the field for one team generally so it would be interesting to identify ‘team’ of latent users that the model outputs.
- **Other Data**
  - Since we are going to be analyzing the latent structure relationships with respect to other aspects of the scouting report data, we needed to process the columns with those pieces of data in the same manner.

LDA: n_components = 4

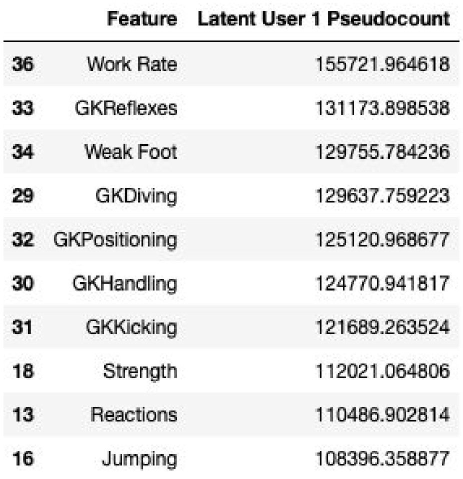




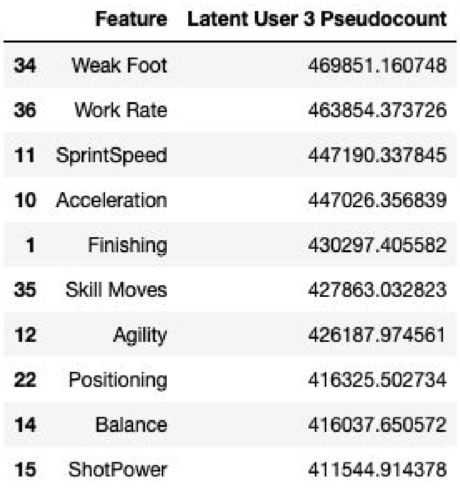
	Feature	Latent User 4 Pseudocount
18	Strength	308107.086626
36	Work Rate	271000.900579
34	Weak Foot	245532.966608
16	Jumping	230851.803252
20	Aggression	227467.080990
26	Marking	226499.976162
2	HeadingAccuracy	220239.999808
27	StandingTackle	219799.918690
21	Interceptions	212670.239590
17	Stamina	201408.823608



	Feature	Latent User 2 Pseudocount
28	SlidingTackle	359211.807549
27	StandingTackle	330647.894560
12	Agility	316504.979383
21	Interceptions	311851.436978
0	Crossing	310023.747035
8	LongPassing	297898.597226
14	Balance	297067.147371
36	Work Rate	293516.761077
5	Dribbling	293477.171396
35	Skill Moves	282795.901772



	Feature	Latent User 1 Pseudocount
36	Work Rate	155721.964618
33	GKReflexes	131173.898538
34	Weak Foot	129755.784236
29	GKDiving	129637.759223
32	GKPositioning	125120.968677
30	GKHandling	124770.941817
31	GKKicking	121689.263524
18	Strength	112021.064806
13	Reactions	110486.902814
16	Jumping	108396.358877

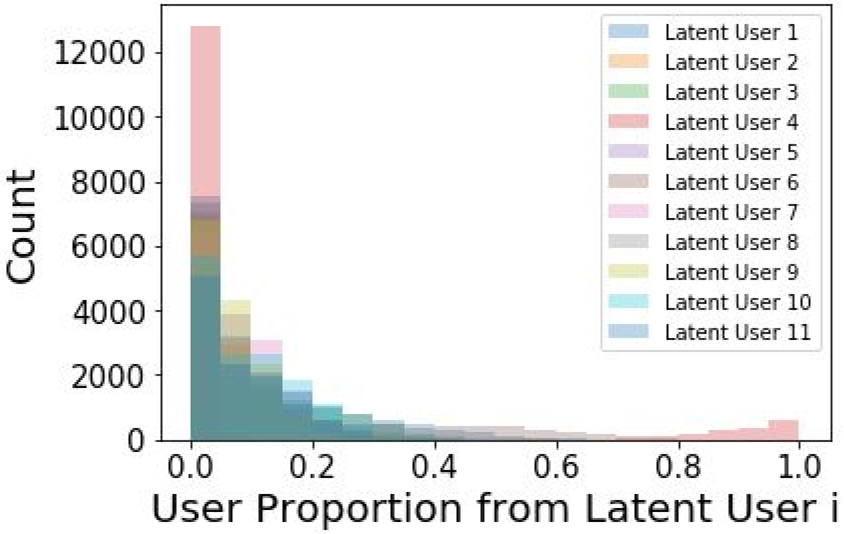


	Feature	Latent User 3 Pseudocount
34	Weak Foot	469851.160748
36	Work Rate	463854.373726
11	SprintSpeed	447190.337845
10	Acceleration	447026.356839
1	Finishing	430297.405582
35	Skill Moves	427863.032823
12	Agility	426187.974561
22	Positioning	416325.502734
14	Balance	416037.650572
15	ShotPower	411544.914378

LDA: n_components = 4

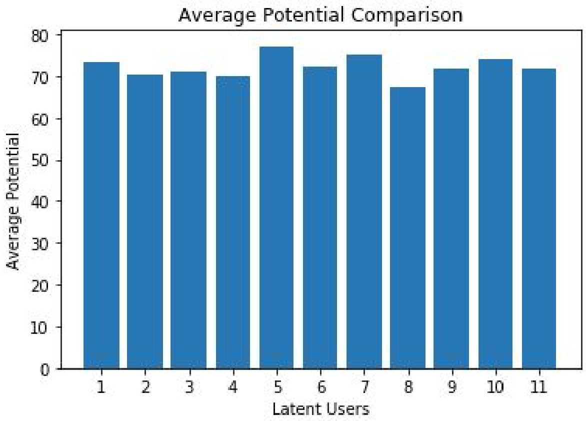
- As predicted each of the latent users represent one of the four general positions in soccer: l.u. 1 (Goalkeeper), l.u. 2 (Midfielder), l.u. 3 (Attacker), l.u. 4 (Defender)
- There are overlaps between the midfield latent user and the defender latent user as well as the midfield latent user and the attacker latent user. This makes sense since in the midfield position you are often required to be impactful both defensively and offensively.
- The difference in how much the 3 non-goalkeeper latent users relate to a wider variety of users versus how much the goalkeeper latent user relates to the users is also reflected by the huge gap in the proportion graph seen by latent user 1.
    - The goalkeeping latent user has the most users with a very high relation to it ~1 but also by far the most users with almost no relation to it and pretty much nothing in between
    - The other latent users have a more spread out proportion graph which demonstrates that many users share certain aspects of each of the latent users
- To test the fit of our model, I compared the score of the model with my training set to the score of the model with my test set.
    - Training Set Score: -6610.3053186717
    - Test Set Score: -6613.015651691198
    - The fact that the proportion of the difference between the two scores to the actual score is less than 1% shows that the model was well fitted and the outputs are accurate.

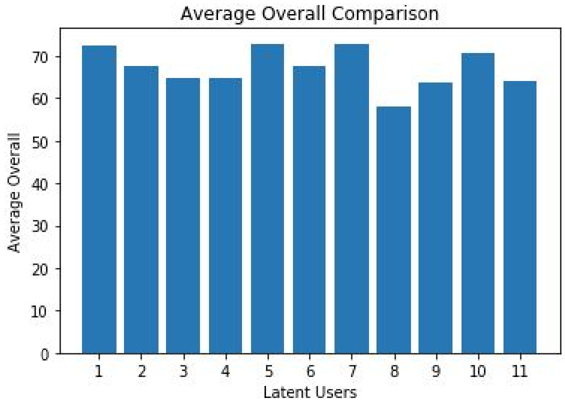
LDA: n_components = 11



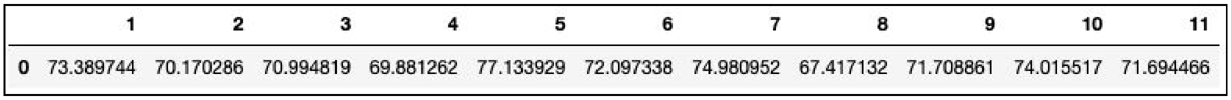
- With the increased number of components it is expected that the graph will take on more of a skewed left shape as the proportions of users would be more greatly distributed among the latent users.
- **Latent User 4:** Latent user 4 has a similar distribution to what we saw with the goalkeeper latent user with n_components = 4 and when we take a look more deeply it turns out that it is a goalkeeper latent user.
- **Fit Score:**
  - Training: -6628.69695522525 Test: -6631.598170223036 (Well-fitted)

LDA: n_components = 11 (Stats Analysis Examples)



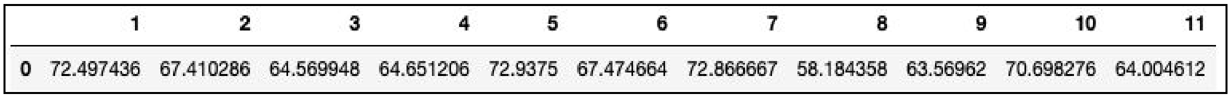


Potential (>.25)

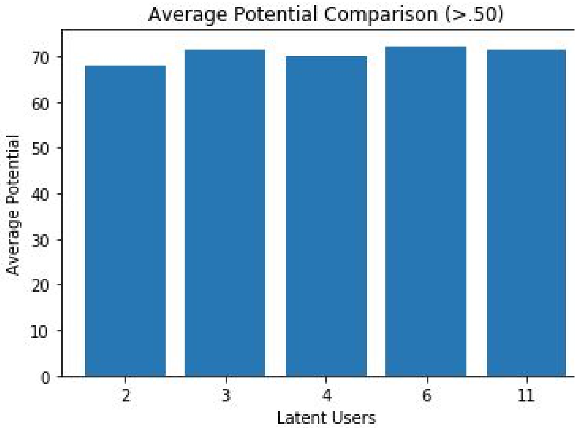


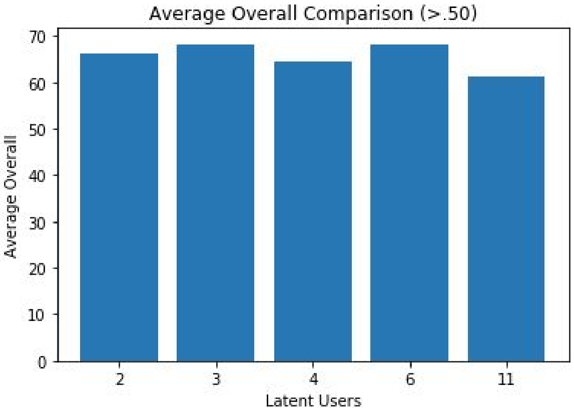
	1	2	3	4	5	6	7	8	9	10	11
0	73.389744	70.170286	70.994819	69.881262	77.133929	72.097338	74.980952	67.417132	71.708861	74.015517	71.694466

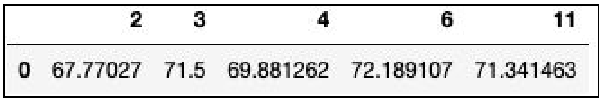
Overall (>.25)



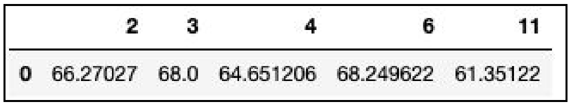
	1	2	3	4	5	6	7	8	9	10	11
0	72.497436	67.410286	64.569948	64.651206	72.9375	67.474664	72.866667	58.184358	63.56962	70.698276	64.004612





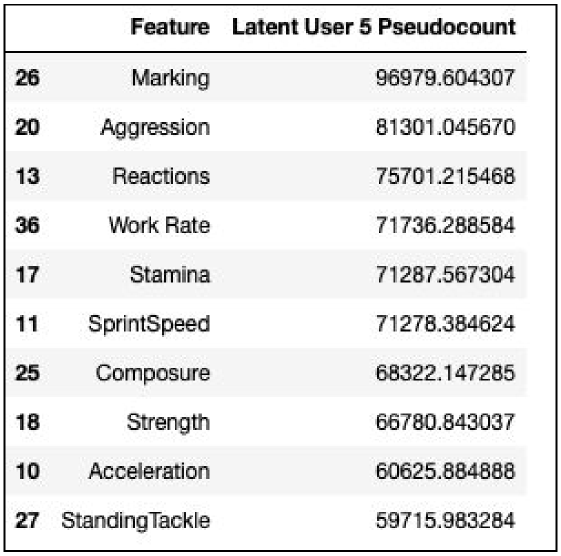


	2	3	4	6	11
0	67.77027	71.5	69.881262	72.189107	71.341463

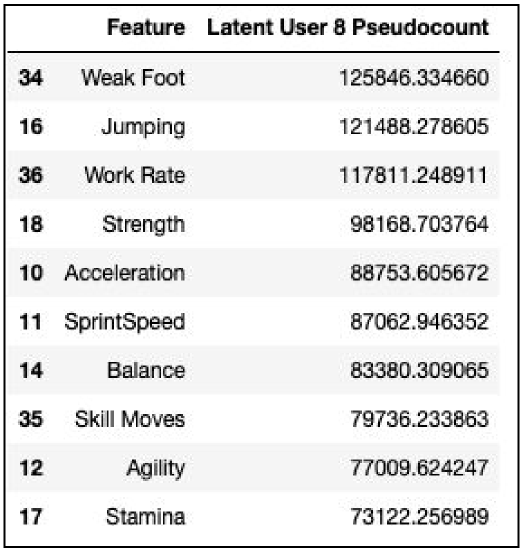


	2	3	4	6	11
0	66.27027	68.0	64.651206	68.249622	61.35122

LDA: n_components = 11 (Stats Analysis Examples)



	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 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

- Using the latent structure of the skill ratings in the report we can analyze how certain skill characteristics impact other aspects of the report. As an example, we will look at how the potential of the players and the overall of the players are impacted.
- The above bar graphs are the average potentials and average overalls calculated from the users whose proportionality rating was greater than .25 with respect to each latent user while the one below is likewise for proportionalities greater than .50. The need for a low proportionality threshold is due to the widespread latent user relationship with the users.
- From there we can delve into all sorts of exploration as to which characteristic seems to affect the overall and potential the most/least or not at all.
- Above is the top 10 characteristics of latent user 5, who had the greatest overall and potential in the (>.25) bar graph, and latent user 8 who had the lowest. We can compare them to see the characteristics that latent user 5 possesses and latent user 8 is missing. The characteristics that 5 possesses in the top 10 but 8 is missing are ‘Standing Tackle’, ‘Composure’, ‘Reactions’, ‘Aggression’, and ‘Marking’. We could go in-depth and see whether there is a trend associated with these characteristics among the other latent users so as to perhaps suggest a particular skill that a team should foremost desire in a player and look players with that skill.

RELATED WORK

[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/.