Supervised Machine Learning

Aarti R Jadhav 09/09/2022

Sports analytics is a field that is growing in popularity and application throughout the world. One of the open problems in this field is the assessment of football players based on their skills, physical attributes and market value.

We will tackle this problem using advanced supervised learning techniques like ridge and lasso regression, logistic regression, KNN, LDA, QDA, decision trees, random forests and additive models depending on the type of problem. Player attributes will be carefully chosen using feature selection techniques and cross-validation will be used to determine the model with the least error. Finally, based on the use case we will evaluate the performance of the respective models on the test set using metrics like Adjusted R^2, MSE, RMSE, confusion matrix and ROC curve.

Overview

Aim: Establish player assessment model to support player transfer decisions for football clubs.

Use Cases:

1. What variables drive the valuation of a player?
2. Do clubs need to look at players from specific nations while making transfer decisions?
3. Classify player work rate for better player management.
4. What physical conditioning should trainers focus on for a player who is transitioning from one position to another?

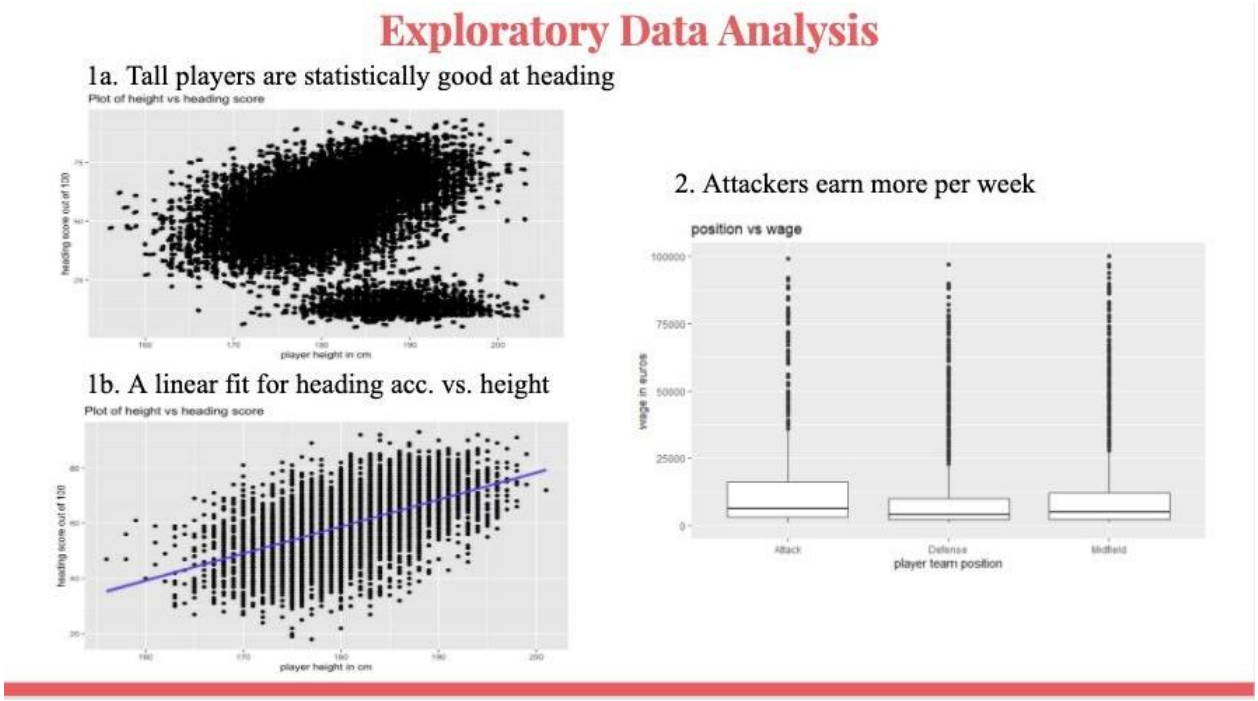
	short_name	work_rate	value_eur	team_position	nationality	preferred_foot	height_cm	age
0	L. Messi	Medium/Low	95500000	RW	Argentina	Left	170	32
1	Cristiano Ronaldo	High/Low	58500000	LW	Portugal	Right	187	34
2	Neymar Jr	High/Medium	105500000	CAM	Brazil	Right	175	27
3	J. Oblak	Medium/Medium	77500000	GK	Slovenia	Right	188	26
4	E. Hazard	High/Medium	90000000	LW	Belgium	Right	175	28

The primary aim of this project is to establish football player assessment models using machine learning techniques to support transfer decisions of football clubs.

We will be using publicly available datasets from Kaggle that contain players' data for the Career Mode from FIFA 2015 to FIFA 2020, where each player record is characterised by 104 features. Some of the few important player features include, year, age, body type, work rate, value, skills (e.g. pace, shooting, passing), wage, traits, position, nationality, club, ratings, preferred foot and physical attributes. These features will enable us to analyse the performance of players across seasons and build player skills assessment models.

Based on domain specific knowledge of the game, we now propose a few use cases that the management of any football team would be interested in to make transfer decisions.

1. What important features make a player valuable? Intuitively, if a player has high value, then the player must be highly skilled, and be good at his respective position (i.e. attacker, defender, midfielder, goalkeeper). This inturn has a positive impact on the rating.
2. If a team is scouting for players for specific positions, does the management look at players from specific nations while making transfer decisions? For example, it is said that players from Spain are good passers of the ball.
3. Attackers and defenders train based on the needs of their respective positions. Is it possible to identify the amount of work (high, medium, low) a player does to improve his attack and defense skills?
4. If a player wishes to transition to other positions, what skills must he improve to make that possible?



Tall players are statistically good at heading. We use a scatter plot for heading accuracy vs. height. The plot confirms that as the height of the player increases, the player's heading accuracy increases (Fig-1a).

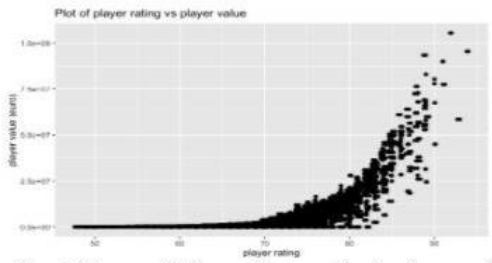
An interesting find here is the presence of two clearly distinguishable clusters, where a small cluster contains tall players with low heading accuracy. Upon further investigation, we observed that the small cluster contained only goalkeepers, reserves and substitute players. Filtering out these records, we observe a good linear fit for height against heading accuracy (Fig-1b). Even though we observe a positive correlation between heading accuracy and height, the residuals suggest that there exists a range for heading accuracy for players with the same height.

Attackers earn more per week compared to defenders and midfielders. We use a box plot of wage for different positions. The plot confirms that the median wage of attackers is greater than that of defenders and midfielders.

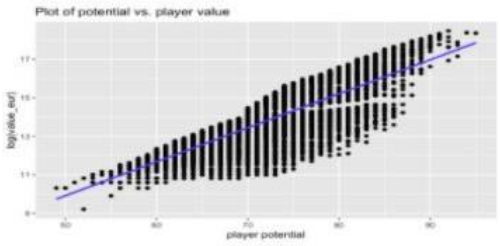
An interesting find here is the presence of many outliers in each position. The reason for this is the presence of highly skilled players in each position who earn more compared to others in the same position. This makes sense in that, for a team to perform well it not only needs good attackers but also good defenders and midfielders.

Exploratory Data Analysis

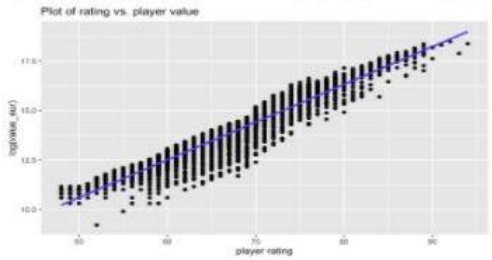
3a. Players with higher rating have higher value



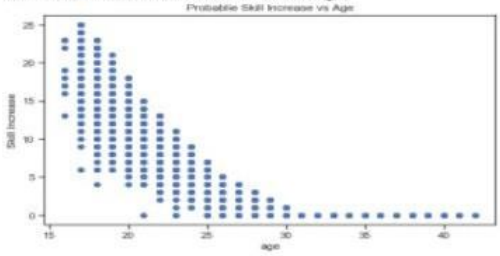
4. Players with higher potential have higher value



3b. A linear fit for rating vs. log(value_eur)



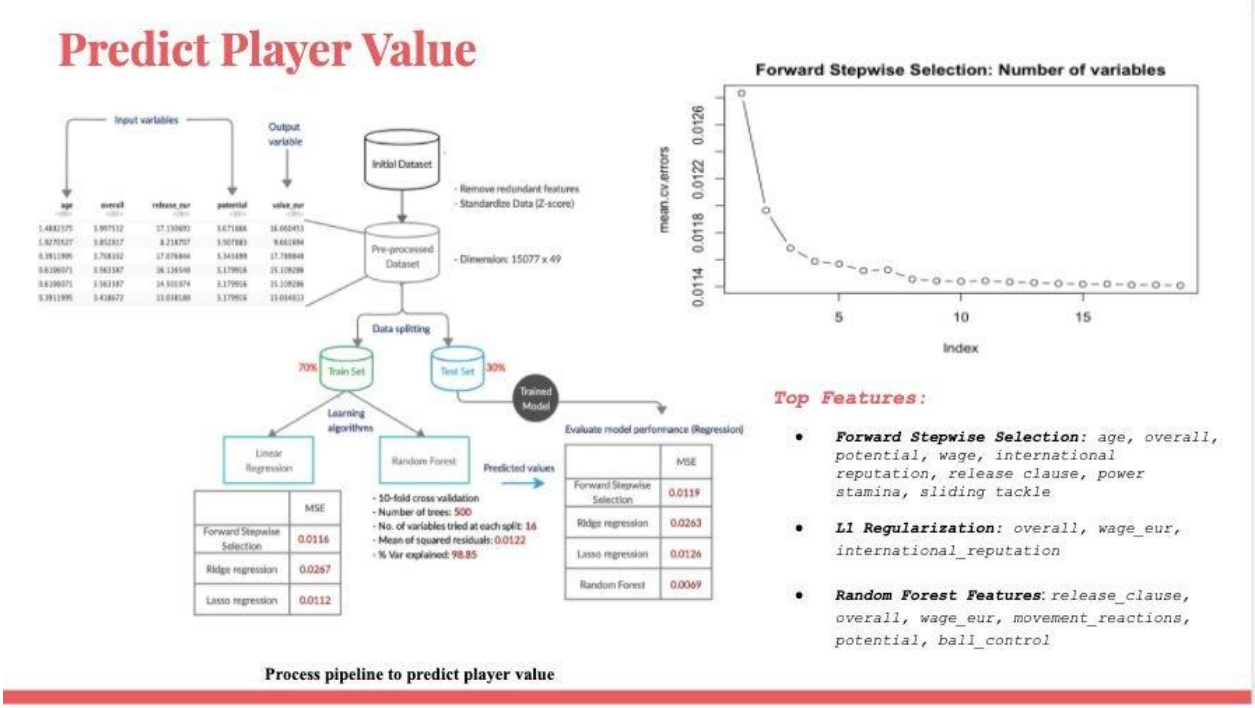
5. Potential decreases with age



There exists a positive correlation between player rating and value. We use a scatter plot for rating vs. value. The plot shows that as rating increases, there is an exponential growth in the value of the player (Fig-3a). This is confirmed by observing a linear fit on the log-transformation of value against rating (Fig-3b). An Adjusted R squared error of 0.9083 shows that a linear fit is a good choice to capture the relation between the predictor and target variable.

Value of a player increases with increase in potential (Fig-4). We use a scatter plot for potential vs. value. Potential is defined as the possible overall improvement of a player as the player's career progresses. The plot shows that as potential increases the value of the player increases. An Adjusted R squared error of 0.6072 shows that a linear fit is a good choice to capture the relation among the variables.

Potential of a player decreases with increase in age (Fig-5). As the age of a player increases, the scope for improvement in the player's abilities decreases because the older a player gets, the more he/ she is set in their ways. We use a scatter plot for potential vs. age. The plot shows that as age increases, potential decreases. After a player reaches an age of thirty, we observe that, potential completely drops to zero indicating no scope of improvement.



Process pipeline to predict player value

Based on domain knowledge, we know that the value of a player can depend on several variables. However, before applying any feature selection technique, we removed redundant features and standardized the data to perform hypothesis testing to verify whether there exists a relationship between player attributes and value. The results obtained from fitting a linear model rejected the null hypothesis allowing us to apply subset selection and regularization techniques.

Applying forward stepwise selection we obtained eight predictors (age, overall, potential, wage_eur, international_reputation, release_clause_eur, power_stamina and sliding_tackle) that sufficiently predicted player value. The eight predictors were chosen from the model that gave the smallest error using cross-validation. Fitting a linear model to predict value using the eight features obtained from forward stepwise selection we achieve a mean squared error of 0.0116 and Adjusted R-squared of 0.9896 on the train set, and 0.0119 on the test set.

L1 regularization chooses overall, wage_eur and international_reputation as features to predict player value, and the coefficients of all the other features shrink to zero. The linear fit using features from L1 regularization gives a mean squared error of 0.0112 on the train set, and 0.0126 on the test set.

Since, L2 regularization does not shrink the coefficients of all the variables to zero, a lot more predictors are part of the linear regression model as compared to the fit using predictors from L1 regularization. The linear fit for L2 regularization gives a mean squared error of 0.0267 on the train set, and 0.0263 on the test set.

Finally, a Random Forest model with 10-fold cross-validation was fit to the train data to predict value. The cross-validated model gives a mean squared error of 0.0069 on the test set.

Evaluate Player Value Model (Test Data)

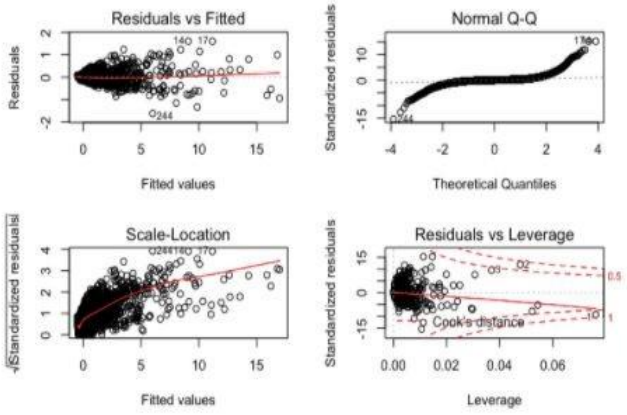
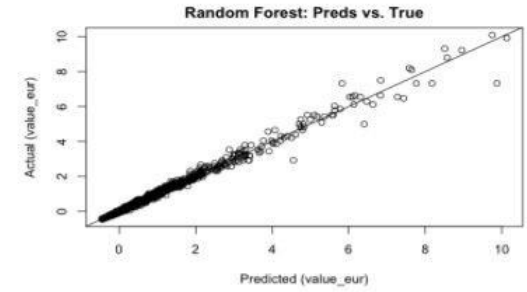


Fig: Diagnostic Plots for Linear Regression Analysis



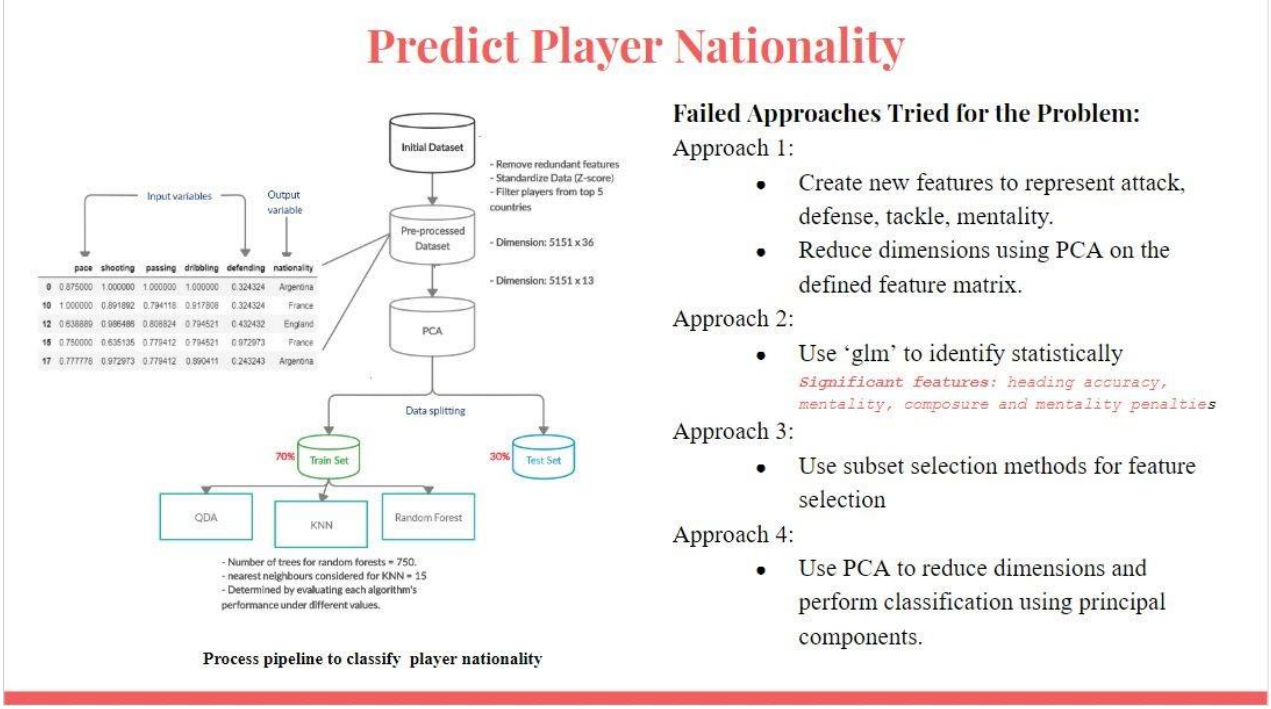

```
Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -6.946e+05  1.213e+05  -5.724 1.07e-08 ***
age                       -1.474e+04  2.880e+03  -5.119 3.12e-07 ***
overall                    3.037e+04  2.578e+03  12.014  < 2e-16 ***
potential                 -2.180e+04  2.426e+03  -8.983  < 2e-16 ***
wage_eur                   3.476e+00  5.400e-01   6.436 1.28e-10 ***
international_reputation   4.735e+05  2.164e+04  21.874  < 2e-16 ***
release_clause_eur         4.935e-01  1.139e-03 433.405  < 2e-16 ***
power_stamina              4.816e+03  6.242e+02   7.715 1.31e-14 ***
defending_sliding_tackle  -2.221e+03  3.327e+02  -6.676 2.58e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 610500 on 10545 degrees of freedom
Multiple R-squared:  0.9896,    Adjusted R-squared:  0.9896
F-statistic: 1.253e+05 on 8 and 10545 DF,  p-value: < 2.2e-16
```

The diagnostic plots for linear regression using predictors obtained from forward stepwise selection convery the following about the fit.

- **Residual vs Fitted plot:** Residuals are equally spread around the horizontal line near zero, hence no model assumptions have been violated.
- **Normal Q-Q plot:** In the normal Q-Q plot the points fall along a line in the middle of the graph, but curve off in the extremities.
- **Scale Location plot:** Residuals spread wider and wider, the red smooth line is not horizontal and shows a steep angle, variance is not equally spread among the predictors. This conveys the presence of heteroscedasticity.
- **Residuals vs Leverage plot:** Players in rows 1, 2 and 3 have high leverage. Not surprised, those players are L. Messi, Cristiano Ronaldo and Neymar Jr.

All the plots convey the presence of several outliers, and this is because there are several highly skilled players who have a higher value compared to other players in the dataset. These outliers influence the model fit. A linear model was fit using the same predictors after removing the outliers, and the performance of the model did not change much on the test data.

Finally, the more important question is what variables have an effect on a player's value. There is a great overlap in the features selected by each feature selection technique. Further, analyzing the regression coefficients from fitting a linear model with predictors obtained from forward stepwise selection we can notice how overall (rating), wage_eur, internationl_reputation are few of the player attributes that have a positive effect on a player's value when when all the attributes are held constant. Whereas, the model suggests that if a player gets a year older his value is bound to decrease by € 14,740 and if his potential (i.e. player's room for improvement) increases by one point, his value will decrease by € 21,800 when all other attributes are constant.



	pace	shooting	passing	dribbling	defending	nationality
0	0.875000	1.000000	1.000000	1.000000	0.324324	Argentina
10	1.000000	0.891892	0.794118	0.917808	0.324324	France
12	0.638889	0.986486	0.808824	0.794521	0.432432	England
15	0.750000	0.635135	0.779412	0.794521	0.972973	France
17	0.777778	0.972973	0.779412	0.890411	0.243243	Argentina

Process pipeline to classify player nationality

Classify nationality of a player based on skill attributes. These include 35 attack and defense attributes like attack crossing, attack finishing, sliding tackle, standing tackle. We tried several approaches to solve this classification problem. There are records of players from 160 countries. There are less than 10 countries with a number of records more than 300. Hence, we are attempting to classify players from top five countries in terms of number of players.

Approach - 1

We created a representative feature(mean) for attack, skill, movement, power, mentality and defense for each player. For example, mean of attacking_crossing, attacking_finishing, attacking_heading_accuracy, attacking_short_passing and attacking_volleys represents attack for a player. We use k-nearest neighbour, linear discriminant analysis and quadratic discriminant analysis to classify players.

Approach - 2

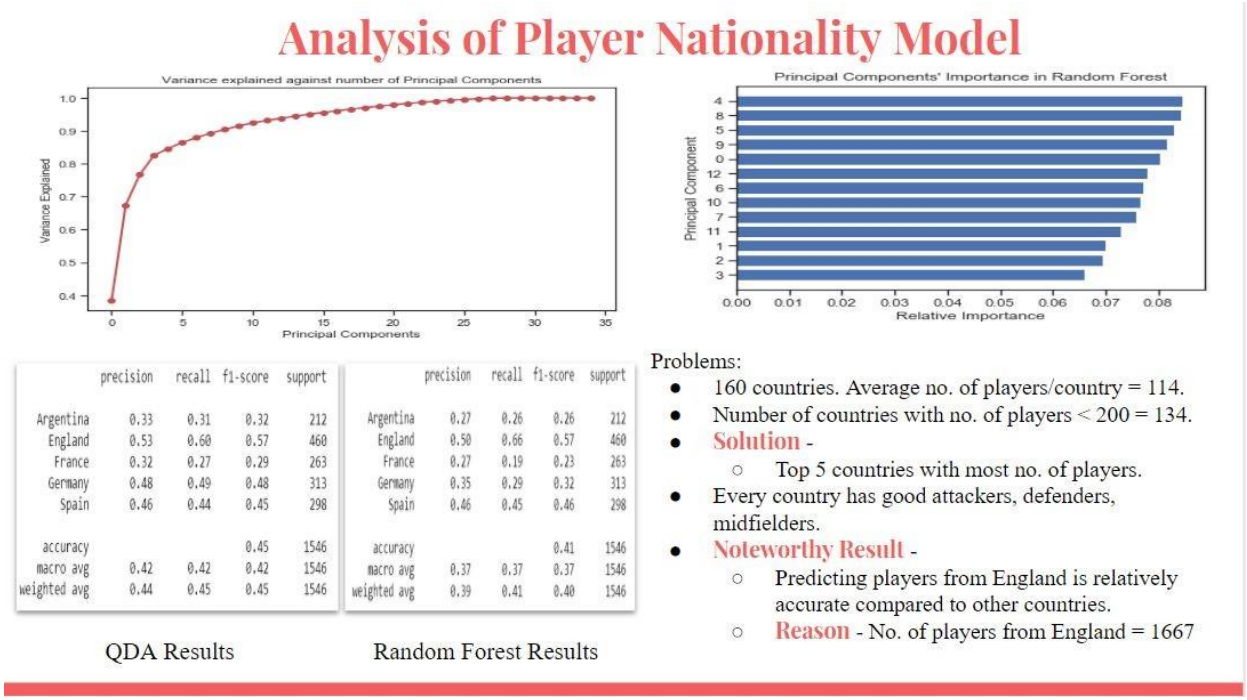
Identify significant features from 35 features that contribute to classifying nationality. Based on domain knowledge, we suspect that a player cannot be classified to a country based on the obtained attributes.

Approach - 3

Forward subset selection behaved similarly with slightly improved results.

Approach - 4

The only approach remaining was to consider every feature in the original matrix and perform PCA on it.



	precision	recall	f1-score	support
Argentina	0.33	0.31	0.32	212
England	0.53	0.60	0.57	460
France	0.32	0.27	0.29	263
Germany	0.48	0.49	0.48	313
Spain	0.46	0.44	0.45	298
accuracy			0.45	1546
macro avg	0.42	0.42	0.42	1546
weighted avg	0.44	0.45	0.45	1546

	precision	recall	f1-score	support
Argentina	0.27	0.26	0.26	212
England	0.50	0.66	0.57	460
France	0.27	0.19	0.23	263
Germany	0.35	0.29	0.32	313
Spain	0.46	0.45	0.46	298
accuracy			0.41	1546
macro avg	0.37	0.37	0.37	1546
weighted avg	0.39	0.41	0.40	1546

The results obtained for the various approaches are described below.

- **Approach - 1**
  - QDA performs best with an accuracy of 35%, but recall for 'France' is 0.02 which is the best among used models.
  - Applying PCA on the feature matrix didn't change results.

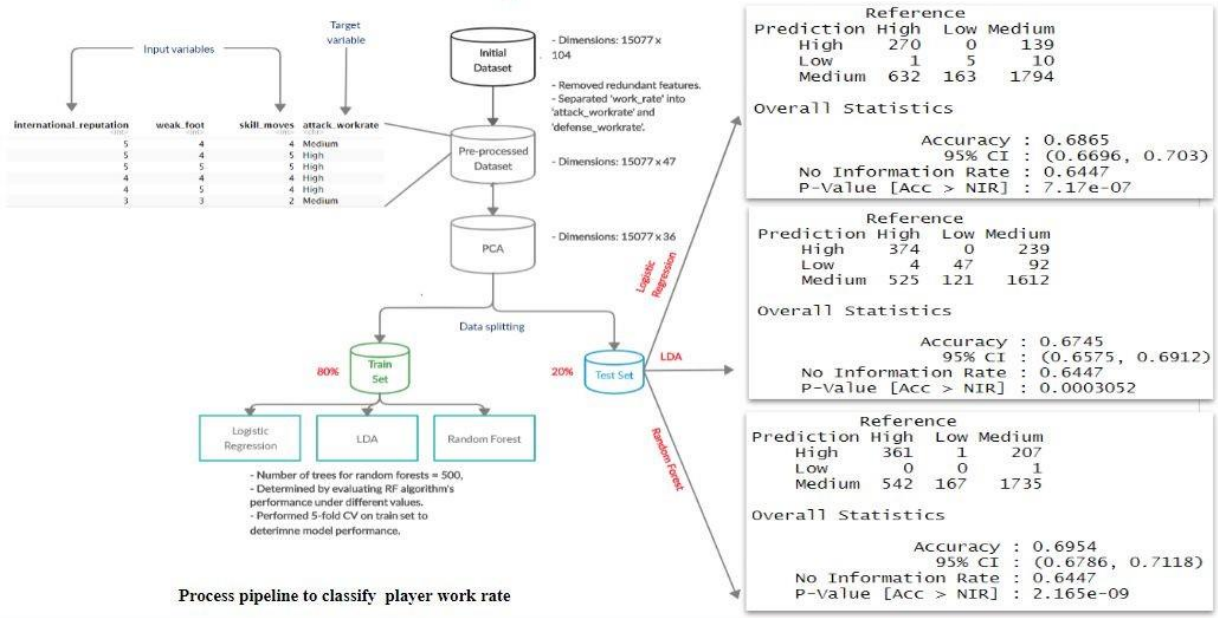
- **Approach - 2**
  - Classifying(using the above models) players using only significant features yielded an accuracy of 30% with recalls for three countries (Argentina, France, Germany) close to zero.

- **Approach - 3**
  - Forward subset selection resulted in an accuracy of 32% with recalls for the same three countries as above close to 0.1 as opposed to 0.

- **Approach - 4**
  - PCA on the feature matrix shows that 90% variance in data can be explained by 13 principal components.
  - QDA and Random Forest perform best in this scenario with QDA slightly better.
  - Random Forest on the feature matrix reveals that no principal component particularly dominates classification.

Players belonging to England are consistently classified better compared to other countries as due to the presence of many players from the country as opposed to other countries.

Predict Player Attack Work Rate



Process pipeline to classify player work rate

Initial dataset had 15077 records and 104 features. Removed all attributes not pertaining to a player's physical, mental and in-game attributes. This resulted in a dataset with 15077 records and 47 features. Split 'work_rate' attribute into 'attack_workrate' and 'defense_workrate'. Examples of attributes included in models to classify 'attack_workrate' : 'weak_foot', 'power_stamina', 'mentality_positioning' and 'pace' etc.

The attribute 'attack_workrate' is an indication of how hard a player works on the pitch to contribute towards his team's goal scoring and possession build-up. "What physical, mental and in-game attributes contribute towards a player's attack workrate?", is a question that a team's manager/coach and analyst often analyse after every match. Tracking and recording these attributes by conducting numerous tests on a player helps answer this question. Building classification models often help a team's analyst to predict a new player's likely 'attack_workrate' given his attributes.

Classification models used to predict player 'attack_workrate':

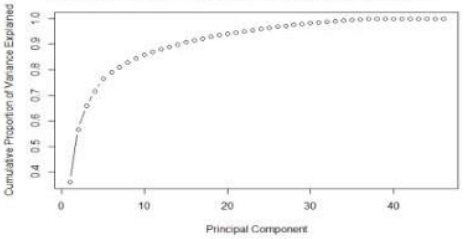
1. Logistic Regression
2. Linear Discriminant Analysis
3. Random Forest (rf model to highlight important features affecting a player's 'attack_workrate')

Metric used to evaluate model performance:

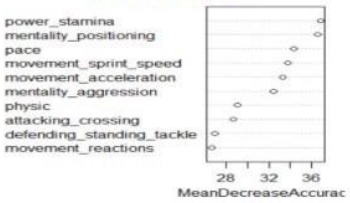
- Confusion matrix
- Sensitivity (TPR) and specificity (TNR) of individual classes ('High', 'Medium' and 'Low')
- AUC of ROC curve for the 3 models

Analysis of Attack Work Rate Prediction Models

Cumulative proportion of variance explained by PCs



Important features in Random Forest model



AUC of ROC on PCA test set

Logistic Regression	LDA	Random Forest
0.7694	0.7912	0.7819

Outline for the approach:

- PCA on pre-processed dataset to reduce feature space to a 36 dimensional space.
- Fit models on PCA train set (**5-fold CV**).
- Determine AUC of ROC for the models on PCA test set.

Results:

- Random Forest does a bad job in predicting **'Low'** class (**TPR=0**).
- LDA model does the best job in predicting **'Low'** class.
- LDA handles class imbalance well.
- Random Forest model again does a bad job in predicting **'Low'** class (**TPR=0**) on the PCA test set.

Fit Logistic Regression, Linear Discriminant Analysis and Random Forest models on pre-processed train set (5-fold CV) to classify 'attack_workrate'. Principal Component Analysis on pre-processed dataset to reduce feature space to a 36 dimensional space (since, 97% of variance in 'attack_workrate' is explained by 36 principal components) . Fit same models on PCA train set (5-fold CV) with 36 PCs and target variable to evaluate performance on PCA test set. Determine AUC of ROC curves for the 3 models on PCA test set.

The results obtained for the models are described below:

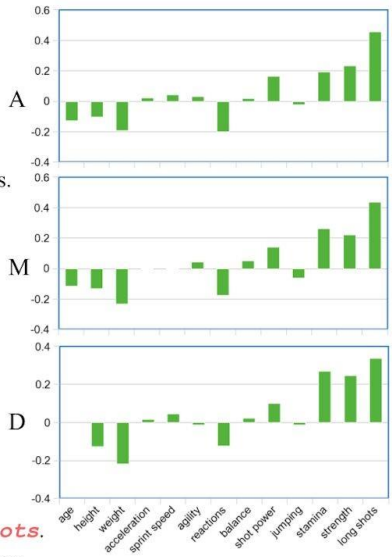
Classification accuracy of 3 models is similar (~67%) on a pre-processed test set. The Random Forest model does a bad job in predicting 'Low' class (True Positive Rate=0) whereas the LDA model does the best job in predicting 'Low' class. No significant increase in accuracy of 3 models after performing PCA. Though the LDA model has the least accuracy, it does the best job in handling class imbalance which is suggested by the AUC of ROC curve of LDA model on PCA test set which is the highest amongst all 3 models. The Random Forest model does a bad job in predicting 'Low' class (TPR=0) on the PCA test set too.

Attributes like 'power stamina', 'mentality positioning', 'pace', 'movement sprint speed' and 'mentality_aggression' affect a player's attack work rate the most. A team's manager or analyst can monitor these attributes to determine if a player needs to improve in order to work harder to contribute to the team's goal scoring and possession build-up. Scouting and signing new players to bolster a team's forward/attack line can be done by predicting the attack workrate given a potential player's aforementioned attributes.

Physical Attributes and Player Position

Part 1: "Given a player's physical condition, which position is he best suited to?"

1. *Consolidate positions* into Attacker (A), Midfielder (M), Defender (D), and Goalkeeper (G).
2. *Isolate features* that reflect individual physical condition/fitness level.
3. *Multinomial logistic regression* (with goalkeepers as reference position).
4. *Analyze coefficients* for insights into relative physical condition between positions.



Confusion Matrix and Statistics

Prediction	Reference G	A	D	M
G	1859	0	22	0
A	1	1389	250	693
D	13	299	4313	810
M	0	1582	1009	4866

Statistics by Class:

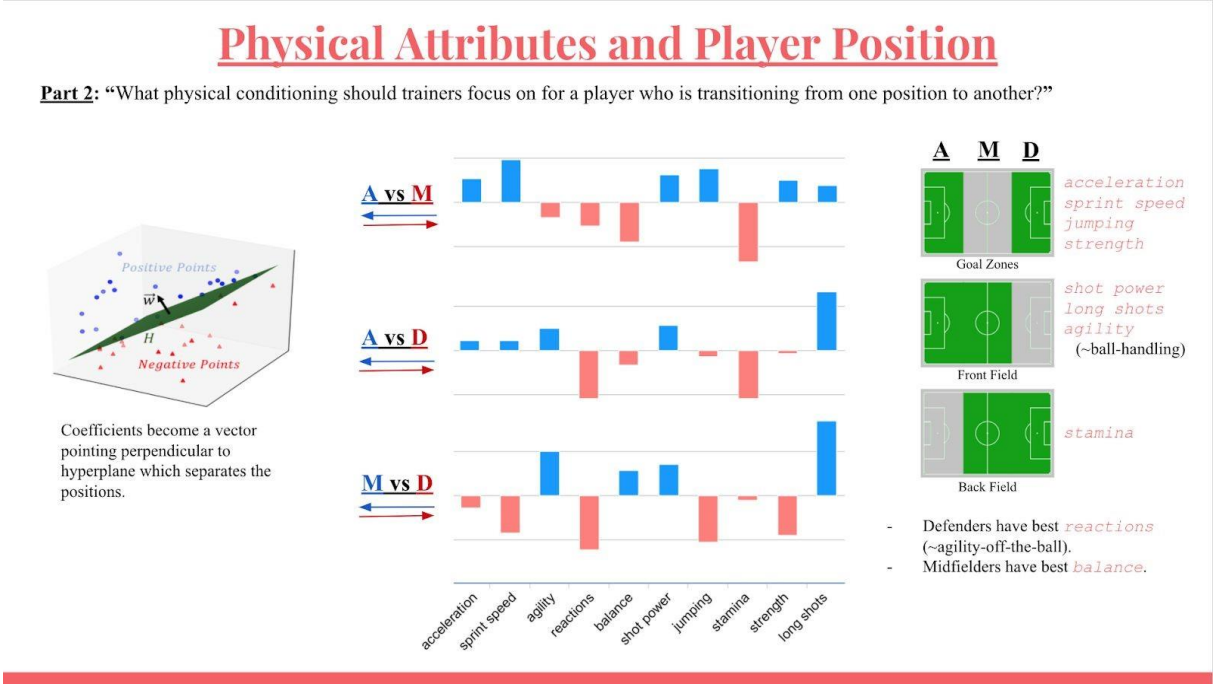
	Class: G	Class: A	Class: D	Class: M
Sensitivity	0.9925	0.4248	0.7710	0.7640
Specificity	0.9986	0.9318	0.9025	0.7587
Pos Pred Value	0.9883	0.5954	0.7936	0.6525
Neg Pred Value	0.9991	0.8727	0.8902	0.8442
Prevalence	0.1095	0.1912	0.3270	0.3723
Detection Rate	0.1087	0.0812	0.2521	0.2845
Detection Prevalence	0.1100	0.1364	0.3177	0.4359
Balanced Accuracy	0.9955	0.6783	0.8368	0.7613

Results:

- Goalkeepers on average older, taller, heavier (heavier because taller), and have quicker ***reactions***.
- A, M, and D beat G in ***shot power***, ***stamina***, ***strength***, and ***long shots***.
- Nearly identical coefficient plots obscure differences between non-G field players.

The question which motivated the following explorations was: "Given a player's physical condition, which position is he best suited to?" In other words, if we take a player and put them through a barrage of physical tests measuring physical fitness along several different dimensions and record their performance, can we determine which position they are most naturally suited to play? To answer this we began by collapsing the plethora of different position types into four basic position types: Attacker (A), Midfielder (M), Defender (D), and Goalkeeper (G). We then selected the features reflecting physical attributes dependent only on the individual player and not on his team. These were *age*, *height*, *weight*, *acceleration*, *sprint speed*, *agility*, *reactions*, *balance*, *shot power*, *jumping*, *stamina*, *strength*, and *long shots*. These categories are fairly self-explanatory. *Agility* can more aptly be described as ball-handling, or agility-on-the-ball whereas *reactions* can more aptly be described as agility-off-the-ball. *Long shots* refers to shot accuracy at a distance.

Once the data was appropriately cleaned, we split the data into a training and testing set and ran a multinomial logistic regression on the data with the Goalkeeper (G) position as the reference. This model yielded strong and meaningful results. It was particularly effective at differentiating Goalkeepers (G) from field players (A), (M), and (D). This makes intuitive sense, as field players all spend considerable time running up and down the pitch and cannot touch the ball with their hands. The goalkeeper has a restricted domain near his goal and can use his hands. This sharp distinction dominates the structure of our variable coefficients. We can see clearly that goalkeepers are on average older, taller, heavier, and have quicker *reactions*. We can also see that field players have greater *shot power*, *stamina*, *strength*, and *long shots*. While *strength* may come as a bit of a surprise, it is clear that the position of goalkeeper does not select for players who have strong *shot power*, accurate *long shots*, or *stamina*. A goalkeeper rarely gets a chance to use these skills. Attackers and defenders were also well-differentiated from one another, whereas midfielders are harder to tell apart from both attackers and defenders, reflecting the gradual transition in player responsibilities from the back of the field to the front.



In order to better understand the differences between attackers, defenders, and midfielders, we were guided by a plausible practical question that a coach might have: "What physical conditioning should players focus on when transitioning to play a new position?" To answer this question we trained three separate binomial logistic regressions on the pairs (Attacker, Midfielder), (Attacker, Defender), and (Midfielder, Defender). Because we wanted to focus only on attributes that a player could conceivably improve through training, we removed *age*, *height*, and *weight* from the features. While *weight* can be affected by training, it is highly correlated with *height*.

The coefficients of any of our pair models constitute a vector perpendicular to the hyperplane separating the two positions. This means that moving along this vector from any point takes us most efficiently towards the positively labeled position, and moving in the opposite direction takes us most efficiently towards the negatively labeled position. Naturally, we do not want to actively sabotage a player's physical abilities, although we might be willing to tolerate some reduction in ability due to underuse. Thus, if Attacker (A) is the positive category and Midfielder (M) is the negative category and we have a midfielder who wants to become an attacker, focusing on the largest positive coefficients (i.e. *sprint speed*) will turn our midfielder into an effective attacker. If, in the same situation, we have an attacker who wants to become a midfielder, focusing on the largest negative coefficients (i.e. *stamina*) will turn our attacker into an effective midfielder. (We omit numbers in the slide because, since these coefficients represent a vector, it is the coefficient ratios that are important and not the coefficient magnitudes.)

Our analysis reveals several very sensible patterns. Attackers and defenders dominate in *acceleration*, *sprint speed*, *jumping*, and *strength*. This reflects periodic rapid and intense activity in the goal zones. Attackers and midfielders dominate in *shot power*, *long shots*, and *agility* (~ball-handling). This reflects the attacking nature of the forward positions and the fact that defenders are most active when their team does not possess the ball. Midfielders and defenders dominate *stamina*. This reflects the constant contest for possession in the midfield and the constant vigilance of the defence.