FIFA Player Assessment Model & Analytics

DS 5220: Supervised Machine Learning

Akshit Jain
Naga Santhosh Kartheek Karnati
Praharsha Singaraju
Thomas Lindstrom-Vautrin

Overview

- The primary aim of this project is to establish a football player assessment model using machine learning techniques to support transfer decisions of football clubs.
- The dataset contains players’ records from FIFA 2015 to FIFA 2020 - (Matrix - 18,278 x 104).
- Important player features include:
    - work rate, value, position, nationality, skills, preferred foot and physical attributes
- These features will enable us to analyse the performance of players.

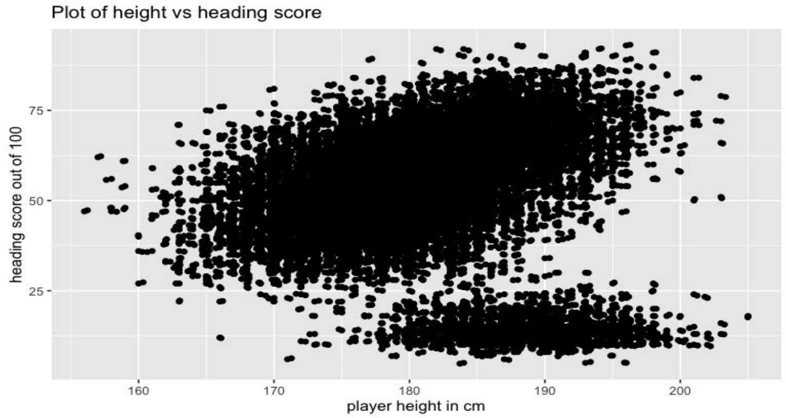
Use Cases:

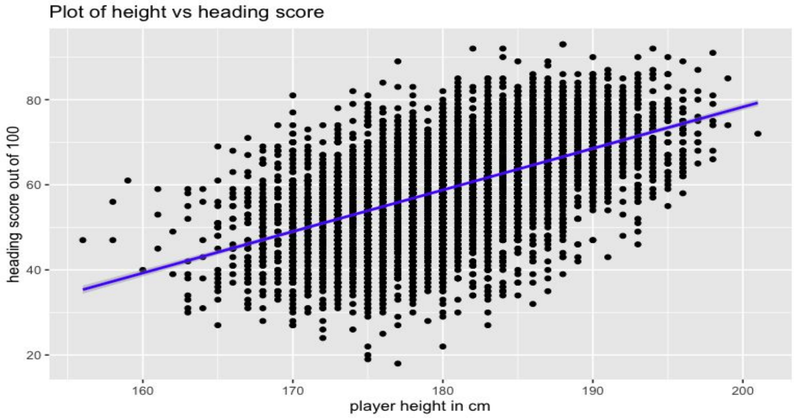
1. What variables drive the valuation of a player?
2. Do clubs need to look at players from specific nations while making transfer decisions?
3. Classify player work rate for better player management.
4. What physical conditioning should trainers focus on for a player who is transitioning from one position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

	short_name	work_rate	value_eur	team_position	nationality
0	L. Messi	Medium/Low	95500000	RW	Argentina
1	Cristiano Ronaldo	High/Low	58500000	LW	Portugal
2	Neymar Jr	High/Medium	105500000	CAM	Brazil
3	J. Oblak	Medium/Medium	77500000	GK	Slovenia
4	E. Hazard	High/Medium	90000000	LW	Belgium

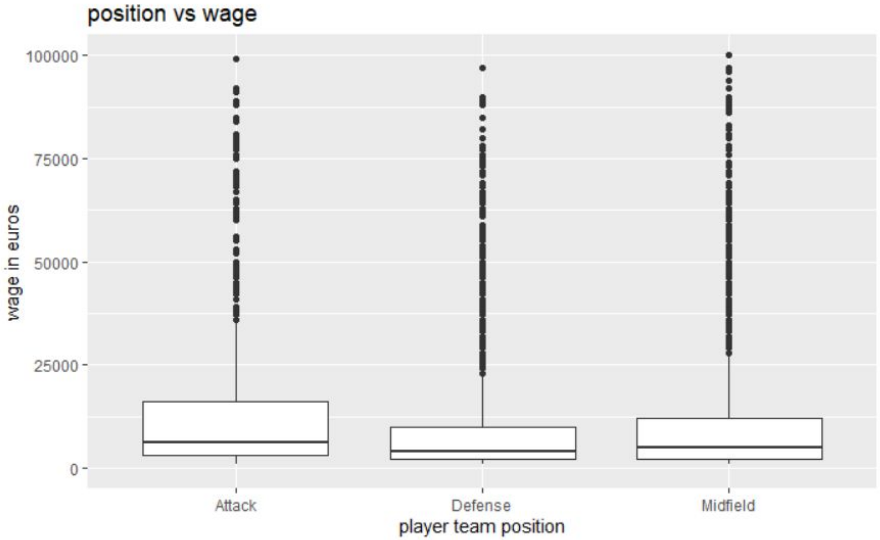
Exploratory Data Analysis

1. Tall players are statistically good at heading



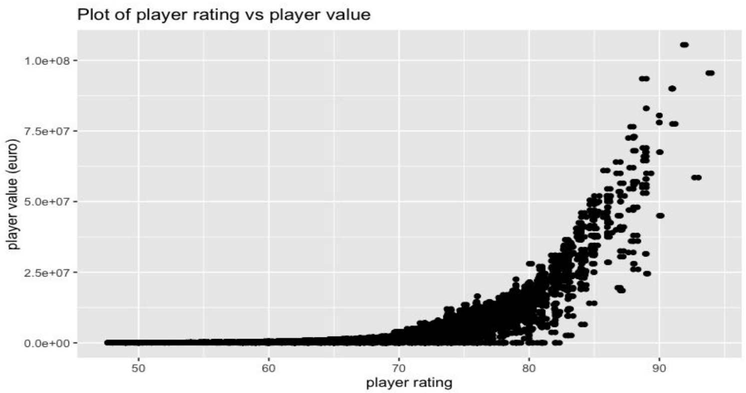


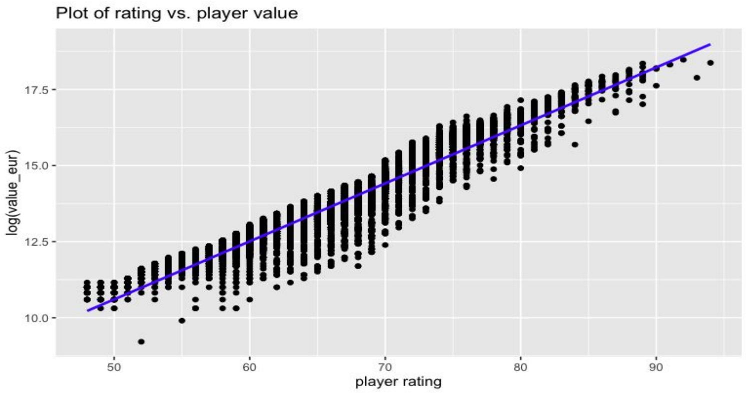
2. Attackers earn more per week



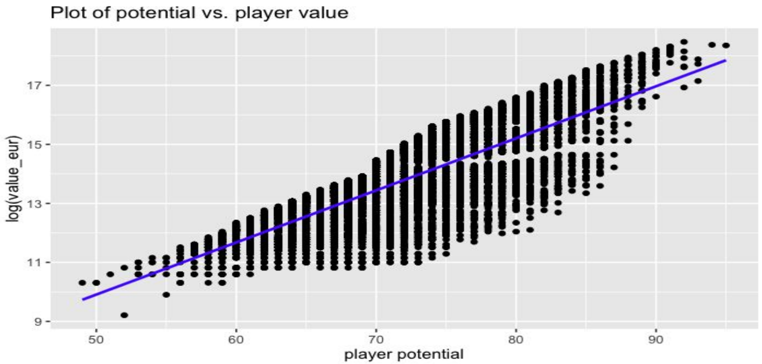
Exploratory Data Analysis

3. Players with higher rating have higher value

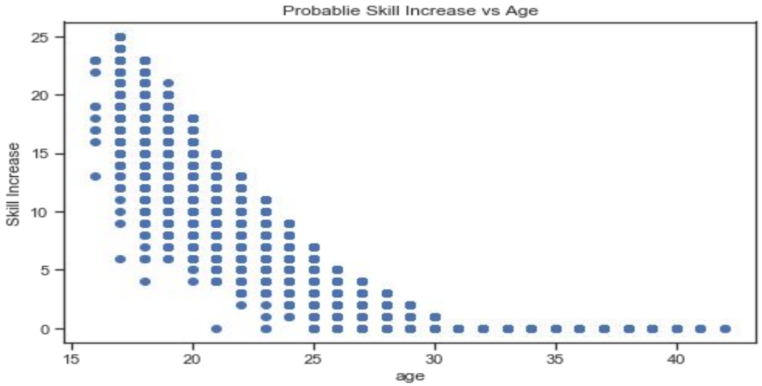




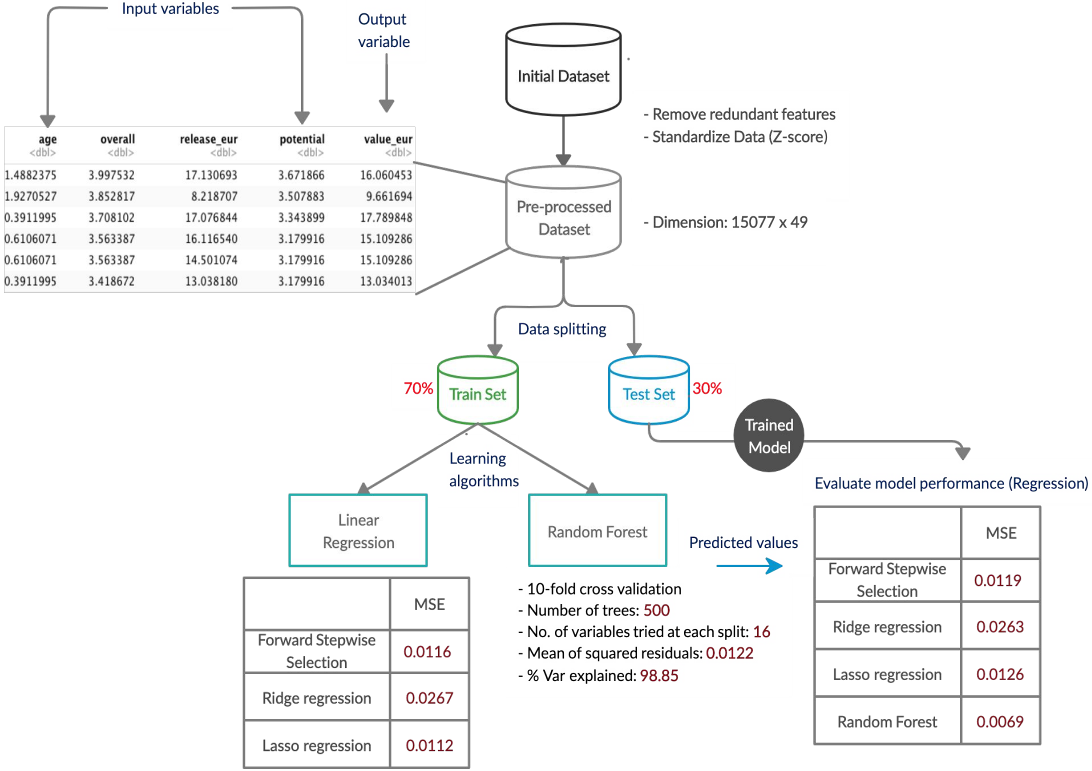
4. Players with higher potential have higher value



5. Potential decreases with age



Predict Player Value

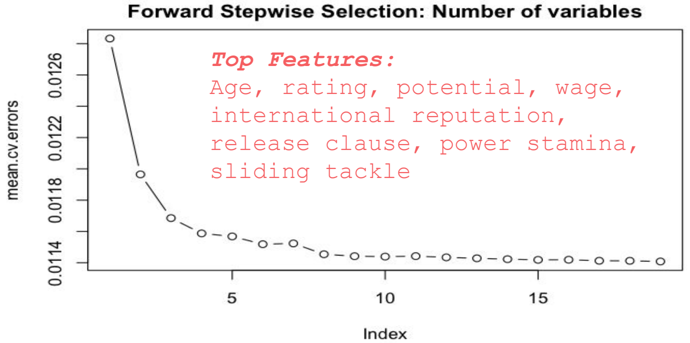


age <dbl>	overall <dbl>	release_eur <dbl>	potential <dbl>	value_eur <dbl>
1.4882375	3.997532	17.130693	3.671866	16.060453
1.9270527	3.852817	8.218707	3.507883	9.661694
0.3911995	3.708102	17.076844	3.343899	17.789848
0.6106071	3.563387	16.116540	3.179916	15.109286
0.6106071	3.563387	14.501074	3.179916	15.109286
0.3911995	3.418672	13.038180	3.179916	13.034013

	MSE
Forward Stepwise Selection	0.0116
Ridge regression	0.0267
Lasso regression	0.0112

	MSE
Forward Stepwise Selection	0.0119
Ridge regression	0.0263
Lasso regression	0.0126
Random Forest	0.0069

Process pipeline to predict player value

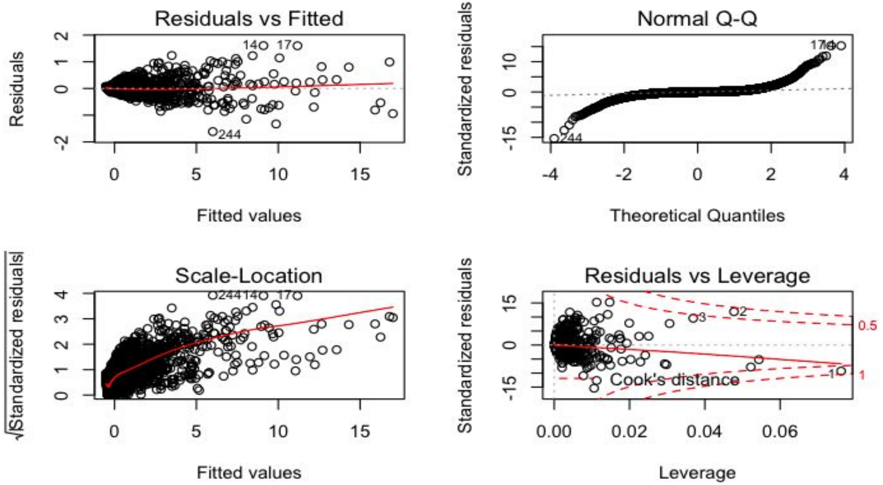


Methods Applied:

1. Performed hypothesis testing to determine whether there exists a relationship between player's attributes and value.
2. Applied forward stepwise selection to obtain a subset of player attributes that help explain player value.
3. Regularization Techniques (L1 and L2 Regression) -
   a. ***L1:*** *overall, wage_eur, international_reputation*
4. Fitted a Linear Regression model with features obtained from subset selection.
   a. ***Adjusted R-squared:*** 0.9896
5. Fitted a Random Forest model with 10-fold cross-validation - split variable decided randomly from 16 variables.
   a. ***Important features:*** *release clause, overall, wage eur, movement reactions, potential, ball_control*

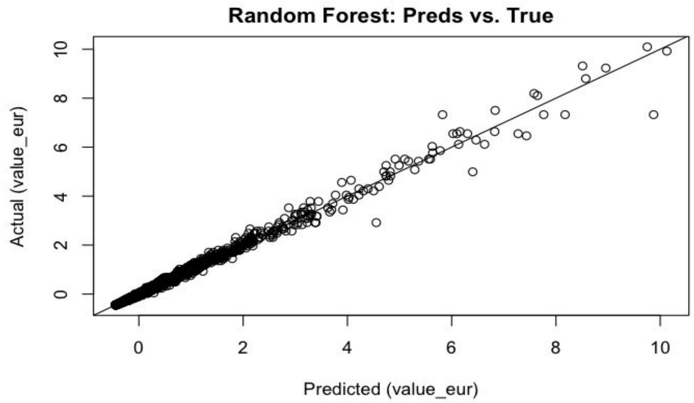
Evaluate Player Value Model (Test Data)

Diagnostic Plots for Linear Regression Analysis



- [T-L] Residuals are equally spread around the horizontal line near zero, hence no model assumptions have been violated.
- [T-R] In the normal Q-Q plot the points fall along a line in the middle of the graph, but curve off in the extremities.
- [B-L] 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.
- [B-R] Players in rows 1, 2 and 3 have high leverage. Not surprised, those players are L. Messi, Cristiano Ronaldo and Neymar Jr.

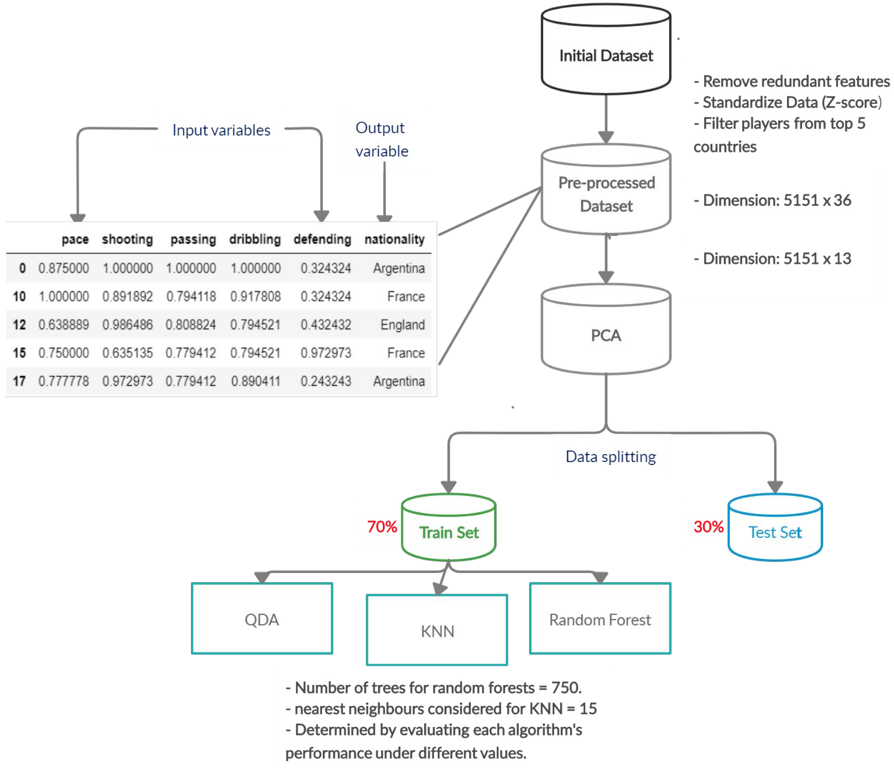
The Predictive Power of Random Forest, MSE=0.0069



Interpreting the Regression Coefficients

```
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                  3.097e+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
```

Predict Player Nationality



	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

Failed Approaches Tried for the Problem:

Approach 1:

- Create new features to represent attack, defense, tackle, mentality.
- Reduce dimensions using PCA on the defined feature matrix.

Approach 2:

- Use 'glm' to identify statistically

  Significant features: *heading accuracy, mentality, composure and mentality penalties*

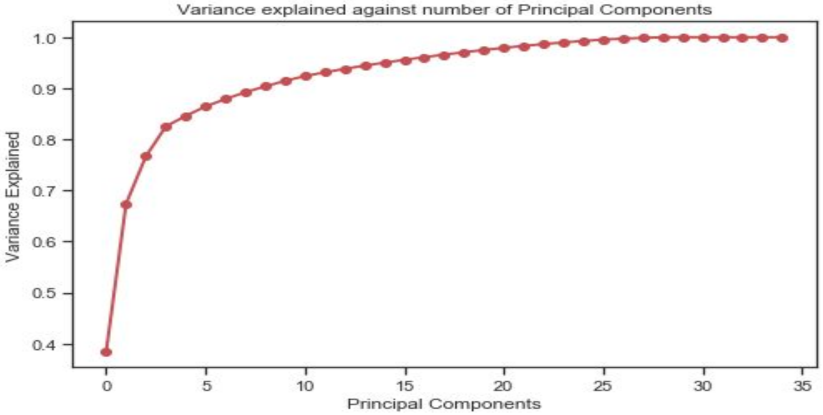
Approach 3:

- Use subset selection methods for feature selection

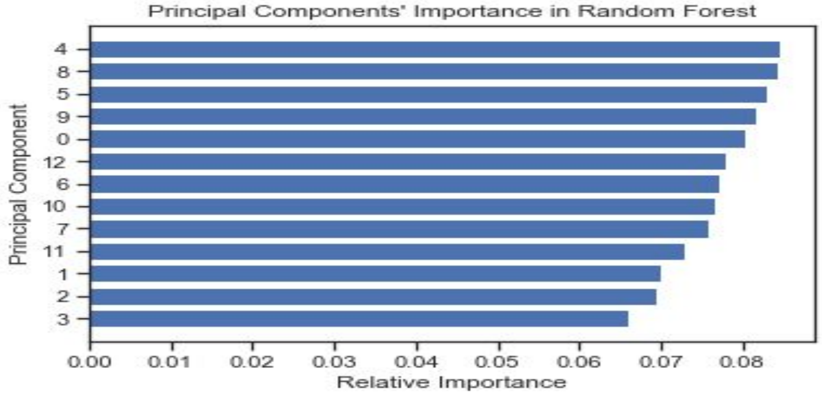
Approach 4:

- Use PCA to reduce dimensions and perform classification using principal components.

Analysis of Player Nationality Model



More than 90% of variance in the data explained by 13 components



No single component is impacting the model drastically

	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

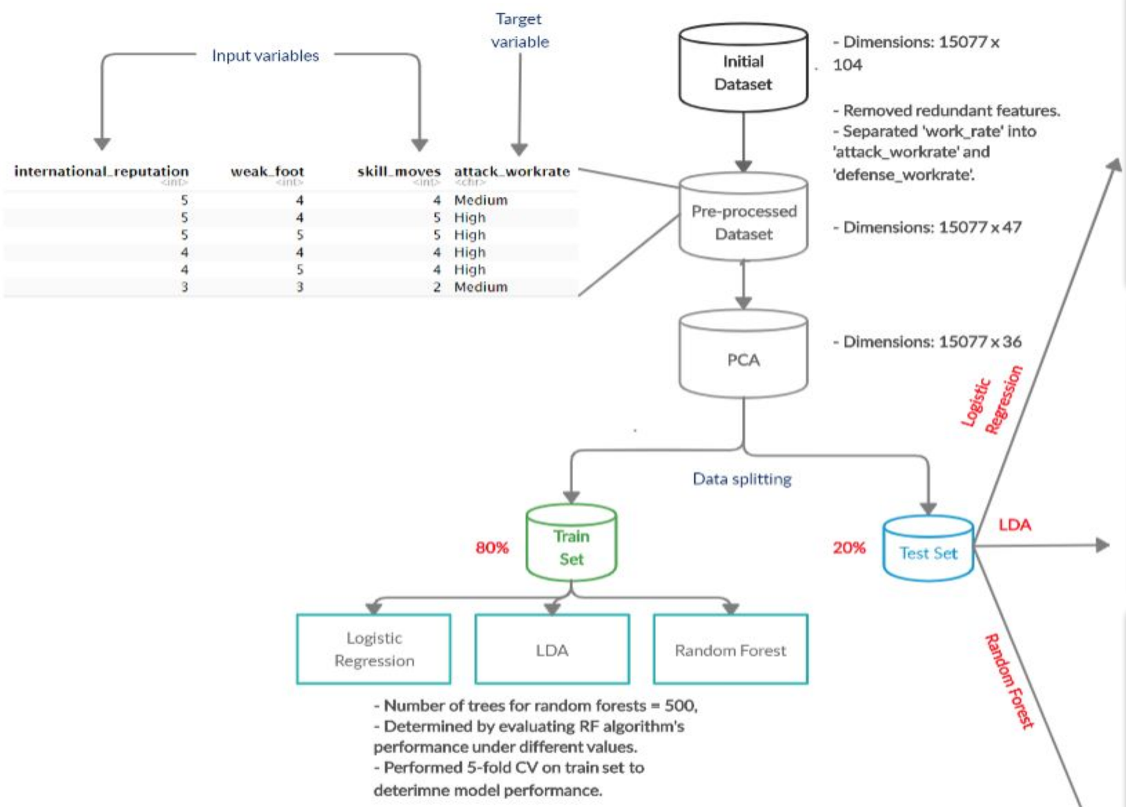
QDA Results

	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

Random Forest Results

- The results of models are almost similar with QDA performing slightly better than Random Forests and KNN classifiers.
- Problems:
    - 160 countries. Average no. of players/country = 114. Number of countries with no. of players < 200 = 134. **Solution** - top 5 countries with most no. of players.
    - Every country has good attackers, defenders, midfielders. So no feature particularly dominates classification of players for a country.
- **Noteworthy Result** - Predicting players from England is relatively accurate compared to other countries. Reason - no. of players from England = 1667

Predict Player Attack Work Rate



Process pipeline to classify player work rate

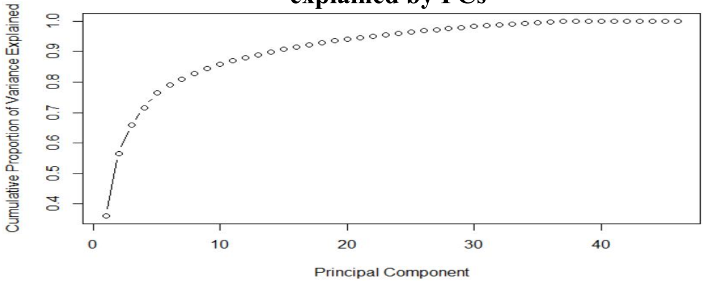
```
            Reference
Prediction High  Low Medium
    High    270    0    139
    Low       1    5     10
    Medium  632  163   1794

Overall Statistics

               Accuracy : 0.6865
                 95% CI : (0.6696, 0.703)
    No Information Rate : 0.6447
    P-Value [Acc > NIR] : 7.17e-07
```

```
            Reference
Prediction High  Low Medium
    High    374    0    239
    Low       4   47     92
    Medium  525  121   1612

Overall Statistics

               Accuracy : 0.6745
                 95% CI : (0.6575, 0.6912)
    No Information Rate : 0.6447
    P-Value [Acc > NIR] : 0.0003052
```

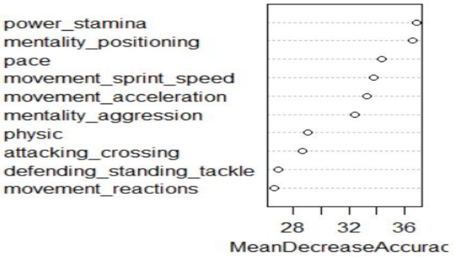
```
            Reference
Prediction High  Low Medium
    High    361    1    207
    Low       0    0      1
    Medium  542  167   1735

Overall Statistics

               Accuracy : 0.6954
                 95% CI : (0.6786, 0.7118)
    No Information Rate : 0.6447
    P-Value [Acc > NIR] : 2.165e-09
```

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:

- Predict player '**attack work-rate**' using various physical, mental and in-game attributes. Fit Logistic reg., LDA and Random Forest models on pre-processed train set (**5-fold CV**) to evaluate performance on test set.
- PCA on pre-processed dataset to reduce feature space to a 36 dimensional space (since, **~97%** of variance in target variable is explained by **36** features).
- Fit same classification models on PCA train set (**5-fold CV**) with 36 PCs and target variable to evaluate performance on test set.
- Determine AUC of ROC for the models on PCA test set.

Results:

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

Physical Attributes and Player Position

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

1. Consolidate various player positions under simplified positions:
   a. `Attacker (A), Midfielder (M), Defender (D), and Goalkeeper (G)`
2. Isolate features that reflect physical condition.
3. Split the data 80/20 into a training and testing set.
4. Trained multinomial logistic regression model (with goalkeepers as reference position).
5. The model coefficients offer 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

Coefficients:

	(Intercept)	age	height_cm	weight_kg
A	1.054069	-0.1284945891	-0.1022747	-0.1942253
D	6.141442	-0.0006128889	-0.1262502	-0.2170574
M	11.683797	-0.1170215674	-0.1304249	-0.2321517

	movement_acceleration	movement_sprint_speed
A	0.023907991	0.044004291
D	0.015081834	0.045635282
M	-0.005734023	-0.004713136

	movement_agility	movement_reactions	movement_balance
A	0.03146402	-0.2018269	0.01888024
D	-0.01175104	-0.1236048	0.02334179
M	0.04434541	-0.1766382	0.05112571

	power_shot_power	power_jumping	power_stamina
A	0.1668292	-0.02313521	0.1930654
D	0.1021718	-0.01272136	0.2688559
M	0.1423758	-0.06077378	0.2620672

	power_strength	power_long_shots
A	0.2346653	0.4582996
D	0.2476670	0.3371736
M	0.2219881	0.4364195

Results:

- Goalkeepers easiest to differentiate from others.
- Goalkeepers on average older, taller, and heavier (heavier because taller).
- By same token, `jumping` and `reaction` time dominated by goalkeepers.
- A, D, and M beat goalkeepers in `long shots` and `stamina`.
- Other categories more mixed and help differentiate between other player positions.

Physical Attributes and Player Position

<u>Part 2</u>: "What physical conditioning should trainers focus on for a player who is transitioning from one position to another?"

- Trained binomial logistic regression models for each pair of positions.
- Ignores difficult-to-change attributes of `age, height, and weight.`
- Focused on differences between A, M, and D. (Most potent differentiating factors for goalkeepers (G) evident in previous slide.)
- For transitioning from M $\longrightarrow$ A:
  - **Sprint speed** is the most strongly positively indicated attribute.
  - **Focus on:** acceleration, sprint speed, shot power, jumping, strength, and long shots
  - **Not Focus on:** agility, reactions, balance, or stamina.
- For transitioning from A $\longrightarrow$ M:
  - **Stamina** is the most strongly positively indicated attribute
- Similar interpretation holds for other pairs.

A vs M

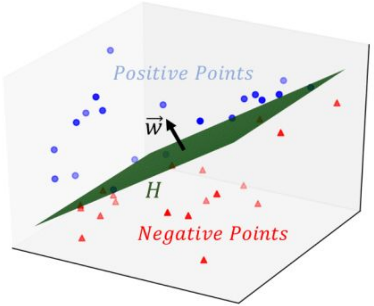
Coefficients:

	Values	Std. Err.
(Intercept)	-2.61190856	0.141007096
movement_acceleration	0.02331771	0.002460068
movement_sprint_speed	0.05083503	0.002335759
movement_agility	-0.01764774	0.001727855
movement_reactions	-0.02488424	0.001756298
movement_balance	-0.04517154	0.001682398
power_shot_power	0.03049375	0.001368221
power_jumping	0.03626619	0.001405170
power_stamina	-0.06612347	0.001446593
power_strength	0.02402223	0.001545503
power_long_shots	0.01891870	0.001530805

A vs D

Coefficients:

	Values	Std. Err.
(Intercept)	0.209328210	0.186285186
movement_acceleration	0.021491613	0.003245827
movement_sprint_speed	0.024129091	0.003069087
movement_agility	0.048319064	0.002293898
movement_reactions	-0.106281502	0.002320553
movement_balance	-0.031177398	0.002220961
power_shot_power	0.056267367	0.001797916
power_jumping	-0.013446511	0.001852726
power_stamina	-0.107832551	0.001919718
power_strength	-0.008442502	0.002048385
power_long_shots	0.131854920	0.002030723



$$H = \{\overline{x} : w_0 + \overline{w} \cdot \overline{x} = 0\}$$
$$w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_k x_k = 0$$
$$w_0 + w_1 y_1 + w_2 y_2 + \ldots + w_k y_k = 0$$
$$w_1(x_1 - y_1) + w_2(x_2 - y_2) + \ldots + w_k(x_k - y_k) = 0$$
$$\overline{w} \cdot (\overline{x} - \overline{y}) = 0$$
$$\overline{w} \perp (\overline{x} - \overline{y})$$

M vs D

Coefficients:

	Values	Std. Err.
(Intercept)	3.069354375	0.164844764
movement_acceleration	-0.012456560	0.002868091
movement_sprint_speed	-0.042971446	0.002727121
movement_agility	0.050342685	0.002021650
movement_reactions	-0.062215837	0.002061181
movement_balance	0.027403017	0.001966934
power_shot_power	0.036265427	0.001599297
power_jumping	-0.051797432	0.001637329
power_stamina	-0.006052985	0.001693145
power_strength	-0.043864764	0.001816785
power_long_shots	0.083850686	0.001792783