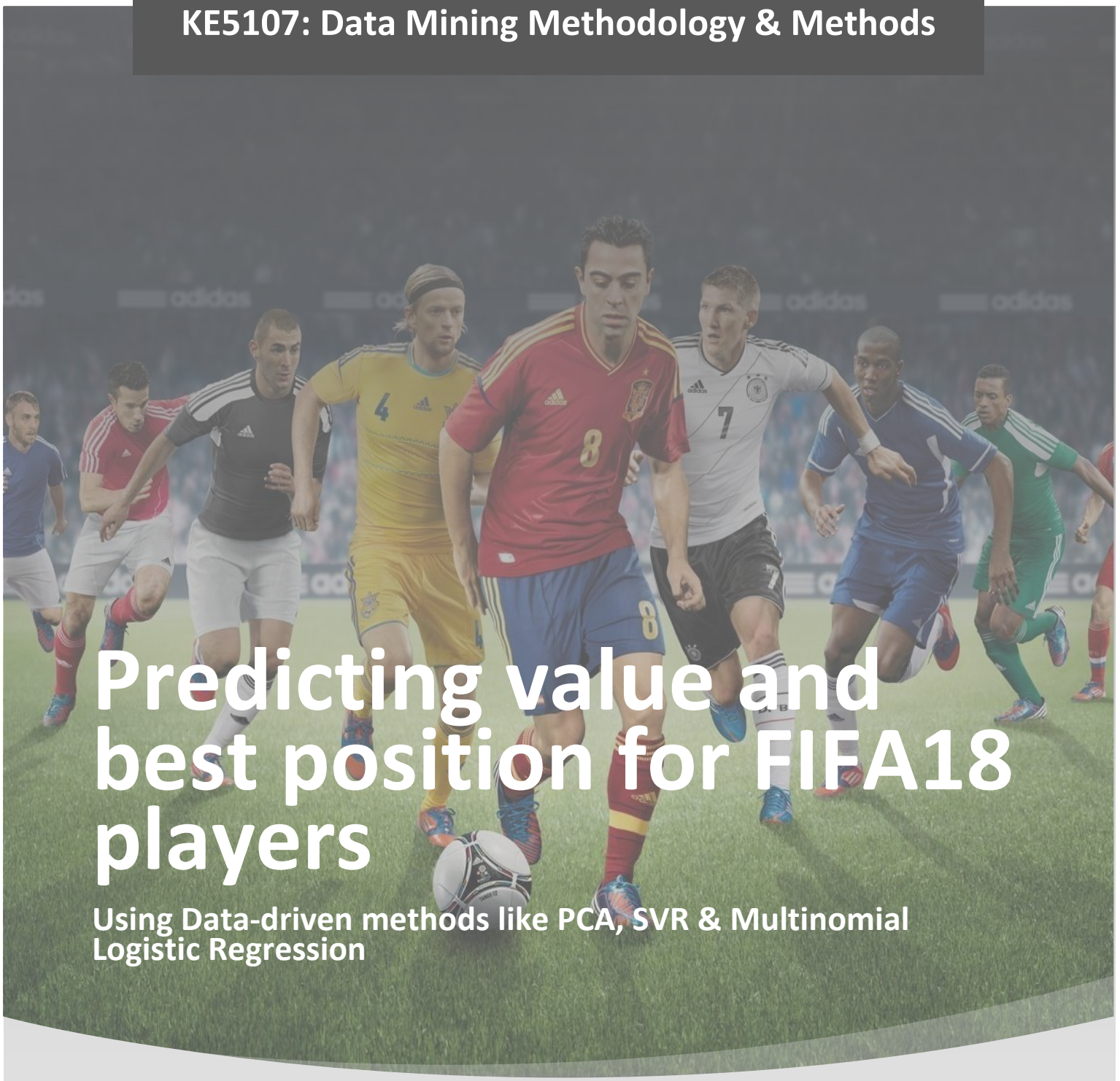
KE5107: Data Mining Methodology & Methods



Predicting value and best position for FIFA18 players

Using Data-driven methods like PCA, SVR & Multinomial Logistic Regression

SUBMITTED TO

Prof. Fan Zhen Zhen & Prof. Zhu Fang Ming

INSITITUTE OF SYSTEM SCIENCE

NATIONAL UNIVERSITY OF SINGAPORE

PREPARED BY

ABU MATHEW THOPPAN	A0178303H
BALAJI NATARAJ	A0178294N
BALAGOPAL UNNIKRISHNAN	A0178398E
PRATEEK KASHYAP	A0178248R
SAURABH SEMWAL	A0178339N
VIKNESHKUMAR BALAKRISHNAN	A0178304E

NUS National University of Singapore | iss INSTITUTE OF SYSTEMS SCIENCE

Table of Contents

TABLES

1. Executive Summary

- **In this project, we aim to predict the value (in euros) of a player and his role (or position) in the team.** This project will help team managers take informed decisions regarding the selection of players and bidding during transfer season.
- The project was planned and executed by following the **CRISP-DM** (Cross-industry standard process for data mining) model. CRISP-DM breaks the process of data mining into six major phases. The model is a cyclic process which allows for back and forth iteration between the six phases. The phases are: business understanding, data understanding, modelling, evaluation and deployment.
- We had 185 attributes in our dataset, for **feature selection** we used our domain knowledge, visual analysis, variance test, bi-directional elimination methods, and predictor screening algorithms to get the best set of features for fulfilling our two objectives. We also derived few columns to further reduce number of features.
- To further reduce dimensionality we used, **Principal Component Analysis (PCA)** which is a popular statistics technique used primarily in the field of data mining for dimensionality reduction. It uses an orthogonal transformation to reduce a set of possibly correlated variables in a multivariate dataset into a set of linearly combination of the original variables, referred to as **Principal Components**. After cleaning and transformation of our data and reducing the dimensions of the data with PCA, we created a **support vector regressor** with a polynomial kernel as our data was non-linear. We used **RMSE, R-Square and sMAPE** to calculate the prediction accuracy for our first objective.
- For second objective, we used **multinomial logistic regression** which calculated probabilities of a player for every position. From the probabilities of all the positions, it assigned the label having the highest probability. We used **confusion matrix, AUC** to estimate the performance of our model.
- We further derived some insights from our result, where we can further classify players into more positions, i.e. players who can play well **at multiple positions** (Attacking and Midfield positions, Midfield and Defending positions) using probability scores calculated by logistic classifier.

2. Background

2.1 Problem Description

Football is the most followed sport in the world. Its popularity can be mostly attributed to the many league tournaments that happen in many countries across the world. Such leagues allow players from different countries to play together as a team under a club. The clubs are usually owned by large conglomerate or sometimes even a single person.

The players are chosen by each club by bidding for them. Once a club gets a player, they come to terms on a contract which stipulates the player's salary and the duration of his contract with that club. Players can also be sold by one club to another during transfer season. This again takes place through a bidding process where the selling team can also choose to reject all bids if it isn't satisfied. This creates a problem for the buying club which has to decide upon an optimum bid based on the player's skills and the expectations of the selling club.

Each team has a manager whose role is similar to a head coach. The manager's responsibilities include:

1. Selecting the team of players for matches, and their formation
2. Planning the strategy, and instructing the players on the pitch
3. Motivating players before and during a match
4. Delegating duties to the first team coach and the coaching and medical staff
5. Scouting for young but talented players for eventual training in the youth academy or the reserves, and encouraging their development and improvement
6. Buying and selling players in the transfer market, including loans
7. Facing the media in pre-match and post-match interviews

Since the manager is responsible for the overall strategy of the team, there is lot of pressure on during transfer season to buy the best players suited for the team's strategy and also to offer the correct bid. If the bid is too low, the player might not be sold and if the bid is too high, the club unnecessarily spends extra money.

Another problem the manager faces is during the initiation of a new, inexperienced player. Since the manager has never seen the player play, he cannot say which is the best position for the player to play in. If the player plays at the wrong position, the whole team's strategy would be thrown of balance.

A system that can help the manager make the correct decisions on the bid amount and choosing the right players will be helpful during transfer season.

Figure 1: Various Playing Positions in Football

The various positions in football can be divided into four main types of players:

Forward – Striker (ST), Left Forward (LF), Centre Forward (CF), Right Forward (RF)

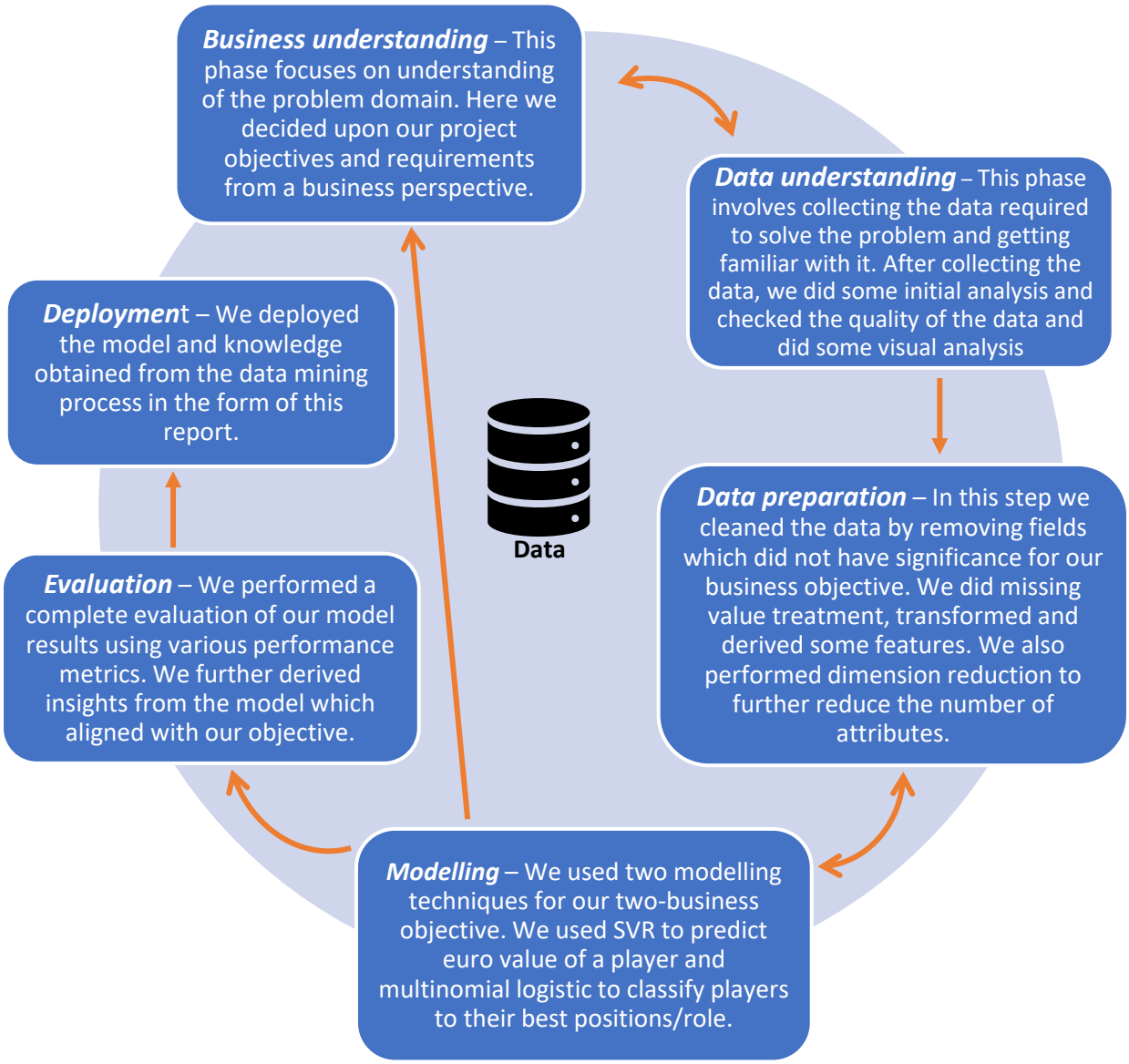
Midfield – Left Winger (LW), Left Midfielder (LM), Centre Attacking Midfielder (CAM), Centre Midfielder (CM), Centre Defending Midfielder (CDM), Right Winger (RW), Right Midfielder (RM)

Defence – Left Wing Back (LWB), Left Back (LB), Centre Back (CB), Right Wing Back (RWB), Right Back (RB)

Goalkeeping – Goalkeeper (GK)

2.2 Data mining process model

The project was planned and executed by following the **CRISP-DM** (Cross-industry standard process for data mining) model. CRISP-DM breaks the process of data mining into six major phases. The sequence is not strict and moving back and forth between the phases is always required. The model is also shown to be cyclic as the process of data mining can be repeated and improved upon based on the previous iteration.



3. Objectives

I. We aim to **predict monetary value of FIFA players** by identifying the main factors that affect the price of a player such as overall statistics, potential and performance metrics. This would enable team owners to take data driven decisions for estimating the correct value of the players and help them for the bidding process during transfer season.

II. Our aim is to **classify player to their ideal positions (or Roles)** on the football field. We will train our classifier to predict whether a player is suitable to play at attacking, defensive, midfield or goalkeeper's positions based on his performance statistics. This would enable coaches to decide best positions for a new, inexperienced player. This can also be used to check if experienced players are suitable to be played in alternate positions to their usual positions.

4. Assumptions

- A Goalkeeper cannot play well at any other position. So, the score of a goal keeper at other positions will be 0. Data has missing values for goal keeper on other positions so added a 0 score
- All the other players cannot play at Goalkeeper's position. So, the score of these players will be 0 at goalkeeper's position. Data has missing values for players on other positions so added a 0 score

5. Preliminary Data Analysis

5.1 Data Source

- The dataset, which is publicly available for research, is related to FIFA 2018
- Source: https://www.kaggle.com/kevinmh/fifa-18-more-complete-player-dataset

5.2 Data Description

- The dataset contains **186** variables and **17,994** observations

Index	Column #	Variable Name	Type	Description
A	1-5	Player description	various	Descriptive attributes of a player and his club such as ID, name, club name
B	6	special	continuous	A score describing specialty of a player.
C	7-16	Physical description	various	Physical attributes of a player, his country name. Ex Height. Weight.
D	17	euro value	continuous	Player's value in euros
E	18	euro wage	continuous	Player's wage in euros
F	19	euro release clause	continuous	Release clause of a player in euros
G	20	overall	continuous	Overall statistics of a player
H	21	potential	continuous	Player's potential index
I	22-27	General stats	continuous	Stats of overall in game performance such as shooting, passing
J	28	International reputation	continuous	International reputation

K	29	Skill moves	continuous	Skill moves score
L	30	Weak foot	continuous	Weak foot score, helps determining how well he plays with his weak foot.
M	31-32	Work Rate	continuous	work rate, describes how often he attends training sessions
N	33-62	Special stats	continuous	Scores of specifics such as crossing, finishing, head accuracy
O	63-67	Goal Keeper stats	continuous	Goalkeeper stats such as diving, handling
P	68-94	Position score	continuous	Performance of a player at various positions on a football pitch
Q	95-144	Trait indices	nominal	Indicator for trait value such as 1 on 1, diver, fancy flicks
R	145-158	Specialty indices	nominal	Indicator for Specialty traits. Ex: Speedster, distance shooter
S	159-185	Preferred position indices	nominal	Preferred positions
T	186	Player_Position	nominal	Players Playing Position

Table 1: Data Dictionary

5.3 Cleaning & Missing Value Treatment

- Dropping **Player description (A)** and **Physical attributes (C)** like player name and other details for hiding sensitive information as a data masking step.
- Removed column **"body_type"** in (C) because of low variance i.e. most of the players have similar body type. It was observed that body type didn't have any effect on the response.
- Removed **"eur_wage"** (E), **"eur_release_clause"** (F) because these can be further derived from our dependent variable (or response) and vice-versa.
- Removed attribute **"work_rate_def"** from **work rate**(M) as it had no effect on the outcome (based on EDA and backward elimination)
- Based on EDA and our Domain knowledge removed preferred foot attribute from **special stats** (N) as it has no effect on the value of a player.
- Removed weak foot attribute from **special stats** (N) based on EDA and including this attribute was reducing the performance of our model.
- Removed "**gk_positioning**" attribute from Goalkeeper stats(O) as including this attribute was reducing the performance of our model.
- Removed **Trait indices** (Q) as they have almost null variance i.e. almost all the players have either possessed or don't possess a trait.
- Removed **Specialty indices** (R) as they also have null variance i.e. except very few top players almost all the players have similar specialty index. Removed **Preferred position index**(S) as it was redundant with Position score(P)
- Removed 260 rows because they had 0 Euro Value, considering it as noise in our data. These rows were only 0.01% of our total data

Factors	Attribute	# of "missing"	Comments	Treatment
Position Scores(P)	All columns in (P) except "gk"	2022	All the goal keepers have missing values for all the position except gk I.e. goal keeper position	Added 0 as the score of a goalkeeper at all the other positions based on our assumption.
	"gk" only	15973	All the players except the goal keeper had missing value for goalkeeper position score.	Added 0 as the score of all the players except the goalkeeper at goalkeeper's position based on our assumption.

Table 2: Missing Value Treatment

5.4 Transformations

- To achieve our business objective of predicting Euro value of a player, we do not require specifics on the position performance of a player so; we derived new feature namely Attack, Mid, Def from position score(P) based on Domain knowledge

  Where;
  Attack = Sum of scores at attacking positions/number of attacking positions,
  Mid = Sum of scores at midfield positions/number of mid field positions,
  Def = Sum of scores at defending positions/number of defending position.

  So, we derived 3 features from 26 features in Position scores

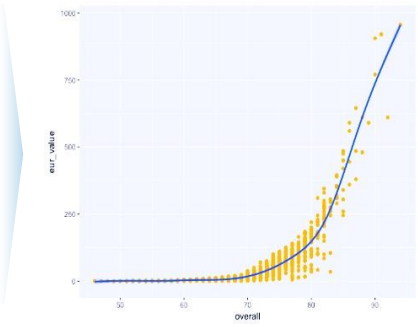
- For classifying players to the role as attackers, midfield players, defenders and goalkeeper we derived a column "**player_position**" which consists of the best position/role for a player

  "player_poition" → **Forward**; if Forward score > Scores at all other positions (defender, midfield, goalkeeper). Similarly, we assigned positions for midfield, defending and goalkeeping positions.

5.5 Inferences from Visual Analysis

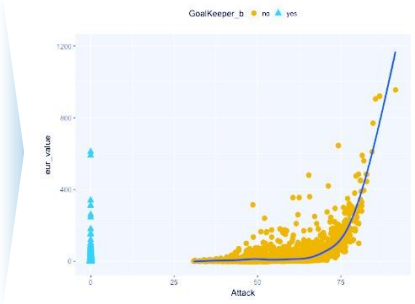
- **Overall stat vs Player Value**

We can refer from the plot that there is a non-linear relation between Overall stat and the Euro Value of the player. Player who have <= 65 score for overall stat, have very less Euro_value limit. Thus, this attribute should significantly affect the dependent variable, in our case, the Euro Value of the player.



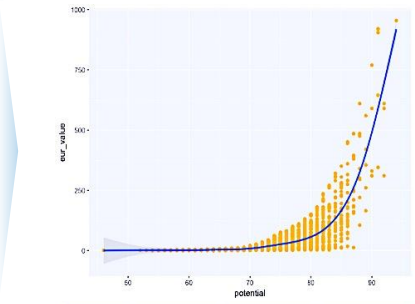
» **Attack vs Player Value**

Like overall stat, "Attack" attribute (which describes how well a player can play at attacking position) also seem to have a non-linear relationship with the value of a player. However, the value increases more rapidly as the "Attack" score increases. Interestingly, goal keepers have 0 attack value but not necessarily low Euro value since they don't play at attacking position



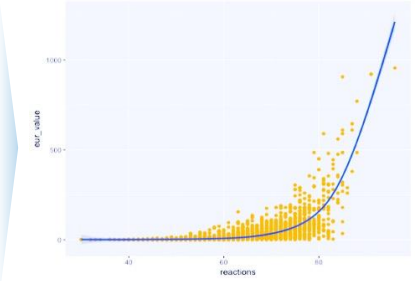
» **Potential vs Player Value**

Though the plot shows that as Potential of the player increases, his value also increases. However, this is true only for the upper limit. Higher potential does not guarantee higher player value, but higher range for the value. Hence, it is significant but in presence of other more significant attributes like Overall score



» **Reactions vs Player Value**

Reaction also plays an important role in determining the value of a player. Better the reactions of an attacker better are the chance of scoring a goal, similarly a goalkeeper with high reaction will save more goals. Again, reaction alone doesn't guarantee good outcomes.



NOTE: Added a column GK for identifying goal keepers on the plots. Euro value has been scaled by 55*10^5, so a point on zero doesn't mean that the player has 0-euro value. For the sake of simplicity, we have created plots on small (random) samples.

6. Predicting Player's Value

After data cleaning, transformations and feature selection we used **48 Features** as the input for the PCA

6.1 Dimensionality Reduction using PCA

» We could represent **~91%** of the data using **10 principal components**

» The first PCA represents **48%** of the data. From the loading matrix, we can infer that it is strongly correlated with the special stats of players (N) for example finishing, crossing etc. It is covering all the Attacking, Midfield, Defending as well as the Goalkeepers. It has a strong negative correlation with the attributes for a goal keeper

Number	Eigenvalue	Percent	20 40 60 80	Cum Percent
1	23.0892	48.102		48.102
2	7.8235	16.299		64.401
3	4.8416	10.087		74.488
4	2.3697	4.937		79.425
5	1.8887	3.935		83.360
6	1.0249	2.135		85.495
7	0.8105	1.689		87.184
8	0.7233	1.507		88.691
9	0.5539	1.154		89.845
10	0.4411	0.919		90.764

Figure 2: Eigen Values of Components

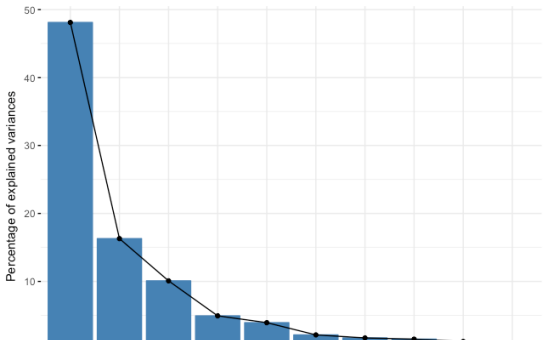


Figure 3: Scree Plot for PCA

» From the biplot we observed that goalkeeper stats have strong negative correlation with Dim1 and most of the special stats of attackers, midfield have strong positive correlation with it, general stats have less loading in Dim1. Dim2 is more correlated with the general stats of the players (except the goalkeepers)

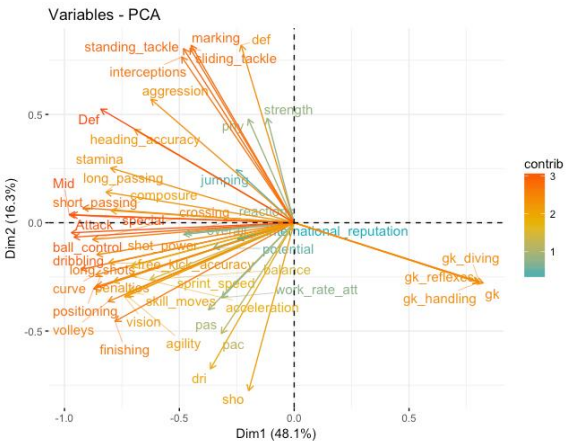


Figure 4: Bipolar Plot

6.2 Profiling of Components

Component	*Loaded On*	*Profile*
PC 1	curve, finishing, crossing, diving etc.	**Special Statistics of attacker, midfielders and Goalkeeper**
PC 2	standing tackle, sliding tackle etc.	**Defender Statistics**
PC 3	pass, dri, vision	**Passing Skills**
PC 4	acceleration, sprint speed, agility scores	**Players with High Agility**
PC 5	strength, phy scores	**Players with High Strength**
PC 6	potential score	**Player's Potential**
PC 7	jumping scores	**Jumping Skills**
PC 8	international reputation of a player	**International Reputation**
PC 9	working rate of a player	**Working rate**
PC 10	shot power	**Shot Power**

Table 3: Profiling of Components

Rotated Factor Loading

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10
work_rate_att	0.249409	-0.114413	0.138437	0.209995	-0.022786	0.015917	0.018547	0.014360	0.923941	0.004582
special	0.773812	0.409019	0.376441	0.219608	0.119171	0.063999	0.122554	0.055978	0.057097	0.044622
overall	0.210266	0.196280	0.641676	0.095499	0.368547	0.446064	0.137963	0.237221	-0.000862	0.012779
potential	0.145578	0.106773	0.350081	0.157933	0.055720	0.838965	0.008567	0.153586	0.018352	0.003117
pac	0.139265	-0.197422	0.235502	0.899725	-0.064704	0.109155	0.032115	0.012505	0.114262	0.055872
sho	0.261215	-0.564574	0.663287	0.014993	0.111515	0.079734	0.023213	0.068475	0.087810	0.314050
pas	0.180930	0.034393	0.944438	0.044918	-0.053512	0.027907	-0.084481	0.073614	0.051832	-0.012092
dri	0.230482	-0.267597	0.803751	0.303306	-0.087729	0.162546	0.004048	0.068676	0.101809	-0.007936
def	-0.119356	0.949425	0.099380	-0.042749	0.155489	0.055734	0.066695	0.040477	-0.040880	0.060935
phy	-0.045889	0.392878	0.211558	-0.006538	0.840177	0.036310	0.196253	0.053848	0.026948	0.010314
international_reputation	0.111262	0.062925	0.278795	-0.006382	0.102112	0.162738	0.043506	0.918989	0.014156	0.004642
skill_moves	0.776112	-0.103180	0.057439	0.213560	-0.124724	0.072424	-0.047325	0.110773	0.020060	-0.189616
crossing	0.743919	0.324532	0.307318	0.267135	-0.103693	-0.060197	-0.059560	0.032337	0.089821	-0.039458
finishing	0.886879	-0.245329	0.175786	0.119516	0.053922	0.071459	0.046158	0.014108	0.079194	0.115205
heading_accuracy	0.652636	0.398417	-0.241340	-0.081010	0.378598	0.209450	0.193004	0.055940	-0.015911	0.016658
short_passing	0.792328	0.391266	0.290926	0.070696	0.012131	0.115706	0.006297	0.019205	0.005381	-0.171311
volleys	0.873835	-0.122839	0.223380	0.088500	0.051107	0.043512	0.052502	0.081753	0.045064	0.156708
dribbling	0.883089	0.117372	0.190131	0.277031	-0.069018	0.083971	-0.011766	0.005191	0.078258	-0.091506
curve	0.817677	0.131533	0.353197	0.147448	-0.080008	-0.032624	-0.030270	0.060278	0.054461	0.099999
free_kick_accuracy	0.775519	0.169513	0.372384	0.036180	-0.077822	-0.085438	-0.041460	0.063076	0.017929	0.182906
long_passing	0.645049	0.497563	0.410535	0.006930	-0.030223	0.026836	-0.043988	0.007690	-0.018811	-0.166035
ball_control	0.886299	0.234806	0.193751	0.169946	0.021311	0.140920	0.024912	0.022514	0.036994	-0.116634
acceleration	0.584260	0.067942	0.003331	0.754053	-0.101829	0.058653	0.068756	-0.009764	0.076539	-0.021587
sprint_speed	0.575048	0.080373	-0.032425	0.764938	-0.008949	0.077853	0.056284	-0.005995	0.076451	-0.001486
agility	0.617363	0.026774	0.233223	0.541353	-0.225587	-0.034064	0.185655	-0.000921	0.065106	-0.131703
reactions	0.231134	0.190532	0.618690	0.043364	0.310502	0.319002	0.188822	0.208531	0.004875	-0.027560
balance	0.545961	0.097451	0.141244	0.424029	-0.473285	-0.075236	0.253769	-0.009698	0.041678	-0.103899
shot_power	0.862801	0.084044	0.202497	0.051284	0.159409	0.068925	0.054496	0.027478	0.040732	0.191658
jumping	0.072078	0.219054	0.002085	0.143101	0.182620	0.033859	0.903671	0.044038	0.015703	0.005935
stamina	0.612351	0.452510	0.042878	0.324281	0.257048	-0.034047	0.151060	-0.046454	0.103472	-0.096367
strength	0.020737	0.245684	-0.039036	-0.142554	0.893923	0.032010	0.045613	0.046077	-0.038599	-0.004023
long_shots	0.876102	0.001721	0.306089	0.083969	0.055291	0.013110	0.032092	0.003847	0.050220	0.182815
aggression	0.412074	0.654287	0.034875	-0.009383	0.389395	-0.007537	0.177638	0.037167	0.027269	-0.051171
interceptions	0.187901	0.932090	0.043709	-0.018249	0.132351	0.024535	0.064776	0.030459	-0.028218	-0.010818
positioning	0.888473	-0.033284	0.196910	0.198416	0.037905	0.023084	0.038296	0.013938	0.125852	-0.009436
vision	0.673684	0.024394	0.582476	0.054430	-0.042777	-0.010858	-0.025625	0.032205	0.052495	-0.152168
penalties	0.867952	-0.072787	0.129365	0.025047	0.043890	0.057782	0.067578	0.083751	0.027974	0.202601
composure	0.662811	0.276865	0.317123	0.019770	0.222699	0.183751	0.134373	0.167076	0.001113	-0.161603
marking	0.160703	0.957004	-0.058117	-0.017748	0.099441	0.037413	0.038334	0.002015	-0.031466	-0.001797
standing_tackle	0.197363	0.956242	-0.045923	-0.020610	0.089613	0.036387	0.023670	0.016717	-0.023777	-0.004630
sliding_tackle	0.162633	0.962491	-0.058966	0.001722	0.053308	0.028134	0.027730	0.013939	-0.017994	-0.000392
gk_diving	-0.830359	-0.364089	0.350413	-0.095059	0.015504	0.000628	-0.023272	0.028253	-0.002876	0.125005
gk_handling	-0.826567	-0.365919	0.353929	-0.098961	0.021246	-0.000878	-0.026616	0.026952	-0.001711	0.121326
gk_reflexes	-0.829406	-0.364999	0.352643	-0.095100	0.017997	0.001243	-0.023962	0.026103	-0.004559	0.124181
Attack	0.948134	0.201500	0.032697	0.193072	0.040796	0.073021	0.042546	0.009368	0.051442	-0.055374
Mid	0.918365	0.312040	0.068701	0.174567	0.010698	0.059852	0.027437	0.004211	0.040401	-0.104182
Def	0.657607	0.723900	-0.072647	0.095326	0.106591	0.053821	0.058806	0.009330	0.002698	-0.081336
gk	-0.842597	-0.373730	0.343913	-0.101322	0.009863	0.000541	-0.030540	0.033515	-0.001717	0.118645

Table 4: Rotated Factor Loadings

6.3 Support Vector Regression

- After cleaning and transformation of our data and reducing the dimensions of the data with PCA we split the data into training and test, where training consisted of 75% of randomly selected observations and test with 25% of the examples.
- Since our problem statement here is predicting the euro value of a player and the relationship between our label and attributes have a nonlinear relationship, we have opted to use a nonlinear regression method using SVM. We used polynomial kernel function in SVR(with degree=3). Our input features were the 10 PCA which were representing 91% of the data.

6.4 Performance Metrics

- **Root Mean Square Error**: we used **RMSE** as a performance metrics as It indicates the absolute fit of the model to the data and indicates, how close the observed data points are to the model's predicted values It is also used to check overfitting of our regressor. If RMSE

for Test is very high as compared to the RMSE for training set, then it can be inferred that the model has over fitted the data and will give bad predictions for new observations.

RMSE		
	Training	0.3429
	Test	0.3529

» **R-Square:** R-Square provides an estimate of the strength of the relationship between our model and the response variable. High R-Square value indicates that the percentage of the response variable variation that is explained by our model is good.

R-Square	88.29

» **SMAPE (Symmetric mean absolute percent error)**: It is another method for calculating prediction accuracy. It is like MAPE the only difference is a symmetric MAPE takes absolute values of actual and forecasted values for calculations and doesn't do a bias penalization i.e. high penalty for negative errors and low penalty for positive.

- $smape = \frac{2}{N}\sum_{k=1}^{N}\frac{|F_k - A_k|}{F_k + A_k}$ ***SMAPE*** 0.49

6.5 Conclusion

» **Principal Component Analysis**: Since the data had high number of dimensions, PCA helped us to reduce the complexity and facilitated us to understand correlation between various attributes in the data. We can also safely conclude how PCA captures variance across the attributes without much information loss and hence doesn't hamper the performance of the model (trade-off between performance and reducing the dimensions is very small).

» **Modelling:** We conclude that euro value of a player depends upon his stats, his potential, International reputation and his role in the team. Attackers have the highest Euro values amongst all other roles. Followed by midfielders and goalkeepers. Defenders seem to have the least euro value amongst other players

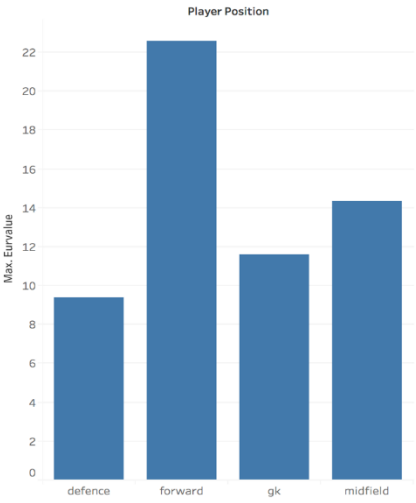


Figure 5: Eur_Value vs Player Position

7. Classifying Players in Playing Position

7.1 Feature Selection

» *From **Bidirectional elimination, p value significance and visual analysis***

- For this objective, we did not remove any attributes in the initial stage as we planned to see significance of each features using their p values and eliminating them if they do not add any value to our objective.
- After running multinomial logistic regression on the baseline model, we calculated the p values for all the features. We considered the features with p values < 0.063 as only those had significance for our objective. We verified this by running the model with and without

these attributes and observed that the accuracy improved with the removal of less significant features. We also visually analyzed the features to see their impact on our goal. Here are some of the observations:

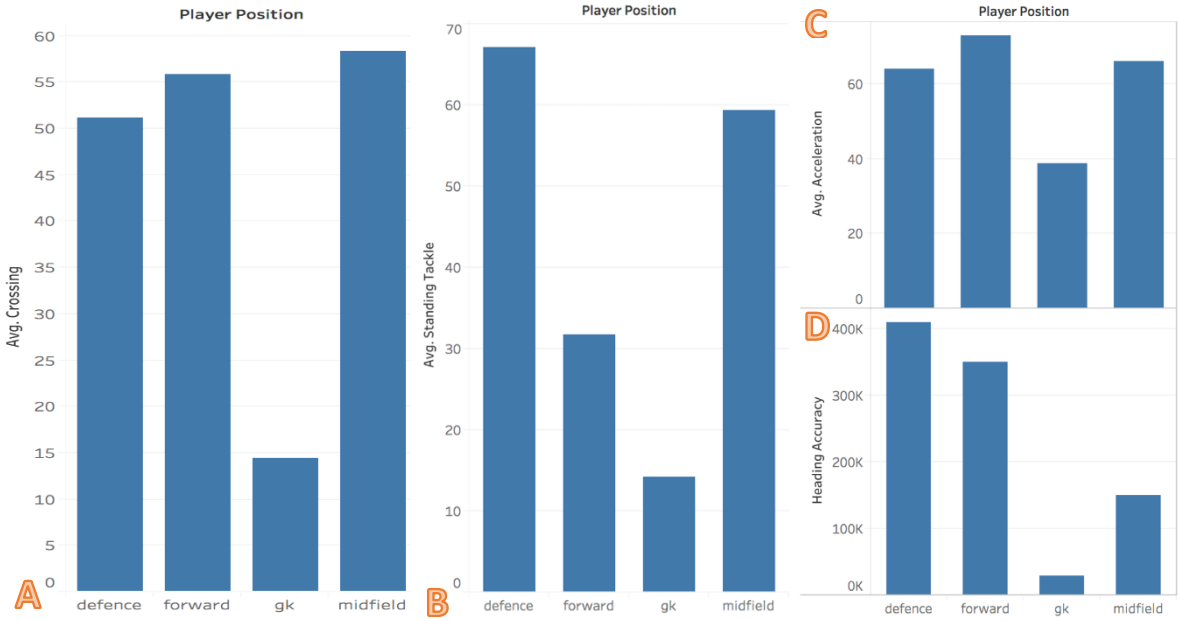


Figure 6: Player Position vs A) Avg. Crossing B) Avg. Standing Tackle C) Acceleration D) Heading Accuracy

No.	Source	LogWorth	PValue
1	crossing	252.503	0.00000
2	overall	150.639	0.00000
3	finishing	65.271	0.00000
4	short_passing	57.908	0.00000
5	long_passing	48.617	0.00000
6	vision	47.755	0.00000
7	stamina	44.133	0.00000
8	heading_accuracy	36.438	0.00000
9	dribbling	27.923	0.00000
10	weak_foot	26.493	0.00000
11	marking	24.352	0.00000
12	acceleration	6.507	0.00000
13	shot_power	2.080	0.00832

No.	Source	LogWorth	PValue
14	interceptions	21.206	0.00000
15	ball_control	21.105	0.00000
16	reactions	18.935	0.00000
17	positioning	18.480	0.00000
18	sprint_speed	16.259	0.00000
19	potential	15.461	0.00000
20	standing_tackle	11.824	0.00000
21	strength	11.384	0.00000
22	aggression	10.579	0.00000
23	long_shots	8.883	0.00000
24	sliding_tackle	8.729	0.00000
25	jumping	8.052	0.00000
26	agility	1.199	0.06330

Table 5: Selected Features for Regression

7.2 Multinomial Logistic Regression

- Our objective here was to classify players as Forward, Midfield players, Defenders and Goal keepers
- We used the derived column "**player_position**" (*refer: 5.4 Transformations*) as the response variable here and used the selected features (*refer: 7.1 Feature Selection*) as the predictors.
- We split our data into test and train set with split ratio=**0.8**
- Multinomial logistic regression calculates probabilities of a player for each of the four positions. From these four probabilities, it assigns the label having the highest probability

Player ID	P(Defence)	P(forward)	P(goalkeeper)	P(midfield)	Label
12722	0.21	0.6	0.04	0.15	*Forward*
5661	0	0.84	0.01	0.15	*Forward*
5225	0.85	0	0	0.15	*Defender*
11515	0.85	0	0	0.15	*Defender*
19	0	0	1	0	*Goal Keeper*
15360	0	0.16	0	0.84	*Mid-Fielder*

7.3 Performance Metrics

» **Confusion matrix**: Confusion matrix gives us the accuracy of our classification and helps us calculate the sensitivity and specificity for each of the class

Predicted position	defence	forward	gk	midfield
defence	**5,360**	15	1	72
forward	5	**5,010**	0	59
gk	0	0	**1,600**	0
midfield	68	41	0	**2,164**

Table 6: Confusion Matrix on Train Dataset

Predicted position	defence	forward	gk	midfield
defence	**1,348**	1	0	10
forward	1	**1,262**	0	14
gk	1	0	**385**	0
midfield	14	13	0	**550**

Table 6: Confusion Matrix on Test Dataset

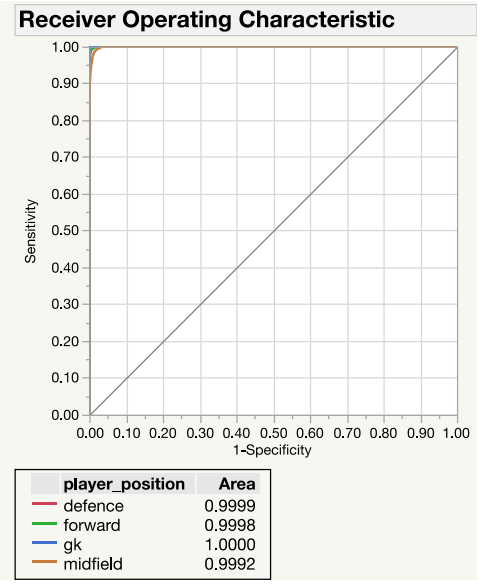


Figure 7: ROC for Various Positions

Player Position	Accuracy		Misclassification Rate		Sensitivity		Specificity	
	Train	Test	Train	Test	Train	Test	Train	Test
Defence	98.3	99.19	1.1	0.7	98.6	98.8	99	99.5
Forward	98.73	98.82	0.8	0.8	98.9	98.9	99.3	99.3
gk	100	99.74	0.006	0.02	99.9	100	100	99.9
midfield	95.2	95.32	1.6	1.4	94.3	95.8	99.1	99.1
Overall	98.19	98.49	1.6	1.7	-	-	-	-

Table 8: Key Performance Indices on Train and Test Dataset

7.4 Conclusion

- We got very high accuracy using just 26 attributes (out of 185) as the scores of a player's general and special stat played an important role in identifying the role of a player. For instance, a player with high score for sliding tackle will most likely be classified as a defender. Similarly, a player with high score for finishing (scoring a goal) attribute will be an attacker(forward).
- We also observed that the model misclassified some midfielder because some mid fielders have characteristics of both defenders and attackers(forward).
- *In our additional insights and further scope, we will discuss how we can further improve the classification by creating two more classes for players who can play at multiple positions.*

8. Additional Insights & Further Scope

- Most of the players play well only at one of the positions i.e. either they are forward (attackers), defenders or midfielders. But, from our domain knowledge, we know that there are some players who play well at multiple positions, so we tried verifying this hypothesis from our modelling results.
- From the predicted probabilities for each class (forward, defense etc.) for a player, we can assign multiple roles to a player or in other words we can derive two more classes;

  A) **Forward and midfielders**: These players can play well at both forward and midfield positions. So, depending on a game's scenario these players can be assigned either forward or mid positions.

  B) **Midfielders and Defenders**: These players can play well at both midfielders and defenders positions. So, depending on a game's scenario these players can be assigned either midfield or defensive positions.

  We calculated P(F-M) = P(F) - P(M) and P(M-D) = P(M)-P(D) from the probability scores of the players and observed that:

  - (only) Forward positions — 1 >= P(F) >= 0.7
  - (only) Midfield positions — 1 >= P(M) >= 0.7
  - (only) Defending positions — 1 >= P(D) >= 0.7
  - Forward and Midfield — 0.2 >= P(F-M) >= -0.2 & 0.6>= P(M-D) >= 0.4
  - Midfield and Defensive — -0.6 >= P(F-M)>= -0.4 & 0.2 >= P(M-D) >= -0.2

Note: A Goal keeper can only play at one position "gk"

P(Defence)	P(Forward)	P(Midfield)	P(F-M)	P(M-D)	Positions
0.8	0	0.2	-0.2	-0.6	only Defending
0.5	0	0.5	-0.5	0	Midfield or Defending
0	0.8	0.2	0.6	0.2	only Forward
0	0.55	0.45	0.1	0.45	Forward or Midfield
0.1	0.1	0.8	-0.7	0.7	only Midfield

Table 7: Predicting Additional Positions

8. References

1. Dataset: https://www.kaggle.com/kevinmh/fifa-18-more-complete-player-dataset
2. Crisp DM : By Kenneth Jensen - Own work based on: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/18.0/en/ModelerCRISPDM.pdf (Figure 1), CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=24930610
3. Player Position: https://www.fifauteam.com/fifa-18-position-change-cards-guide/
4. Manager's responsibilities: https://en.wikipedia.org/wiki/Manager_(association_football)
5. SMAPE: http://www.vanguardsw.com/business-forecasting-101/symmetric-mean-absolute-percent-error-smape/
6. PCA: https://en.wikipedia.org/wiki/Principal_component_analysis
7. SVR: http://www.saedsayad.com/support_vector_machine_reg.htm
8. Multinomial Logit: https://en.wikipedia.org/wiki/Multinomial_logistic_regression
9. Specificity and sensitivity: https://en.wikipedia.org/wiki/Sensitivity_and_specificity