Prescriptive Analytics Challenge Report

Business Case : FIFA

Daviano Almeyda Kustaryanto - 3212610

Contents

Introduction

Nowadays, Data has become the crucial part in professional world. Many company used data in order to make decision in the future therefore they won't be making any wrong decision. Data also can be used to predict something is likely to happen in the future. Based on this statement, we believe that data can also predict which player who have high potential to be a star in the future since many club is willing to invest millions of euro in order to buy a talented player to increase their team value or to be sell into bigger club in the future. Therefore we believe that predicting player based on data is more effective rather than using traditional way.

In Business Proposal Document, we already talked about the general information of the project that we are going to do that cover purpose, benefit and goals. Meanwhile, this document is going will be explained how we do our analysis, what method that we are going to used and the result of the analysis.

Assumption

First, we are going to assume that the data that we get from Kaggle is from an official data and we get the data from FIFA itself. Otherwise, this entire project will be illegal since we're using data from a third-party that the source of the data is not very credible or possible a stolen data, hacked data and etc. We make an assumption that data that we get from Kaggle is legal and free to use since for us Data Ethics and Laws are very important.

Second, we are going to assume that the data that we get from Kaggle is based on real life observation from FIFA. Therefore, if there any mis value from the data that we get from Kaggle and the official data from FIFA, it's not our fault.

Third, We are assuming that all professional football club is already a data-driven company, therefore predicting player based on data is already pretty common. Because, up until now many professional club is still using traditional way to recruit or scout a player by looking their performance directly.

Fourth, We are assuming that any result that we get from machine learning that we made, our clients will consider to recruit it and try to make their player fulfil their potential.

Data

We assumed that the dataset is originally from FIFA that we retrieved from Kaggle. The dataset contains the attributes and profile of a player such as Name, Date of Birth, Speed and etc. The data set is contains 89 Columns and 18.159 rows however we're not going to used all of them. The Rough data will look like the picture Below :

Number	ID	Name	Age	Photo	Nationality	Flag	Overall	Potential	Club
0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	https://cdn.sofifa.org/flags/52.png	94	94	FC Barcelona
1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	https://cdn.sofifa.org/flags/38.png	94	94	Juventus
2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	https://cdn.sofifa.org/flags/54.png	92	93	Paris Saint-Germain
3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	https://cdn.sofifa.org/flags/45.png	91	93	Manchester United
4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	92	Manchester City
5	183277	E. Hazard	27	https://cdn.sofifa.org/players/4/19/183277.png	Belgium	https://cdn.sofifa.org/flags/7.png	91	91	Chelsea
6	177003	L. Modrić	32	https://cdn.sofifa.org/players/4/19/177003.png	Croatia	https://cdn.sofifa.org/flags/10.png	91	91	Real Madrid
7	176580	L. Suárez	31	https://cdn.sofifa.org/players/4/19/176580.png	Uruguay	https://cdn.sofifa.org/flags/60.png	91	91	FC Barcelona
8	155862	Sergio Ramos	32	https://cdn.sofifa.org/players/4/19/155862.png	Spain	https://cdn.sofifa.org/flags/45.png	91	91	Real Madrid
9	200389	J. Oblak	25	https://cdn.sofifa.org/players/4/19/200389.png	Slovenia	https://cdn.sofifa.org/flags/44.png	90	93	Atlético Madrid

Club Logo	Value	Wage	Special	Preferred Foot	International Reputation	Weak Foot	Skill Moves	Work Rate	Body Type	Real Face
https://cdn.sofifa.org/teams/2/light/241.png	€110.5M	€565K	2202	Left	5	4	4	Medium/ Medium	Messi	Yes
https://cdn.sofifa.org/teams/2/light/45.png	€77M	€405K	2228	Right	5	4	5	High/ Low	C. Ronaldo	Yes
https://cdn.sofifa.org/teams/2/light/73.png	€118.5M	€290K	2143	Right	5	5	5	High/ Medium	Neymar	Yes
https://cdn.sofifa.org/teams/2/light/11.png	€72M	€260K	1471	Right	4	3	1	Medium/ Medium	Lean	Yes
https://cdn.sofifa.org/teams/2/light/10.png	€102M	€355K	2281	Right	4	5	4	High/ High	Normal	Yes
https://cdn.sofifa.org/teams/2/light/5.png	€93M	€340K	2142	Right	4	4	4	High/ Medium	Normal	Yes
https://cdn.sofifa.org/teams/2/light/243.png	€67M	€420K	2280	Right	4	4	4	High/ High	Lean	Yes
https://cdn.sofifa.org/teams/2/light/241.png	€80M	€455K	2346	Right	5	4	3	High/ Medium	Normal	Yes
https://cdn.sofifa.org/teams/2/light/243.png	€51M	€380K	2201	Right	4	3	3	High/ Medium	Normal	Yes

LAM	CAM	RAM	LM	LCM	CM	RCM	RM	LWB	LDM	CDM	RDM	RWB	LB	LCB	CB	RCB	RB	Crossing	Finishing	HeadingAccuracy	ShortPassing	Volleys
93+2	93+2	93+2	91+2	84+2	84+2	84+2	91+2	64+2	61+2	61+2	61+2	64+2	59+2	47+2	47+2	47+2	59+2	84	95	70	90	86
88+3	88+3	88+3	88+3	81+3	81+3	81+3	88+3	65+3	61+3	61+3	61+3	65+3	61+3	53+3	53+3	53+3	61+3	84	94	89	81	87
89+3	89+3	89+3	88+3	81+3	81+3	81+3	88+3	65+3	60+3	60+3	60+3	65+3	60+3	47+3	47+3	47+3	60+3	79	87	62	84	84
																		17	13	21	50	13
88+3	88+3	88+3	88+3	87+3	87+3	87+3	88+3	77+3	77+3	77+3	77+3	77+3	73+3	66+3	66+3	66+3	73+3	93	82	55	92	82
89+3	89+3	89+3	89+3	82+3	82+3	82+3	89+3	66+3	63+3	63+3	63+3	66+3	60+3	49+3	49+3	49+3	60+3	81	84	61	89	80
87+3	87+3	87+3	86+3	88+3	88+3	88+3	86+3	82+3	81+3	81+3	81+3	82+3	79+3	71+3	71+3	71+3	79+3	86	72	55	93	76
85+5	85+5	85+5	84+5	79+5	79+5	79+5	84+5	69+5	68+5	68+5	68+5	69+5	66+5	63+5	63+5	63+5	66+5	77	93	77	82	88

Position	Jersey Number	Joined	Loaned From	Contract Valid Until	Height	Weight	LS	ST	RS	LW	LF	CF	RF	RW	LAM	CAM	RAM	LM
RF	10	01/07/2004		2021	5'7	159lbs	88+2	88+2	88+2	92+2	93+2	93+2	93+2	92+2	93+2	93+2	93+2	91+2
ST	7	10/07/2018		2022	6'2	183lbs	91+3	91+3	91+3	89+3	90+3	90+3	90+3	89+3	88+3	88+3	88+3	88+3
LW	10	03/08/2017		2022	5'9	150lbs	84+3	84+3	84+3	89+3	89+3	89+3	89+3	89+3	89+3	89+3	89+3	88+3
GK	1	01/07/2011		2020	6'4	168lbs												
RCM	7	30/08/2015		2023	5'11	154lbs	82+3	82+3	82+3	87+3	87+3	87+3	87+3	87+3	88+3	88+3	88+3	88+3
LF	10	01/07/2012		2020	5'8	163lbs	83+3	83+3	83+3	89+3	88+3	88+3	88+3	89+3	89+3	89+3	89+3	89+3
RCM	10	01/08/2012		2020	5'8	146lbs	77+3	77+3	77+3	85+3	84+3	84+3	84+3	85+3	87+3	87+3	87+3	86+3
RS	9	11/07/2014		2021	6'0	190lbs	87+5	87+5	87+5	86+5	87+5	87+5	87+5	86+5	85+5	85+5	85+5	84+5
RCB	15	01/08/2005		2020	6'0	181lbs	73+3	73+3	73+3	70+3	71+3	71+3	71+3	70+3	71+3	71+3	71+3	72+3

Interceptions	Positioning	Vision	Penalties	Composure	Marking	StandingTackle	SlidingTackle	GKDiving	GKHandling	GKKicking	GKPositioning	GKReflexes	Release Clause
22	94	94	75	96	33	28	26	6	11	15	14	8	€226.5M
29	95	82	85	95	28	31	23	7	11	15	14	11	€127.1M
36	89	87	81	94	27	24	33	9	9	15	15	11	€228.1M
30	12	68	40	68	15	21	13	90	85	87	88	94	€138.6M
61	87	94	79	88	68	58	51	15	13	5	10	13	€196.4M
41	87	89	86	91	34	27	22	11	12	6	8	8	€172.1M
83	79	92	82	84	60	76	73	13	9	7	14	9	€137.4M
41	92	84	85	85	62	45	38	27	25	31	33	37	€164M
90	60	63	75	82	87	92	91	11	8	9	7	11	€104.6M

Volleys	Dribbling	Curve	FKAccuracy	LongPassing	BallControl	Acceleration	SprintSpeed	Agility	Reactions	Balance	ShotPower	Jumping	Stamina	Strength	LongShots
86	97	93	94	87	96	91	86	91	95	95	85	68	72	59	94
87	88	81	76	77	94	89	91	87	96	70	95	95	88	79	93
84	96	88	87	78	95	94	90	96	94	84	80	61	81	49	82
13	18	21	19	51	42	57	58	60	90	43	31	67	43	64	12
82	86	85	83	91	91	78	76	79	91	77	91	63	90	75	91
80	95	83	79	83	94	94	88	95	90	94	82	56	83	66	80
76	90	85	78	88	93	80	72	93	90	94	79	68	89	58	82
88	87	86	84	64	90	86	75	82	92	83	86	69	90	83	85
66	63	74	72	77	84	76	75	78	85	66	79	93	84	83	59

Cleaning Data

Removing and Changing Columns

In this section we're going to explain how we cleaned the dataset. Since the datasets contains many columns and thousands rows of data that we don't need to do this project.

```
#Removing Unecessary Column
df1.drop(['Joined','Contract Valid Until',
        "Photo","Flag","Club Logo","Special","ID",'Preferred Foot','Real Face',
        'Jersey Number','Loaned From','LS','ST','RS','LW','LF','CF',
        'RF','RW','LAM','CAM','RAM','LM', 'LB','LCB','CB','RCB',
        'RB','LCM','CM','RCM','RM','LWB','LDM','CDM','RDM','RWB',
          'StandingTackle','SlidingTackle','GKDiving','GKHandling','GKKicking','GKPositioning','Number','Release_Clause'  ],
        axis=1, inplace=True)
```

After we remove some unnecessary column, we are going to change the value and wage into numeric since the value from the dataset is messy.

```
## Converting to numeric Format

def convert_to_numeric(df_value):
    try:
        value = float(df_value[1:-1])
        suffix = df_value[-1:]
        if suffix == 'M':
            value = value * 1000000
        elif suffix == 'K':
            value = value * 1000
    except ValueError:
        value = 0
    return value
```

```
## Appliying The Function

df1['Value'] = df1['Value'].apply(convert_to_numeric)
df1['Wage'] = df1['Wage'].apply(convert_to_numeric)
df1.Value = df1.Value.replace(0, np.nan)
df1.Wage = df1.Wage.replace(0, np.nan)
df1.head(3)
```

On the dataset, we also found a anomaly column that called work rates, in order to make it more suitable we separated into two different column called Attacking Rates and defensive rates.

In [8]:
```
##Separating Work Rate
df1['Work_Rate'] = df1['Work_Rate'].astype(str)
df1['Work_Rate'] = df1['Work_Rate'].str.split('/')
df1['Attacking_rates']=df1['Work_Rate'].str.get(0)
df1['Attacking_rates'] = df1['Attacking_rates'].str.strip()
df1['Defensive_rates'] = df1['Work_Rate'].str.get(1)
df1['Defensive_rates'] = df1['Defensive_rates'].str.strip()
```

The last thing in this section, we are going to fix the body type in this datasets because we found some miss input in the datasets.

In [10]:
```
#Fixing Body Type
df1.loc[df['Body Type'] == "Lean", 'Body Type'] = 1
df1.loc[df['Body Type'] == "Normal", 'Body Type'] = 2
df1.loc[df['Body Type'] == "Stocky", 'Body Type'] = 3
df1.loc[df['Body Type'] == "Messi", 'Body Type'] = 2
df1.loc[df['Body Type'] == "C. Ronaldo", 'Body Type'] = 2
df1.loc[df['Body Type'] == "Neymar", 'Body Type'] = 1
df1.loc[df['Body Type'] == "Courtois", 'Body Type'] = 2
df1.loc[df['Body Type'] == "Shaqiri", 'Body Type'] = 3
df1.loc[df['Body Type'] == "Akinfenwa", 'Body Type'] = 3
df1.loc[df['Body Type'] == "PLAYER_BODY_TYPE_25", 'Body Type'] = np.nan
```

In [11]:
```
##Droping Body Type that have "nan" values and changing Body Type column into int
df1.dropna(subset=['Body Type'], inplace=True)
df1['Body Type'].astype('int64')
```

Analysis Report

In order to complete this project, we are going to used some machine learning modules. Below is the type of machine learning that we used :

1. KNN Algorithm(K Nearest Neighbours)

The K-Nearest Algorithm is a supervised classification algorithm, It takes a lot marked points and use them to learn how to label another point. The reason we used this technique is this technique is quite common among data scientist, It is very simple to implemented and many major company used this technique to predict Something.

2. Euclidian Model

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2}$$

Euclidian is a method for machine learning where they spotted a common divisor of a number, where the largest number can divided without leaving a remainder. Euclidian method is pretty common when we used KNN as a method.

3. Decision Tree Classifier Accuracy

Decision Tree is a machine learning method where they break down a data set into smaller subset. We also use Decision Tree Classifier to test our accuracy score therefore we can make a comparison between accuracy score with decision tree and with KNN.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{f_{11} + f_{00}}{f_{11} + f_{10} + f_{01} + f_{00}}. \quad (4.1)$$

4. F1 Score

F1 Score is a tool to measure the accuracy of a test. F1 Score is often used in Information Retrieval department for measuring search, document classification, and query classification.

$$F_1 = \left(\frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} \right) = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

5. Decision Tree Regressor

R2 is a statistical Method who have a purposes to either predict a future of the outcomes or testing the hypothesis based on the machine learning. We try this method to test our hypothesis about the data

R-squared (R2) Coefficient of Determination

$$\text{R-Squared} = 1 - \frac{\text{First Sum of Errors}}{\text{Second Sum of Errors}}$$

Result

K Nearest Algorithm

The KNN algorithm was successfully applied in our machine learning. The algorithm show the potential overall of a player. Picture of proof are shown below :

```
Player Potential number 1 :
Predict Player potential : [88]
Player Potential number 2 :
Predict Player potential : [78]
Player Potential number 3 :
Predict Player potential : [88]
Player Potential number 4 :
Predict Player potential : [87]
Player Potential number 5 :
Predict Player potential : [75]
```

To get this result, we predict the potential Based on the age and current overall of the player and test it with dummy data. For Example : Player number 1 have overall 75 and the age is 17 so the potential he will received in the future is 88.

As a side goals of this project is to predict attacking rates of a player, we also going to shown the prediction as proof below :

```
Player  number 1 :
Predict Player Attacking_rates : [3]
Player  number 3 :
Predict Player  Attacking_rates : [2]
Player  number 3 :
Predict Player  Attacking_rates : [2]
```

This algorithm is going to predict the attacking rates of a player, because attacking rates determine the energy of player to do an attacking move towards the enemies. The number on the picture above stands for High, Medium and Low where 3 equals to High, 2 equals to Medium and 1 equals to Low. This calculation is based on Finishing, Shot Power and Heading Accuracy. For instance : Player 1 have 85 Finishing, 76 Shot Power and 83 on Heading accuracy so he will get the High attacking rates.

Accuracy Score and F1 Score of Potential Player and Attacking Rates

```
F1 Score
0.1293916023993145
Accuracy Score
0.1293916023993145
```

The F1 and Accuracy score that we applied for predicting potential player is both have 12,9 % accuracy score, which is quite low for an accuracy test. However predicting a player based on data is quite a challenge since there is also external factor that determined the potential of a player.

```
F1 Score
0.7377892030848329
Accuracy Score
0.7377892030848329
```

The F1 and Accuracy score that we used on Predicting attacking rates is quite goof where both of them have 73,7 % accuracy.

As you can see above the, F1 score determined that the prediction near perfection so the higher the score the higher chance of the prediction is coming true, meanwhile Accuracy score determined that that the prediction is nearly accurate therefore the higher the score the higher also it will happening in the future.

Decision Tree Classifier Accuracy For Potential Player Accuracy

```
              accuracy                           0.13      1167
             macro avg       0.11      0.13      0.11      1167
          weighted avg       0.11      0.13      0.11      1167
```

To make sure that we had the same result we going to used Decision tree accuracy test, which our prediction using this machine learning resulted 13% which is quite similar with The K Nearest Algorithm.

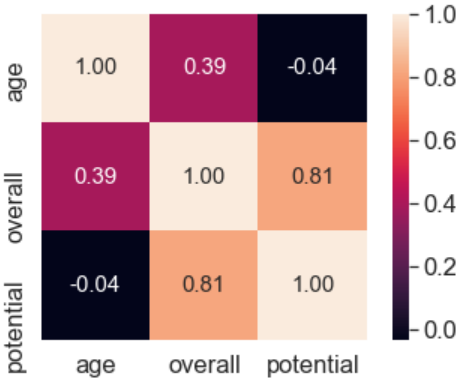
Decision Tree Regressor For Predicting Potential Player

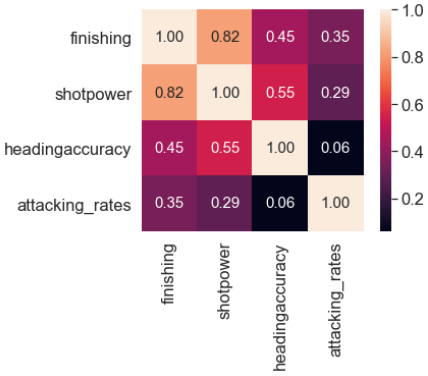
For this section we are going to measure if there is correlation between the variable that we choose to predict the Player potential using R2.

```
Mean Absolute Error: 2.0717596975390955
Root Mean Squared Error: 2.5957667499968777
R2 accuracy 0.8175975696727882
```

The R2 calculation in this machine learning is 81,7% which is quite high where in the KNN and Decision Tree classifier the accuracy is rather low. We also calculated the MAE (Mean Absolute Error) where we get the calculation 2.07 which is quite good where the mean of the potential is 73 so MEA of the data is less than 10% of the actual mean.

Conclusion





Based on these two diagram, I can make a conclusion that our machine learning is working perfectly since the result between KNN and Decision Tree have similar accuracy score. However the result that we get for Potential accuracy is quite disappointing since the accuracy is quite low. On the other hand our predictions of Attacking Rates show a good result since the accuracy looks as we expected.

Based on this conclusion, we want give a recommendation for FIFA to add few column as a factor to predict potential such as sport intelligence and Awareness therefore, it will help us to give them more accurate score.

Appendix

Our complete data set will look like this :

Number	17790 non	null int64
ID	17790 non	null int64
Name	17790 non	null object
Age	17790 non	null int64
Photo	17790 non	null object
Nationality	17790 non	null object
Flag	17790 non	null object
Overall	17790 non	null int64
Potential	17790 non	null int64
Club	17557 non	null object
Club Logo	17790 non	null object
Value	17790 non	null object
Wage	17790 non	null object
Special	17790 non	null int64
Acceleration	17790 non	null int64
Aggression	17790 non	null int64
Agility	17790 non	null int64
Balance	17790 non	null int64
BallControl	17790 non	null int64
Body Type	17790 non	null object
CAM	15806 non	null object
CB	15806 non	null object
CDM	15806 non	null object
CF	15806 non	null object
CM	15806 non	null object
Composure	17790 non	null int64
Contract Valid Until	15800 non	null object
Crossing	17790 non	null int64
Curve	17790 non	null int64
Dribbling	17790 non	null int64
FKAccuracy	17790 non	null int64
Finishing	17790 non	null int64
GKDiving	17790 non	null int64
GKHandling	17790 non	null int64
GKKicking	17790 non	null int64
GKPositioning	17790 non	null int64
GKReflexes	17790 non	null int64
HeadingAccuracy	17790 non	null int64
Interceptions	17790 non	null int64
International Reputation	15811 non	null float64
Jersey Number	15800 non	null float64
Joined	14629 non	null object

Jumping	17790 non	null int64
LAM	15806 non	null object
LB	15806 non	null object
LCB	15806 non	null object
LCM	15806 non	null object
LDM	15806 non	null object
LF	15806 non	null object
LM	15806 non	null object
LS	15806 non	null object
LW	15806 non	null object
LWB	15806 non	null object
Loaned From	1258 non	null object
LongPassing	17790 non	null int64
LongShots	17790 non	null int64
Marking	17790 non	null int64
Penalties	17790 non	null int64
Position	17778 non	null object
Positioning	17790 non	null int64
Preferred Foot	17790 non	null object
RAM	15806 non	null object
RB	15806 non	null object
RCB	15806 non	null object
RCM	15806 non	null object
RDM	15806 non	null object
RF	15806 non	null object
RM	15806 non	null object
RS	15806 non	null object
RW	15806 non	null object
RWB	15806 non	null object
Reactions	17790 non	null int64
Real Face	17790 non	null object
Release Clause	16292 non	null object
ST	15806 non	null object
ShortPassing	17790 non	null int64
ShotPower	17790 non	null int64
Skill Moves	17790 non	null int64
SlidingTackle	17790 non	null int64
SprintSpeed	17790 non	null int64
Stamina	17790 non	null int64
StandingTackle	17790 non	null int64
Strength	17790 non	null int64
Vision	17790 non	null int64
Volleys	17790 non	null int64
Weak Foot	17790 non	null int64
Work Rate	17790 non	null object