FIFA 21 – Players Analysis & Prediction

Vikas Ranjan

Date: 03/03/2021

DSC 550

Bellevue University, Bellevue

Abstract:

FIFA (Fédération Internationale de Football Association) is the governing body of the association football. FIFA periodically releases players statistics. These statistics are very important and depicts every aspect of the players performance.

As part of this case study, I will be applying data mining and visualization techniques learnt during this course. With graph analysis, various trends and patterns related to players and their performance will be identified. Based on the various players performance data, I'll try to predict if a player will be able to have an overall score of greater than 80. The intent is also to find which all features are most important in order to make that prediction.

Player's performance data is used by various soccer clubs to sign up new players or buy players contracts from other clubs. This prediction model can be used by club managers and recruiters to identify new players and predict if they will have high overall scores.

Dataset Details

Source - https://www.kaggle.com/ekrembayar/fifa-21-complete-player-datasets

This data set has players data for 2021 season and contains 107 columns and 17125 rows. The data describes players attributes like name, height, weight, age, club association, nationality, playing position, wage, contract details, and various performance data.

Describe Data

	ID	Age	OVA	BOV	POT
count	17125.000000	17125.000000	17125.000000	17125.000000	17125.000000
mean	219388.716204	25.272934	66.965022	67.900204	72.489810
std	37499.197507	4.942665	6.864329	6.637538	5.769949
min	2.000000	16.000000	38.000000	42.000000	47.000000
25%	204082.000000	21.000000	62.000000	64.000000	69.000000
50%	228961.000000	25.000000	67.000000	68.000000	72.000000
75%	243911.000000	29.000000	72.000000	72.000000	76.000000
max	259105.000000	53.000000	93.000000	93.000000	95.000000

	Total Stats	Dominant_Right_Foot
count	17125.000000	17125.000000
mean	1631.256175	0.753635
std	260.357024	0.430906
min	731.000000	0.000000
25%	1492.000000	1.000000
50%	1659.000000	1.000000
75%	1812.000000	1.000000
max	2316.000000	1.000000

Summarized Data

	Name	Nationality	Club	BP	Height	Weight	foot	Value
count	17125	17125	17102	17125	17125	17125	17125	17125
unique	16176	167	917	15	21	57	2	216
top	J. Rodríguez	England	Chelsea	CB	6'0"	154lbs	Right	€1.1M
freq	10	1707	45	3252	2583	1342	12906	500

	Wage	W/F	SM	A/W	D/W	IR	Hits	Position
count	17125	17125	17125	17036	17036	17125	17125	17125
unique	142	5	5	3	3	5	593	4
top	€2K	3 ★	2★	Medium	Medium	1 ★	3	Midfielder
freq	2453	10567	7524	11044	12225	15136	3253	7229

Objectives:

As the objective of this case study, I was looking towards answering below questions, identifying patterns and building prediction model as listed below:

- Check the distribution of Age, Overall Averge, Total Stats and potential.
- Compare players count on categories of playing position, international reputation, Left foot vs right foot and Skills moves.
- Apply pearson corelation on the Age, Overall Averge, Total Stats and potential to see how these features are correlated.
- Find and plot agewise Player distribution in FIFA 21 season.
- Identify and plot nation wise players distribution
- Identify and visualize comparison of left foot vs right foot on categories of playing postion, skills move, international reputation and weak foot.
- Then train and test data to predict if the player would have a dominant right foot or left foot.
- Perform feature selection using sklearn's VarianceThreshold with threshold of 0.5.
- Determine which are top 10 features which would be best to predict if a player is right foot dominant or left foot dominant.
- Conduct feature selection using SelectKBest.
- Perform model evaluation and selection (I performed it among RandomForestClassifier, DecisionTreeClassifier, SGDClassifier and LogisticRegression).

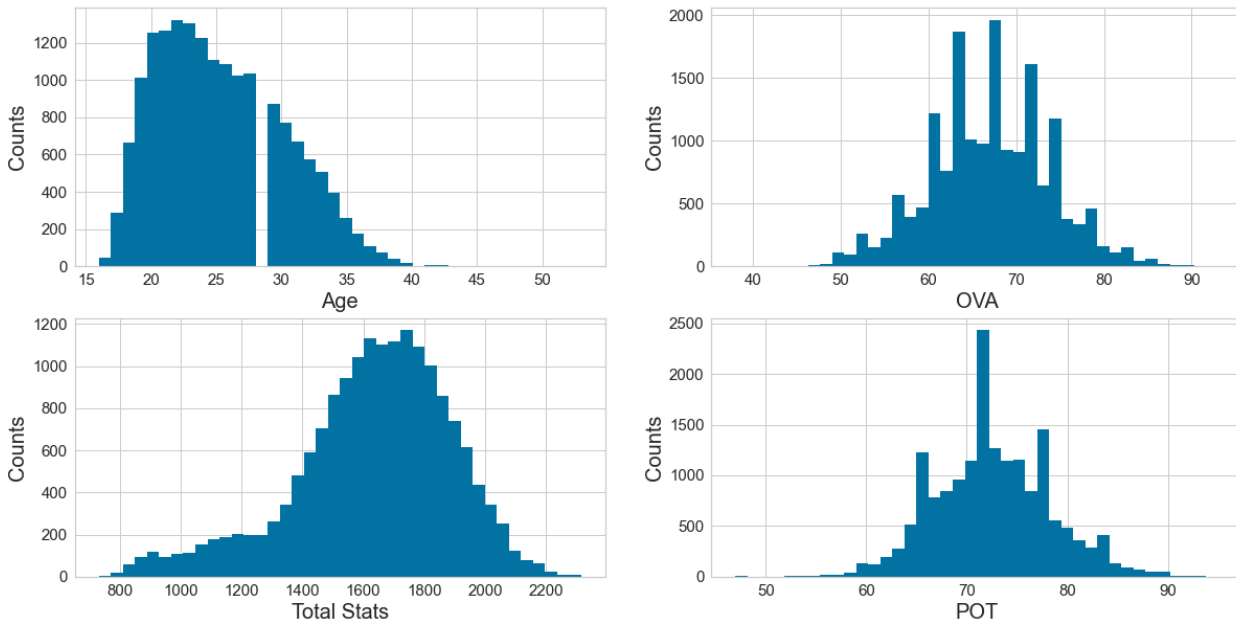
Steps:

Part 1: Data Load, Cleanup & Graph Analysis

- Load the data and perform data cleanup.
- Prepare graph & charts - I've applied some visualization techniques to demonstrate a few key insights in terms of trends/patterns. The idea is to perform initial analysis and understand dataset with graphs and charts.

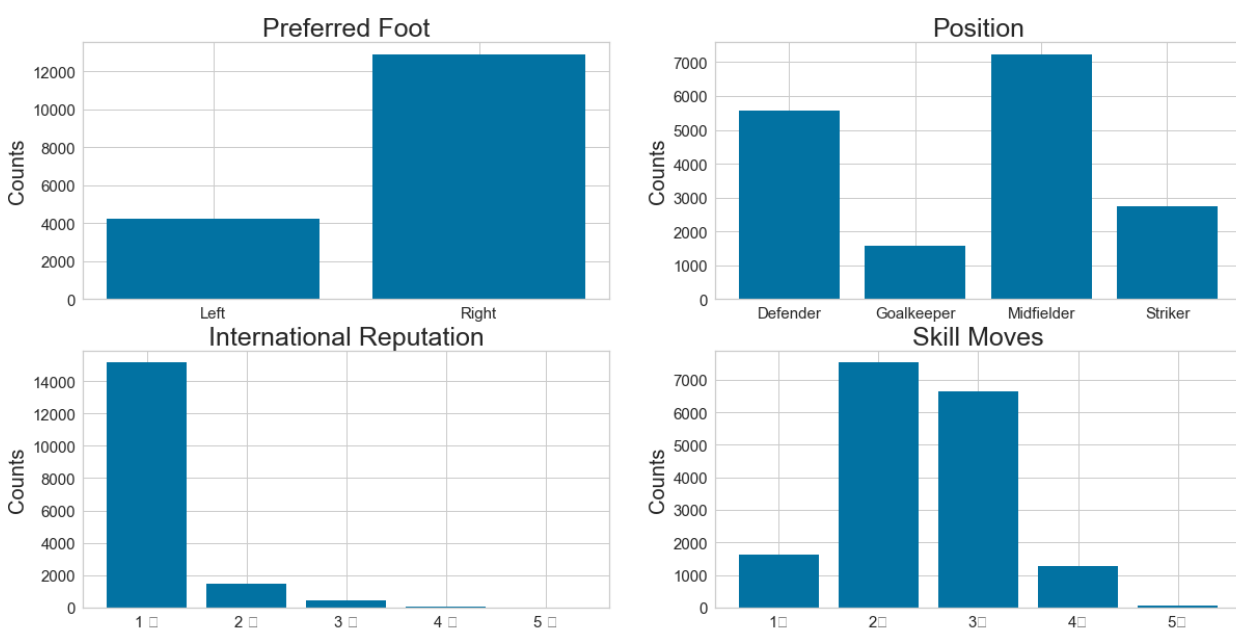
Histograms:

Below set of histograms show the distribution of players age, overall score (OVA), total stats, and potential (POT). It shows very trends such as maximum players are within age range of 20-24.

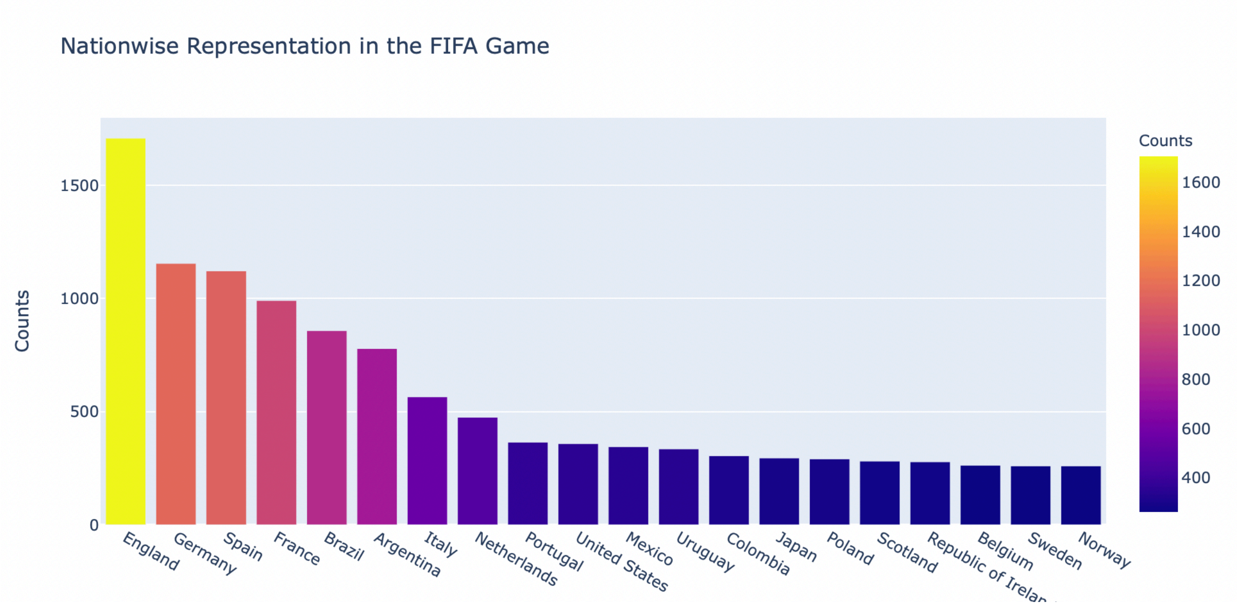


Barcharts:

I did convert playing positions into categorical variable and converted them into 4 categories of Defender, Goalkeeper, Midfielder and Striker. Below set of bar charts do show the comparison of counts for left foot vs right foot, Playing positions, International reputation, and Skills moves.
It is very evident here that a large percentage of players are right foot dominant and there are large number of midfielders compared to other playing positions. International reputation of 3*/4* are not so common among players and very few have 5 * rating.

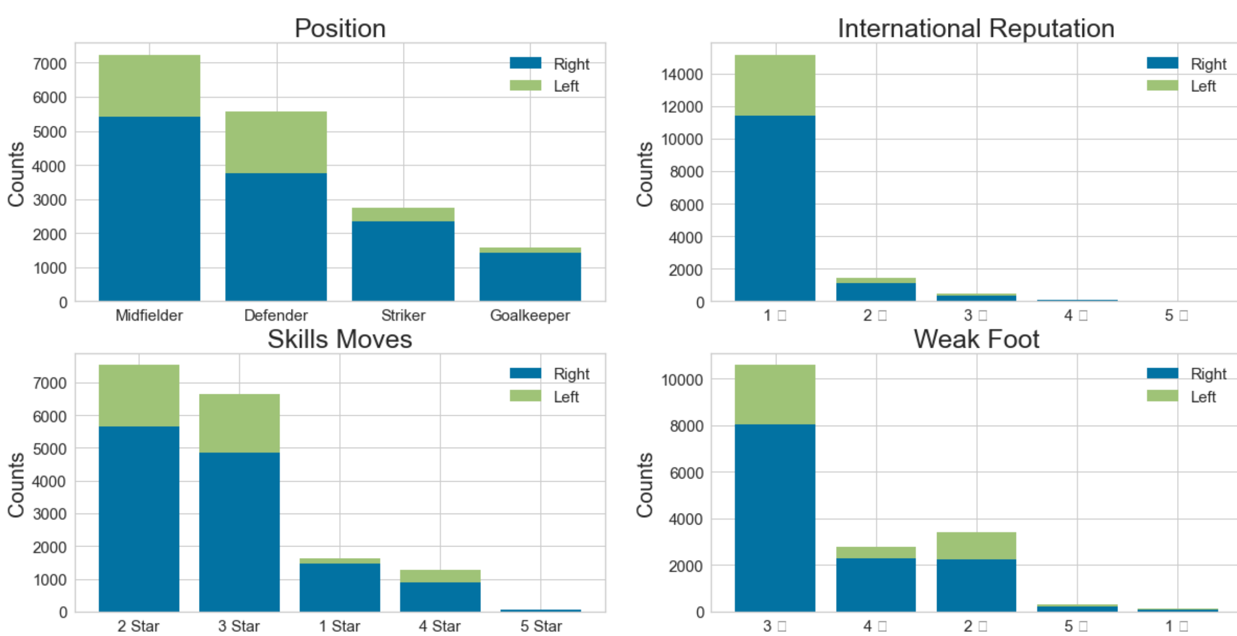


Below chart shows nation wise count of players and UK has the highest number of players.



Stacked Bar Charts:

Below stacked bar charts show how left foot vs right foot players are distributed across features of playing position, skills moves (SM), International reputation (IR). As we can see, in each of these categories, there is a large percentage of right foot dominant players compared to left foot dominant players.



Part 2: Feature Reduction

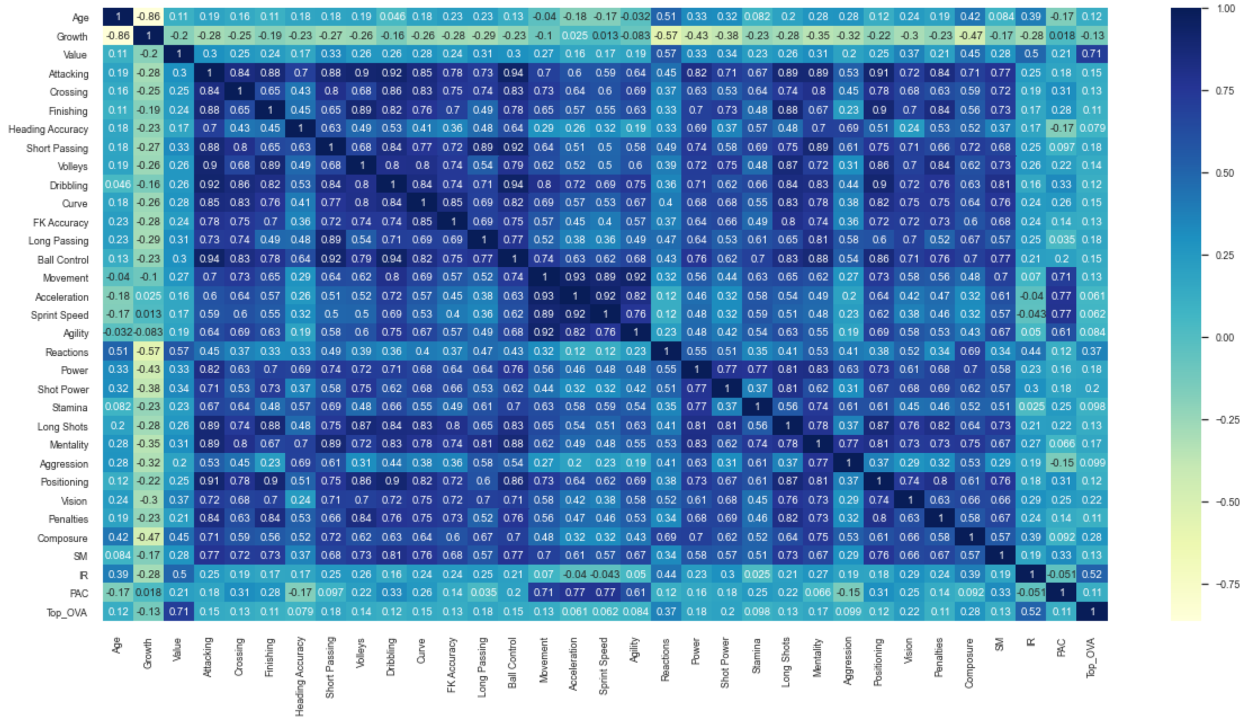
My original dataset consists of data/stats of FIFA 2021 player's and the dataset has 17125 rows and 107 columns. I've done following steps for feature reduction:

- Drop the column with no significance.
- Format data to consistent format such as M(millions) and K(thousands) to money* 1000000 or money*1000, height to inches, etc.
- Drop all the text variables which can't be translated into numbers.
- Verify if any column has nulls. In my dataset, I had 300 null values for Composure. Therefore, applied median to missing values for Composure.
- After performing these steps, my reduced data frame has 13334 rows and 32 columns.
- Then train and test data to predict if the player would have a great overall score (> 80).
- Perform feature selection using sklearn's VarianceThreshold with threshold of 0.5. Variance threshold is calculated based on probability density function of a particular distribution. The values with True are the features selected using Variance threshold technique and since all the features are showing True, therefore none of the columns need to be removed.
- Conducted Feature selection using SelectKBest.

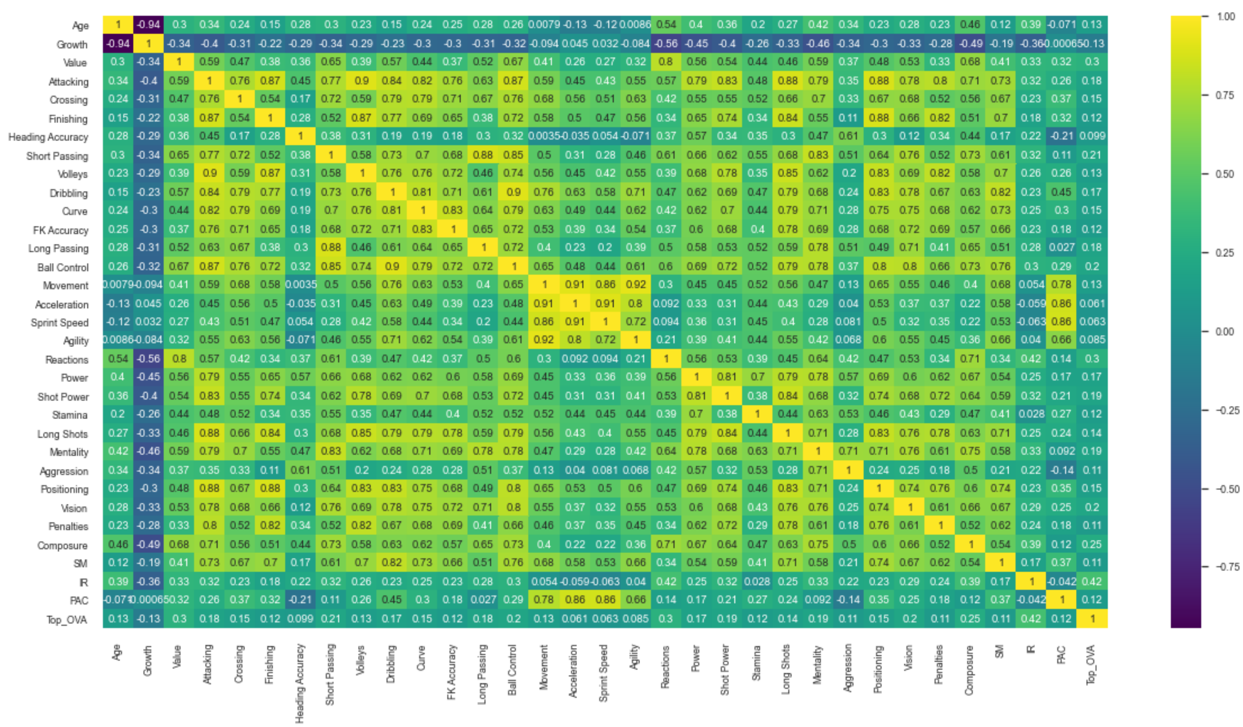
  Based on feature selection, 10 best features in my Fifa 21 dataset to predict if a player would have an overall score of more than 80:

```
     Feature_Name         Score
2           Value  13252.998808
30             IR   4752.118873
18      Reactions   2159.213356
28      Composure   1095.681369
26         Vision    716.005257
20     Shot Power    528.308178
7   Short Passing    438.133000
12   Long Passing    429.927948
19          Power    406.844763
23      Mentality    395.134172
```

Pearson Correlations:



Spearman Correlation:



Feature Selection

My original dataset consists of data/stats of Fifa 2021 players and the dataset has 17125 rows and 107 columns. I've done following steps for feature reduction:

- Drop the column with no significance.
- Format data to consistent format such as M(millions) and K(thousands) to money* 1000000 or money*1000, height to inches, etc.
- Drop all the text variables which can't be translated into numbers.
- Verify if any column has nulls. In my dataset, I had 300 null values for Composure. Therefore, applied median to missing values for Composure.
- After performing these steps, my reduced data frame has 13334 rows and 32 columns.
- Then train and test data to predict if the player would have a great overall score (> 80).
- Perform feature selection using sklearn's VarianceThreshold with threshold of 0.5.
- Conducted Feature selection using SelectKBest.

  Based on feature selection, 10 best features in my Fifa 21 dataset to predict if a player would have an overall score of more than 80.

```
     Feature_Name          Score
2           Value  13252.998808
30             IR   4752.118873
18      Reactions   2159.213356
28      Composure   1095.681369
26         Vision    716.005257
20     Shot Power    528.308178
7   Short Passing    438.133000
12   Long Passing    429.927948
19          Power    406.844763
23      Mentality    395.134172
```

Part 3 - Model evaluation and selection

In the dataset, long with data formatting, cleanup and preparation I did introduce a Boolean column "Top_OVA". The intent is to evaluate a model for being able to predict if a player would have a overall score of 80 or above or not.

When evaluated with dummy classifier, got mean auc as .50 which means that the classifier is not able to distinguish between positive and negative class points. I picked up below four models to compare and evaluate:

- Logistic Regression
- Random Forest
- Decision Tree
- SGD

Below is how models evaluated.

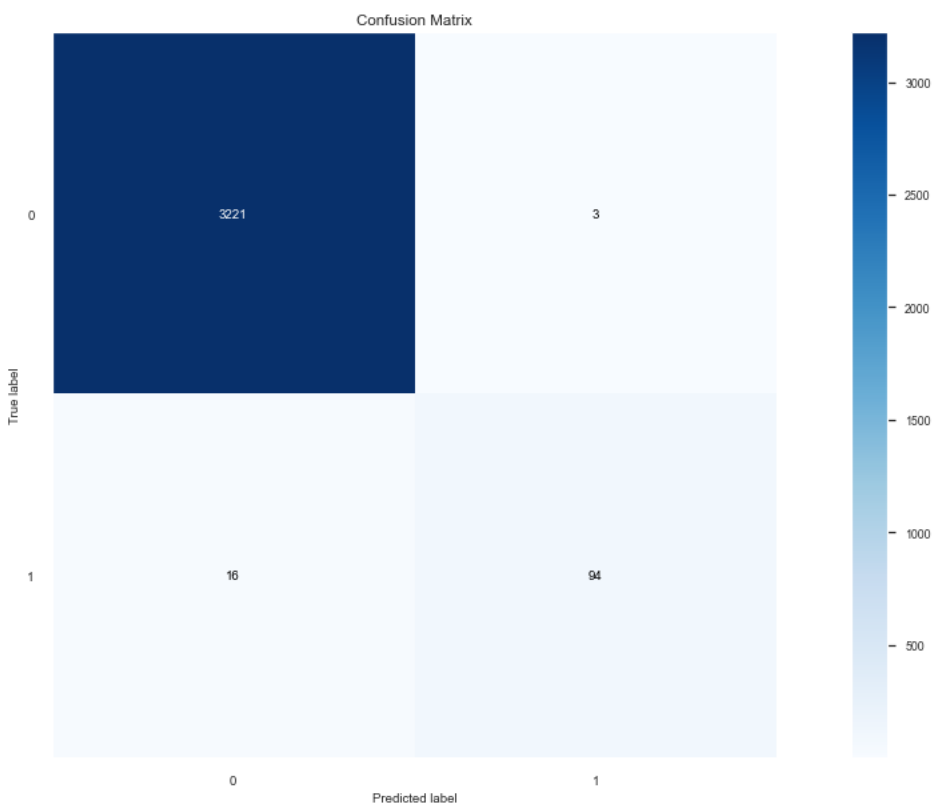
Model	roc_auc
RandomForestClassifier	0.998109
DecisionTreeClassifier	0.925596

SGDClassifier	0.998718
LogisticRegression	0.998994

Below is the result with **logistical regression**:

```
Confusion Matrix
[[3221    3]
 [  16   94]]

Classification report
              precision    recall  f1-score   support

           0     0.9951    0.9991    0.9971      3224
           1     0.9691    0.8545    0.9082       110

    accuracy                         0.9943      3334
   macro avg     0.9821    0.9268    0.9526      3334
weighted avg     0.9942    0.9943    0.9941      3334

Scalar Metrics
        AUROC = 0.9986
```

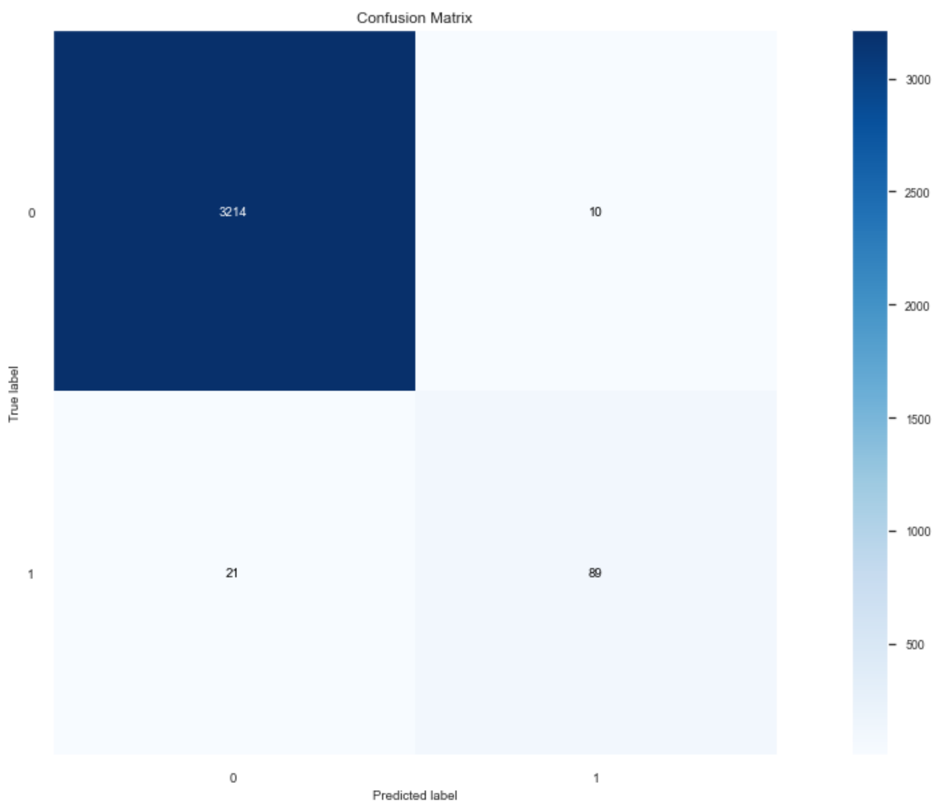
Plot Confusion matrix for logistical regression:



Below is the result with **Random Forest model:**

```
Confusion Matrix
[[3214   10]
 [  21   89]]

Classification report
              precision    recall  f1-score   support

           0     0.9935    0.9969    0.9952      3224
           1     0.8990    0.8091    0.8517       110

    accuracy                         0.9907      3334
   macro avg     0.9462    0.9030    0.9234      3334
weighted avg     0.9904    0.9907    0.9905      3334

Scalar Metrics
        AUROC = 0.9968
```

Plot Confusion matrix for Random Forest:



Conclusion:

Below are the conclusions of this case study:

- Highest number of players are within age range of 20-24.
- It is evident here that a large percentage of players are right foot dominant and there are large number of midfielders compared to other playing positions.
- International reputation of 3*/4* are not so common among players and very few have 5 * rating.
- UK has the highest number of players based on Nation wise distribution.
- When compared between models, logistical regression and random forest models were the best model.
- And looking at the confusion matrix between logistical regression and random forest model, random forest model seems to have performed slightly better.