Soccer Player Attribute Analysis

Group anonymous

Shakti Singh
B00779881
Masters in Applied CS
Dalhousie University
Halifax, NS

Tushar Gupta
B00782699
Masters in Applied CS
Dalhousie University
Halifax, NS

Prashant Pandey
B00779835
Masters in Applied CS
Dalhousie University
Halifax, NS

***Abstract* - Sports analytics has been already be deemed as a revolution in sports, but it is not being used enough. Our project aims to replace the 'conventional wisdom' used in sport with something tangible like factual results from data. We have explored the importance of player attributes and how their playing style is correlated with their attributes. Player attributes were visualized to assign a visually distinguishable pattern which can classify the player according to its position . We have created best squad predictor which can extract the best possible squad within the roster list. To understand the prominence of the attributes among players we have plotted scatter plots for each attribute against each other. The applet 'Player position predictor' can help the coaches find the best suited player for a position using machine learning.**
Another aspect of this project is the exploration of some 'less obvious' factors which can directly influence the game.

***Keywords* - Football, Machine learning, Random Forest, attributes, FIFA**

I. INTRODUCTION

Modern Sports team heavily relies on statistics to help them improve their game. Players are valued based on their overall stats and the potential they show. Player skills have been quantified and are used to value them. Each player gets have their attributes quantified so that their performance can be measured accurately. Different attributes represent skill prominence in that player and groups of attributes can be seen prominent in athletes with similar playing style.
Our project focuses on finding out those group of attributes which can be used to find out players playing style or what style should suit the player best.Another aspect is findi

We have tried to achieve our goals through Machine Learning and Visualization. We have trained a model on a dataset of 18000 players with each having 17 attributes that contribute to their overalls and playing style. On the other hand, we have made visual model for each player to visualize their attributes.
Another aspect of the project aims at using the data to explore the possible factors that could influence the outcome of a football match.In the world of club Football, teams spend millions of dollars on perfecting gameplay strategies. But sometimes there are less obvious factors that could be involved in the outcome of a football match. Thus exploring such phenomenons should also be an important part of the analysis.We have visually analyzed weather 'home advantage' exist in football and does paying more wages to the players ensures good performance or not. As a future reference, a blog[2] is studied which can be used as a model to analyse the effect of strategies on match outcome.
The project is divided into various modules, each module focuses on player attributes and tries to achieve a conclusion. For the Best Squad predictor players attribute and their overall are used as the criterion for their selection in the best squad, for position in the given formation the player with the highest overall is chosen for that position. For figuring out which attributes are prominent in different playing style, we have used scatter plots to differentiate attribute dominance between playing positions. By plotting graphs against one another we were able to achieve.

II. METHODS AND MATERIALS

DATASET

1. 'CompleteDataset.csv'[1](Kaggle.com, 2017) comprises player attributes, pictures, overall, potential, clubs, nationality, value, and their preferred position. We have done data

pre-processing to get the value and wage into a code friendly format, preferred position was also modified to get a single position. Minor changes were done to the attribute column some rows had a mathematical expression which needed to be evaluated before fetching its value. The dataset has data for 18000 players in 676 football clubs from across the world.

2. 'SoccerLeagues.csv' [1](Kaggle.com, 2017) has he home and away performance of 8365 soccer teams from a few dozens of countries during the years 2010-2016.
3. 'Country_facts.csv' [1](Kaggle.com, 2017) contains facts about 88 countries including soccer data such as their FIFA rank, the average attendance of soccer matches.

Machine learning approach

One of our major objectives of this paper has been to discover how player attributes are affecting player position and playing style. Therefore, from the FIFA 18 dataset we have extracted all the attributes of the players in a Pandas Dataframe to train a Random Forest Classifier model.

Random Forest Classifier [4](Scikit-learn.org, 2017) is the optimal classifier in our scenario since we have 15 positions in a soccer field and the data is sparse we needed a bagging algorithm which can take average of all the best cases.

We have used the following flags in our model :
max_Features : Auto, n_trees= 50,
criterion='entropy' (Gini index gave similar performance ,so there was no effect of criterion).

We have trained our model on 70% of the dataset and tested on the rest 30% , we were able to achieve a score of 60%. This was the best score we could achieve among the classifiers we tested : Decision Tree, SVM (kernel: linear, poly), Linear Classifier and Naïve Bayes Classifier.

The lower result is attributed to low number of training cases, the sparsity of the dataset and high number of label classes (15).

Random Forest : 60 %
SVM, kernel- linear :50%
Decision Tree : 46 %
Naive Bayes : 32 %

Best Squad Finder

Using Python, we have done data analysis to find the best possible team in each formation. It is useful to visualize the team in each formation as well as it can be used to try out various combinations in the team with just a click of a button. Players are selected in the team based on their overall score. More features can be considered into this selection criteria on user demand.

Visualization for this tool was inspired from the FIFA 18 game, where a similar interface is used to represent the team formations and the overalls of the team. A Football field is used, and the player picture is positioned on the position he plays in. The whole field is divided into 15 positions representing the 15 positions in soccer. The whole visualization gives clear visualization of the best squad in the selected formation to the user. The backend provides with the best 11 players for the selected formation and in the front end those 11 players are places on to their respective positions, depicting the best squad in the selected formation.

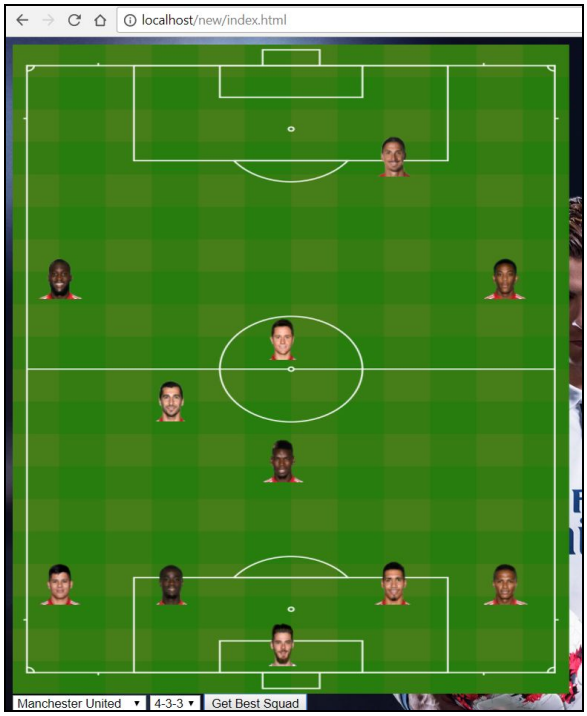


Fig 1. Best Squad Finder

Attribute comparator

Attribute comparator is used to differentiate the attribute prominence in 4 playing positions (GK, Defender, Midfielder, Striker). We have used scatter plots with colours to easily observe the clusters of each playing style. Different clusters can be observed with some outliers depicting the prominence of that attribute among the defenders, forwards, midfielders or GK. Instead of plotting all graphs at the same time we have created a UI to help the user select the

attributes and then a graph would be generated according to the attribute selected.

Player position predictor

To find the best position suitable for the player with a given set of attributes we have used a Random Forest Classifier which is trained on our dataset to predict the best position of a player with a given set of attributes.

Sliders have been used to take inputs from the user. Sliders are a fast and error-free way of taking values from the user.

B. ATTRIBUTE PROFILING

The dataset [1](Kaggle.com, 2017) we are working on consist of attributes of the players. These attributes signifies the skill of a player in different areas like dribbling, crossing, balance, marking, ball control etc. These attributes when visualized through a radar plot give a certain shape to the graph according to the playing position of the player. The attributes chosen for the radar plot are a mixture of skills that are found in players playing at different positions like as a striker, a goalkeeper, a midfielder or a defender. After strategically aligning the different attributes, we were successful in visualizing different playing positions as distinct shapes in the radar plot. These distinct shapes are termed as “attribute profiles” of the players.

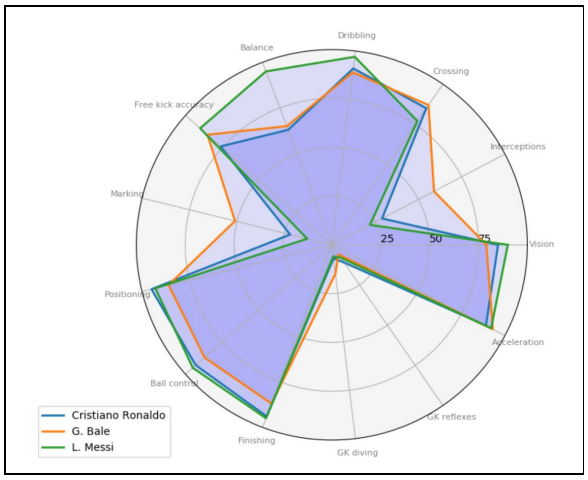


Fig 2. Attribute profiles of Strikers

The above graph shows the attribute profiles of some players playing at striking positions. There are some attributes that hold high values for the strikers like positioning, ball control, and dribbling while some attributes like marking, interceptions and goalkeeper reflexes hold low values. These specific values give a distinct shape to the attribute profiles of thee strikers in general.

Similarly, the attribute profile of goalkeepers gets its distinct shape from the high values of the goalkeeper attributes and low values in the skills not related to goalkeeping.

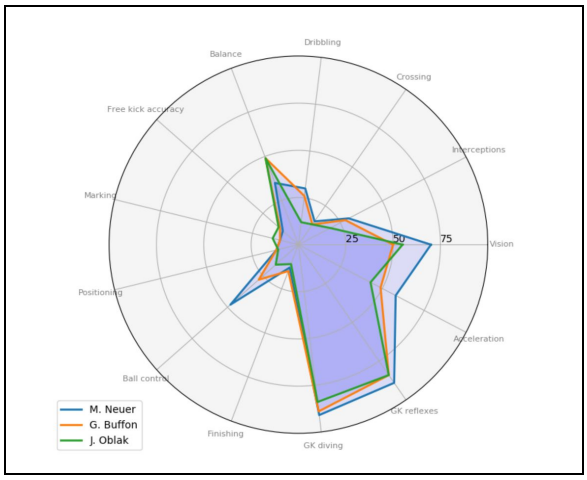


Fig 3. Attribute profiles of GoalKeepers

The Players playing at mid-field position and defending positions also exhibit this kind of specific shape of radar plot as shown in the figures below.

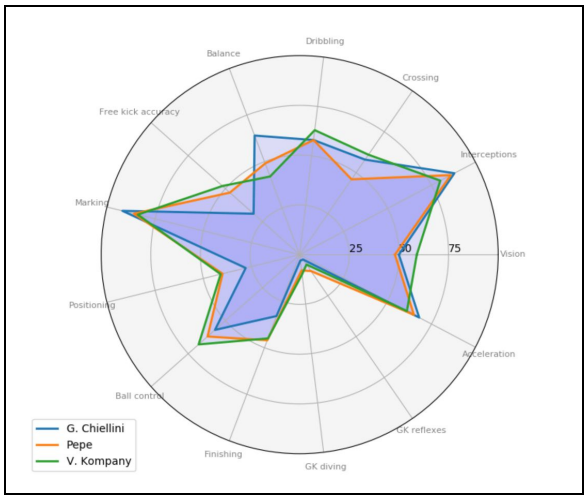


Fig 4. Attribute profiles of Defenders

The main motive behind this attribute profiling was to conveniently visualizing the attributes of a new player and just by looking at the shape of the graph, we can easily determine the appropriate playing position of the player. This tool can also be used to

compare the abilities of certain playing with respect to each other.

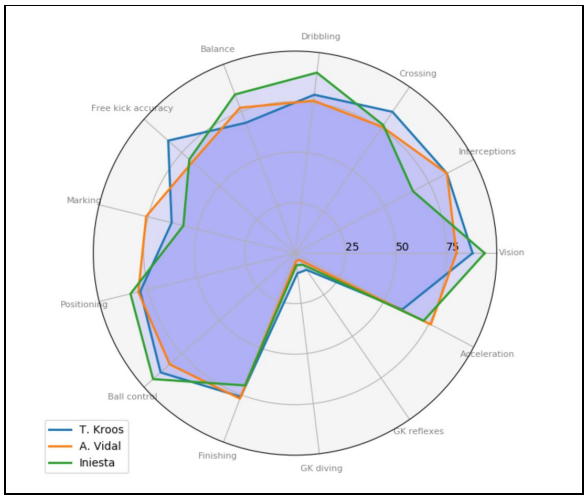


Fig 5. Attribute profiles of Midfielders

The development of this tool was accomplished in python programming language. "Tkinter" library is used for creating a simple UI that plots the graph and clears it according to the user input through a button. The list of players include all of the in our database and any of those players can be visualized through this tool. This tool can also be used in integration to other systems easily.

Radar plots are specifically chosen for this task because it is easy for even a layman to understand them. A lot of information can be visualized using these plots. Moreover, radar plots create a visual similarity between the playing positions yet distinguish them easily.

WHY PYTHON ?

The choice of a programming language for any project depends upon the functionality and ease that a programming language can provide towards the project. Python is a very robust language that comes with several in-built libraries and external packages that assist in the task of data processing and visualization. "Matplotlib" library in python provides several kind of plotting functions that can be used for a myriad of data. The radar plot that is used in this tool is constructed using the polar-plot subplot of matplotlib library. For adding the user interactivity to the plot, a simple UI is created using the "Tkinter" external library.

Tkinter

Tkinter is a GUI (graphical user interface) widget set for Python. *Tkinter* [3](Docs.python.org, 2017) is a set of wrappers that implements tk widgets as python classes. It can be used to create many types of user interactive applications.

C. DATA VISUALIZATION

Country-Wise Attribute Density

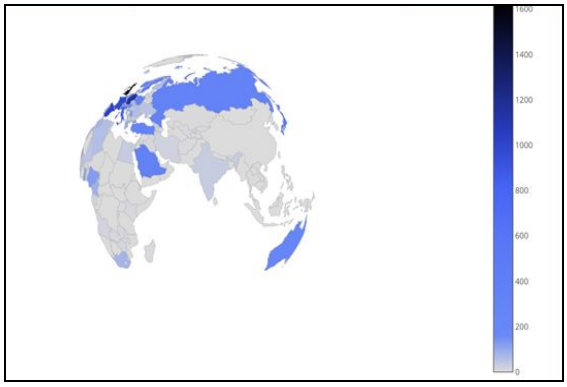


Fig 6. Nation-Wise Player density

The attributes of the players derived from the data-set were visualized on the basis of player's Country. The visualization allows user to select statistical values of any attribute. This visualization is useful for seeing which country has significantly better or worse physical and mental abilities related to the game. Plotly library in python was used to create the 'Choropleth'

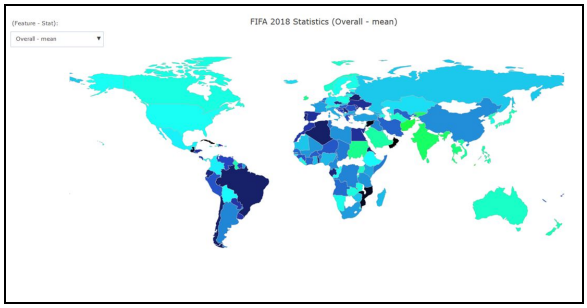


Fig 7. Home vs Away Points Difference

Home Advantage Exploration

Home advantage is a phenomena observed in sports , where the home-ground winning percentage of a team is higher than the away-winning percentage. We visualized our data in order to find out weather this holds out true for our football or not.

Home goal difference is the difference between goal conceded on home ground subtracted by the goals Scored.

This ploat was created with matplotlib in python.

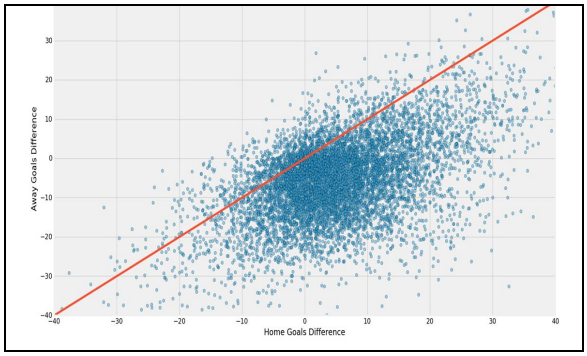


Fig 8. Home vs Away Points Difference

Similarly , to show qualitative difference we could plot the winning percentages.

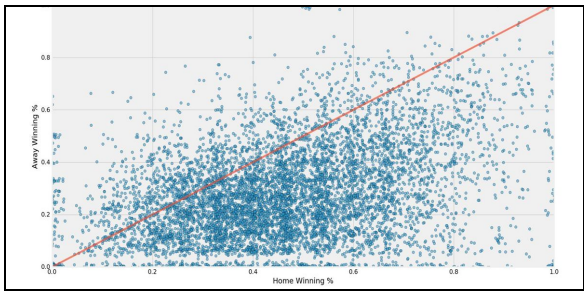


Fig 9. Home vs Away Winning Percent

Analysing the advantage further we tried to plot the home winning percentage along with the audience attendance to check whether 'audience motivation' is a factor deciding the win or not.

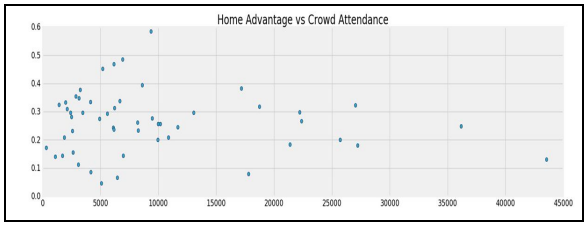


Fig 10. Attribute profiles of Defenders

Wages Effect on Performance

We tried to gauge the effect of wages on the performance of the player using the results of English Premier league.

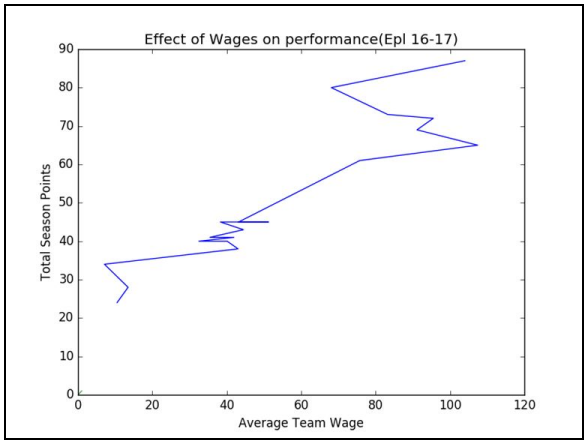


Fig 11. Wages effect on Performance (EPL 16-17)

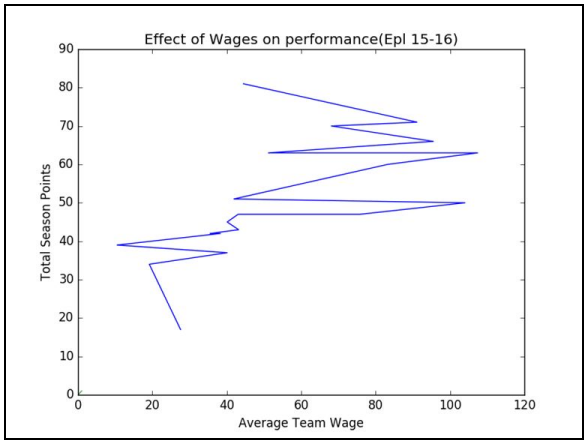


Fig 12. Wages effect on Performance (EPL 15-16)

We can seen that the team with wages lower than the average wage was able to gain the maximum point that season. The next section presents the analysis of that team.

Leicester City Analysis

The data used is not public hence such analysis was not possible in this project . However , it perfectly describes the future direction of this project (blog.acolyer.org).

In this analysis Leicester's 15-16 performance was compared to the previous to evaluate the changes in the gameplay tactics.

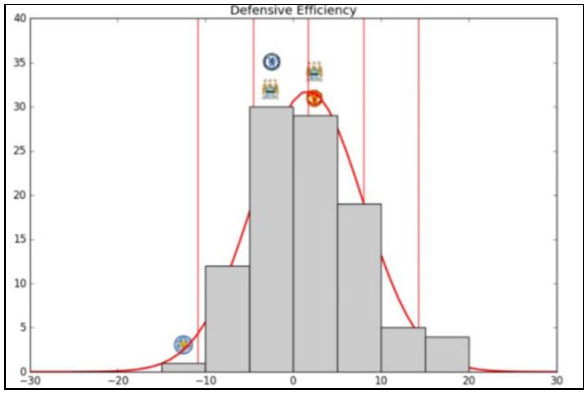
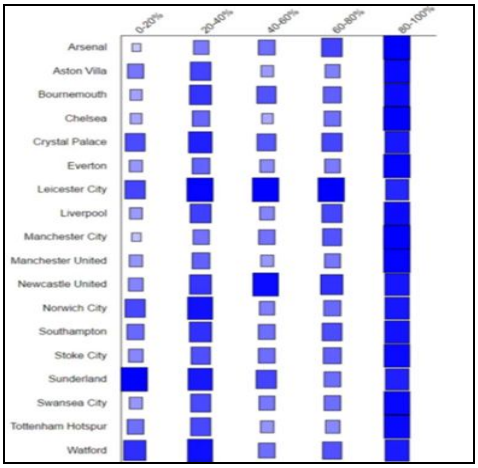


Fig 14. Leicester defences Analysis

The ‘defensive efficiency’ is defined as ratio of goal conceded to goals expected to concede. The later value was calculated on the basis of the previous years performance.



.

Fig 15. Pass interception analysis

In this visualization , the ease of intercepting the pass is given in percentage . The color intensity of the box shows how many passes were intercepted.

III. RESULTS AND USE CASES

Use Cases

Best Squad finder:
When a new Coach joins the team, they are unaware of the squad, players playing style their strengths and weaknesses, how would they implement their playing style. Our tool helps the coach to select the best squad for their desired position and give a neat visualization on a football field. The tool makes all the work of going through the roster list go away and does all the work for you.

Position Predictor:
As a coach when you teach young talent, you can be guide them into any playing position. How would you decide which position the player should play in? our tool helps the coach by giving them a statistical answer to this problem by looking at the current attribute of the player we can either create an attribute profile and compare it with other player’s playing style and see which it matches the best OR we can use the player position predictor functionality to find out which position is best for player which will be based on their attributes.

Attribute profiling:
The attribute profiling module can be used by a trainer or instructor to closely monitor the performance of a player in different parts of the field. This rich insight gained can be applied to assign a better place to play according to their skills.

Results

Home Advantage Exploration
The home advantage is proven if maximum teams lies below the red line which is visibly true in our data-set.The increase in audience attendance did not cause any significant change to the home winning percentage.The precise cause of this advantage remains hidden. Maybe in the end it comes down to human psychology. This problem can be quantified with personalized player data which is out of the scope of this project.

Wages Effect on Performance
Although linear correlation can be drawn for the ‘16-17’ season , the same cannot be said for the ‘15-16’ season. It can be seen that the team with wages lower than the average wage was able to gain the maximum point that season.

Leicester City Analysis
Defense analysis shows that efficiency of leicester can be described as a ‘statistically extreme’ finding.
Leicester also shows excellent performance in all pass difficulty levels.
This analysis reveals key points in Leicester City’s gameplay. The defense was significantly improved as compared to the attack .High interception rate signifies high midfield pressure. These results shows that Leicester significantly improved their defense and increased midfield pressure. Attacking rate was same as last year. Hence there tactics can be defined as counter attacking with high pressure in midfield.

IV. CONCLUSION

The results of this project can be used as an inference for various purposes in the world of soccer. The machine learning approach that we introduced can be utilized to find a relation between the attributes of a player and the playing position they perform best on. By looking at the various scatter plots that certain attributes are prominent in particular playing position for example defenders have higher strength than rest of the players while forwards have high finishing and shooting than other players. For future work we would like work on the cost analysis of players by figuring out on what basis big teams decide to buy young players for prices as high ah 100 million pounds. We would even plan to scale our tools on different datasets to find out more in depth inferences about teams and its players and what can be done to improve their game a level more.

Analysis of Leicester city shows a data analysis model whose findings can be directly related to football tactics. The home team advantage analysis has left us in a limbo with ample scope of future work regarding the 'psychological' aspects too.

REFERENCES

[1] Kaggle.com. (2017). *FIFA 18 Complete Player Dataset | Kaggle*. [online] Available at: https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset [Accessed 11 October. 2017].

[2] blog.acolyer.org (04 September. 2017)."The Leicester City fairytale?[online]Available at: https://blog.acolyer.org/2017/09/04/the-leicester-city-fairytale-utilizing-new-soccer-analytics-tools-to-compare-performance-in-the-1516-and-1617-epl-seasons/.

[3] Docs.python.org. (2017). *Graphical User Interfaces with Tk — Python 3.6.4rc1 documentation*. [online] Available at: https://docs.python.org/3/library/tk.html [Accessed 6 Nov. 2017].

[4] Scikit-learn.org. (2017). *sklearn ensemble RandomForestClassifier — scikit-learn 0.19.1 documentation*. [online] Available at: http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html [Accessed 19 Oct. 2017].

[5] Kaggle.com. (2017). *Home advantage in Basketball| Kaggle*. [online] *— scikit-learn 0.19.1 documentation*. [online] Available at: https://www.kaggle.com/drgilermo/home-advantage-in-soccer-and-basketball [Accessed 19 Oct. 2017].

.