Soccer Analytics

Aniket Giriyalkar
Department of Computer Science
Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, NY 14586
aag5405@cs.rit.edu

***Abstract*—Soccer is the most widely followed sport all over the world. Decisions like selecting the starting lineup, promoting players from academies, transfers, etc. used to be purely based on gut feeling. Since the introduction of Soccer Analytics, such decision making has changed drastically. Nowadays, teams depend heavily on data science and analytics for making crucial judgements. Soccer Analytics is not just restricted to the management, team, media, and the players. Even the fans use Analytics to select the best possible squad for the fantasy leagues they are enrolled in. In my capstone project, I propose to create a list of top players for each position who played in the world cup based on their performances over the entire year and compare them with the players that featured in the Team of the tournament. and create an Expected Goal(xG) model that predicts the number of goals scored based on the xG information of all the shots recorded in the World Cup 2018.**

***Index Terms*—Scraping; StatsBomb; Soccer Analytics; xG; Advanced visualizations**

I. Introduction

Soccer has always been a traditional sport and that is why the growth of analytics in this sport has not been as rapid compared to other sports. Unlike other sports, on-the-ball actions of players in soccer often provide less insights into the strategies and player evaluations.[18] As one of the soccer's greatest player Johan Cryuff once famously said, "When you play a match, it is statistically proven that players actually have the ball 3 minutes on average. So, the most important thing is: what do you do during those 87 minutes when you do not have the ball. That is what determines whether you're a good player or not." [1] Early researchers mainly concentrated on the events at a general level such as home-away advantage, strength of teams based on their standings and performances against the best teams from other leagues. Later, with on-the ball data collection led to analysis based on possession and individual player statistics. This gave rise to several strategies, one of which was the direct football(long ball tactics) [5], in which the physical vulnerabilities of a player are exploited by playing a long distance ball down the field generally bypassing the midfield of the opposition. Next step in the evolution of Soccer Analytics was the ability of the teams to track off-the-ball player movements over the entire game. This was a pivotal point that led to the popular use of Soccer Analytics. After winning the UEFA Champions league last season, Liverpool Football Club was leading the English Premier League by 25 points when the league was temporarily suspended due to COVID-19 last month. Yes! You read that right, they were leading the most competitive league in the world by 25 points, the same league in which they finished just one point behind the Champions Manchester City last season. Their manager Jurgen Klopp totally supports the idea of integrating the insights generated by their data science teams into his team decisions. [11] Liverpool is not the only team that has departments dedicated to data science. Most of the other teams competing at higher levels have it too, but looking at Liverpool's exceptional performances this season, they certainly seem to be doing better than others in utilizing the insights they are provided.

I have been a soccer fan for almost two decades now. In 2016, I came across a book - "The Numbers Game: Why Everything You Know about Soccer is Wrong" [12] which completely changed my perspective to look at soccer games. Since then I have been reading as many soccer analytics discussions/ blogs/ tweets as I can and researching on this topic in my free time. So, when I had to select a topic for my capstone project to complete my Masters program. I decided to combine my life-long passion with my academic ventures. In this project I will be working on the World Cup 2018 dataset from Statsbomb [9]. I will be scraping the player rating information from FIFA's website [13] and team information from ESPN's website [15] to create a team of best players from 2018(overall season) and then compare this team with the team of the tournament using advanced visualization. Next, I want to utilize the xG [7] factor associated with each shot to generate models which predict how many goals were scored in the tournament based on the data we had. In another model, I plan to use the xG from the dataset predicted by StatsBomb as a measure to compare my results with.

II. Background

Apart from the team management and the players, Soccer analytics is also common among fans, commentators and pundits. As a fan I have used analytics in one way or the other every season to select my Fantasy Premier League team. Pundits use advanced visualizations to discuss the various strategies the teams playing could use before the game in the pregame show and then talk about where they could improve after the game in the post-game analysis. The visualizations used by them are very interesting and collecting data for such analytical work is a complex task. There are companies

Fig. 1. 4-2-3-1 Formation in Soccer[8]

like StatsBomb, Opta, etc. which specialize in collecting detailed match related data. Any meaningful data analysis on Soccer data requires data to be of good quality and accurate. Maintaining such detailed and high quality data turns out to be expensive and these companies offer such data for use at some cost. Thankfully for me, StatsBomb [9] released a lot of high quality match data of various tournaments for research purposes for upcoming analysis like me. I selected the FIFA 2018 men's World Cup data for my project. The data is extremely detailed consisting of information like location of the shot, information of the event that occurs prior to a shot, locations of all the players on the pitch at instance of a shot, the information on the type of a pass leading to the shot, etc. Such data can be used to generate heat maps using which one can visualize the movement density of a player during a particular game or even the entire tournament. One can also compare different players based on their positional performances in the tournament using this data.

Figure 1 shows the name of positions in a 4-2-3-1 formation which is also the formation of the team of the tournament from the World Cup 2018. The Goalkeeper is the player who guards the goal and prevents the opponent from scoring, goalkeeper is not included in the formation. Defenders form the line of defence in front of the goalkeeper. Left Back and Right backs often contribute in the attacking play as well by playing higher up in the field and combining with the attackers. Good crossing ability is expected from players at this position. Center backs usually hold their positions throughout the match and hardly go further up the field. They pose a goal threat from free-kicks and corners. Good heading ability and physicality is expected from defenders. Defensive midfielders are the heart of the team and these players dictate the pace and flow of the game. They protect their defenders whenever they are needed. They are also responsible for moving the ball between the attackers. Players with good composure and passing abilities are preferred at these positions. Next come the Left Winger and Right Winger who are responsible for carrying out the attacks and for creating space for the Forward with their intelligent runs. Their movements are important for the other player to find them. Good finishing ability is also expected from these wingers. Attacking midfielder is the main creator in the team, this player is responsible for the attack flow of the team and their role is to set up the forward and wingers and even go for the goal themselves if they are in a good position. Forward is the player who plays farthest up in the pitch. Their role is to finish the chances created by other players and trouble the opposition defenders with their movements.

Expected Goal(xG) is the probability factor associated with a shot that determines if a shot might result in a goal or not based on the characteristics of the leading up to that shot. The characteristics are type of pass that lead to the shot, type of attack the shot originated from, the body part that was involved in the shot and location of the player from where they took the shot. xG is in the range 0-1. xG of 0 means that the shot is a certain miss. xG of 1 indicates a sure goal. [17] The freeze frame feature of StatsBomb model distinguishes it from others. Freeze frame is used to indicate the location of all the players on the pitch at the moment in which the shot was taken. xG does not take into account the quality of the player that is taking the shot or of the player who is involved in the shot. It is an estimate of how a normal player or a team would perform in a certain situation. xG can indicate whether a player has above average finishing / conversion ability if they consistently score more goals than their total xG. A team having a higher xG difference (xG - xG allowed) is bound to perform better than a team with negative difference. A negative goal difference but a positive xG difference indicates that the team has struggled with finishing. xG can also be used to evaluate strategies of defending/ attacking on set pieces. In this project, I will be using xG from StatsBomb to predict the number of goals in the dataset. In another model, I plan to utilize this StatsBomb predicted xG as a metric to compare with my model generated xG. Then use the best model to analyze the player and team performance in the World Cup based on the xG calculated.

III. Methodology

My capstone project is divided into three parts. The first part is forming a Best XI from the players that played in the world cup based on their performance in the World Cup 2018. Next part comprises an xG model which predicts the xG metric based on various features and the result is compared with StatBomb's xG metric. I will also be discussing the applications of this model. Lastly, I will create a model using

StatBomb's xG variable to predict the number of goals scored from the information available in the dataset.

A. Comparing my Best XI with World Cup XI

The World Cup 2018 StatsBomb dataset [9] consists of detailed data for each event and match from the tournament. Since, the World Cup is usually a short tournament with a total of 64 matches being played and a team playing a maximum of 7 games, if they make it to the semi-finals. It is very difficult to fairly predict a team for the tournament. For the scope of this project, I will be going ahead with the official team of the tournament and then compare it with the Highest Rated team of the season 2017-18. The aim of the comparison is to see if the players who performed well throughout the season continued their good form in the World Cup. If not, with the help of advanced visualizations, I will try to analyze what went wrong for them.

1) My Best XI for the 2017/18 season:
Since I will be comparing my team with the team of the tournament [7]. I will be only collecting the 2017/18 seasonal data of the players who played in the world cup. The data will consist of ratings for various skills based on their performances over the entire season. First, I collect the team data of the teams that played in the World Cup from ESPN's website [15]. From the squad info, player details are extracted and the skills for that respective player are obtained from the FIFA website [13]. All the player specific information from this website is obtained and stored into a data frame.

Now, I have data for each player that played in the world cup. I plan to use these ratings in a couple of ways. Firstly, I want to use a weighted average of qualities that give higher weights to qualities that are must for a position and then pick the player with best value for that position. Like for a goalkeeper, qualities like ball distribution, claiming crosses and long passes, and contribution to the attack [4] are given higher weights than other characteristics. Second way would be to find a sum of certain numbers of skills that are required for each position and then pick the player with the highest value for that position. In my opinion, the second version seemed a lot more fairer as FIFA's ratings resemble the players performance over the entire season and using them as it is would make more sense. Also, using the weighted average method might add an extra bias and might work only for players from a certain league as the weights are assigned keeping in mind a particular type of game play.

So, following the second method I will pick the best player for each position by adding the value of ten features that are essential for each position. Also, the positions will be based on a 4-2-3-1 formation as it would be easier to compare players directly with their respective positions. Before starting with the analysis, I drop all the rows containing Null values. For the Goalkeeper(GK) position, the features that I considered were 'GK Reflexes', 'GK Diving', 'GK Handling', 'GK Positioning', 'GK Kicking', 'Strength', 'Jumping', 'Reactions', 'Balance', and 'Composure'. For the Right back position, 'Acceleration', 'Sprint Speed', 'Stamina', 'Balance', 'Agility', 'Jumping', 'Stand Tackle', 'Slide Tackle', 'Aggression', and 'Strength' are the features that are selected for player comparison, these features require the Right back(RB) to be active both in attack and defense. Similarly for the Left back(LB) position same qualities are expected. The center backs are expected to be strong at the back and should have a good heading capability so that they can provide an attacking option from set pieces. 'Acceleration', 'Heading', 'Composure', 'Balance', 'Agility', 'Jumping', 'Stand Tackle', 'Slide Tackle', 'Aggression', 'Strength' are the features I have used to rank the center backs(CB). Central midfielders(CDM) are expected to provide support to both attackers and defenders and are supposed to keep the ball moving and hence the features that I selected for their comparison were - 'Acceleration', 'Sprint Speed', 'Dribbling', 'Balance', 'Agility', 'Ball Control', 'Long Pass', 'Short Pass', 'Marking', and 'Vision'. For the Right wingers(RW) and the left wingers(LW), qualities like dribbling, crossing, finishing and ball control are the skills that describe how good a player is, the features that I used for comparison are 'Acceleration', 'Sprint Speed', 'Dribbling', 'Balance', 'Agility', 'Ball Control', 'Shot Power', 'Short Pass', 'Crossing', and 'Finishing'. For a Central Attacking midfielder(CAM) vision and finding spaces for forward and the wingers, along with composure and finishing is the main requirement, so I used the skills 'Acceleration', 'Vision', 'Dribbling', 'Balance', 'Composure', 'Ball Control', 'Shot Power', 'Short Pass', 'Crossing', and 'Finishing' to compare the players at this position. For the Central Forward(CF), finishing, composure, heading and pace are the most important qualities. I have used the skills 'Sprint Speed', 'Acceleration', 'Strength', 'Balance', 'Agility', 'Ball Control', 'Shot Power', 'Stamina', 'Heading', and 'Finishing' for comparing players at this position. Table 1 summarises the best XI and that I evaluated using the above method. It also displays the Team of the tournament for the World Cup.

TABLE I
BEST XI AND WORLD CUP 2018 XI

Best XI	**Position**	**World Cup XI**
Manuel Neuer	GK	Hugo Lloris
Marcelo	LB	Ashley Young
Kyle Walker	RB	Kieran Trippier
Sergio Ramos	CB	Raphael Varane
Thiago Silva	CB	Dejan Lovren
Luka Modric	CM	Luka Modric
Thiago	CM	Paulinho
Eden Hazard	LW	Neymar
Kevin De Bruyne	CAM	Antoine Griezmann
Lionel Messi	RW	Eden Hazard
Cristiano Ronaldo	CF	Kylian Mbappe

2) Comparing the players - Best XI vs World Cup XI:
From Table 1 it is evident that most of the players who had a great season with their respective clubs did not continue their form going into the World Cup. However, I can also confirm that Antoine Griezmann, Hugo Lloris, Neymar, Rapheal Varane and Kylian Mbappe appeared in the top 5 list for their respective position while computing my Best XI, which

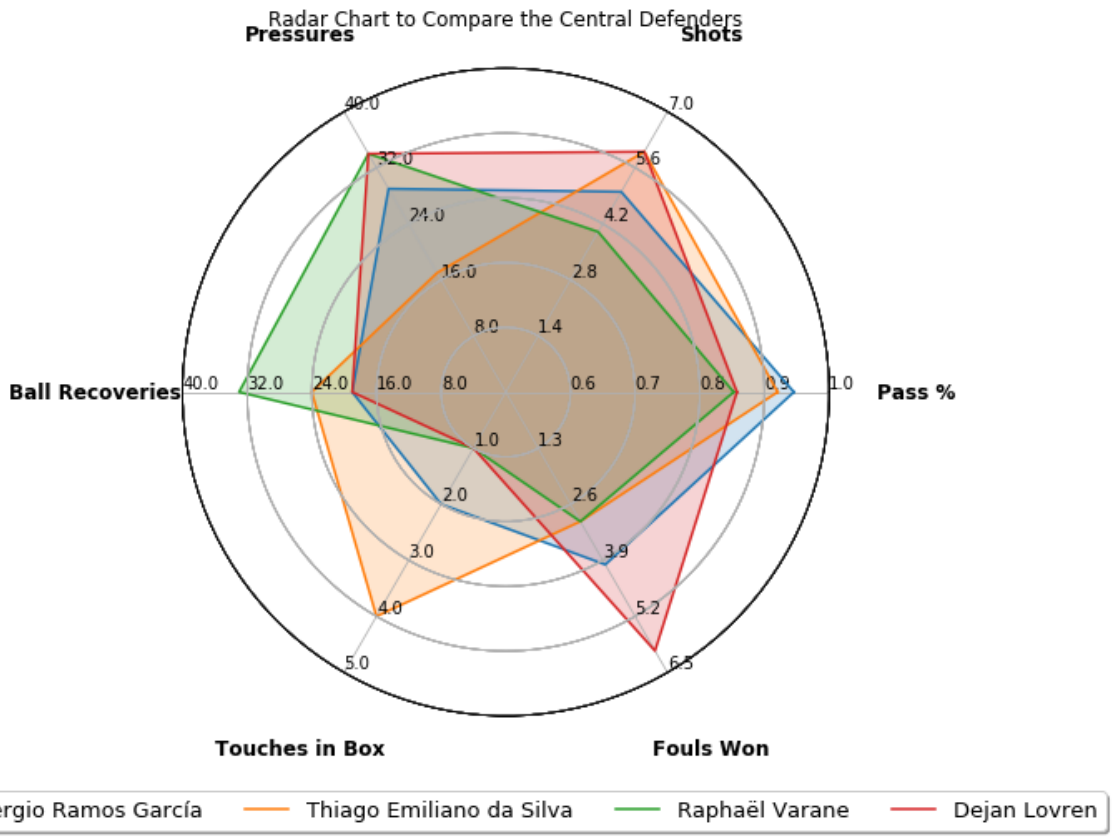

Fig. 2. Comparing CB's from my Best XI and World Cup XI using a Radar Chart

indicates that these players along with Luka Modric and Eden Hazard carried forward their good form into the tournament. Let us see how the players from my Best XI fared in the World Cup.

Manuel Neuer did not have a great tournament as Germany were knocked out in the Group Stages and he did not get to do much in their three games. He was also responsible for conceding their tournament ending goal against South Korea. Marcelo had a decent world cup, his team got knocked out in the quarter finals. Kyle Walker whose original position was Right back but played as a Central Defender in the back three defensive lineup for England had a good world cup. But the defensive and attacking numbers of his teammate, Kieran Trippier who played as a Right Back for England were too good and he was picked in the World Cup XI over Kyle Walker. If they progressed further he would have surely been in contention for the LB position in the team of the Tournament.

Thiago Silva and Sergio Ramos had a good world cup, but as their teams couldn't progress deeper in the tournament they did not make it to the World Cup XI. On comparing the defensive statistics(see Figure 2) like Pressures, Shots, Pass percentage, Ball Recoveries, Touches in the Box and Fouls won with the Central defenders Raphael Varane and Dejan Lovren who made it to the World Cup XI, it was observed that Varane and Lovren had better numbers in important categories like shots, fouls won, ball recoveries and pressures. Also, the fact that their teams made it to the final was an important criteria for them getting picked up in the team of the tournament.

Luka Modric who was present in both, the Best XI and World Cup XI was also awarded the Golden Ball for his performance in the World Cup, this award is usually given to the best player of the tournament. Thiago who played only in two games for Spain in the group stages did not get picked for the remaining games. He had a disappointing world cup compared to the club season he had. Paulinho, whose Brazil made only till the quarter finals was picked up in the World Cup XI for his good performances throughout the tournament.

Moving on to the Right and Left Wingers, Eden Hazard who played the entire club season as a Left Winger for Chelsea, played for Belgium in Right Wing position and was awarded for his performances by getting picked up in the World Cup XI as a Right Winger. Neymar who had a good season with his club PSG also had a good world Cup too. He had most shots on target in the world cup. He was unlucky as most of his shots were either blocked or saved. He ended up with only two goals, in spite of that his performances were strong enough for him to be picked in the team of the tournament. Messi who was the Right Winger in my Best XI played at a Central position for Argentina, he ended the tournament with

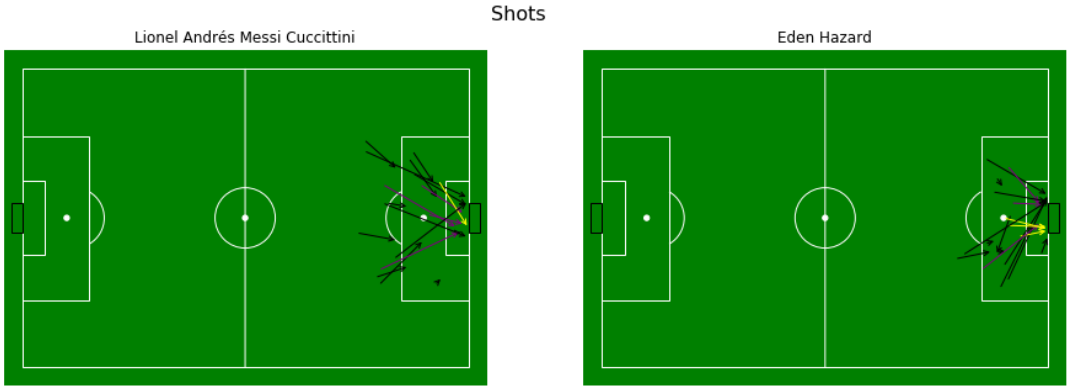

Fig. 3. All the shots of Lionel Messi and Eden Hazard from World Cup 2018

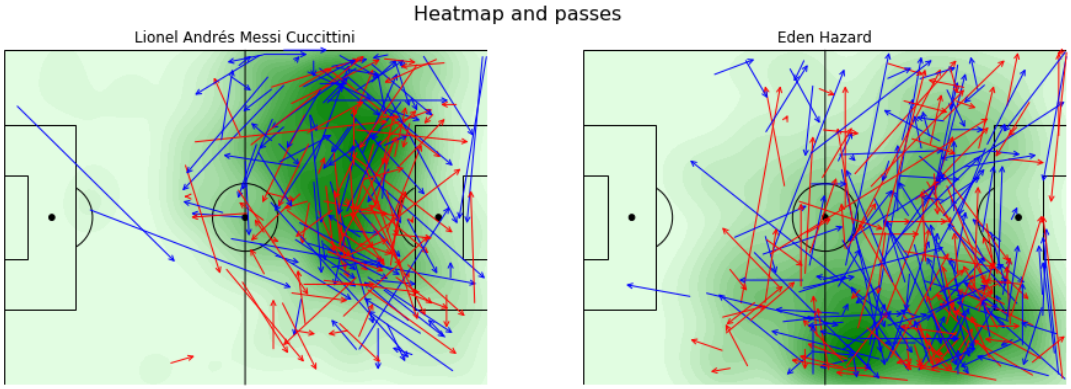

Fig. 4. Heat maps and passes for Lionel Messi and Eden Hazard

one goal and an assist which is poor going by his standards. On looking at the shots taken by Messi and Hazard over the entire world cup(see Fig. 3), where the yellow arrows indicate the goals scored. Purple arrows indicate the shots that were saved by the keeper and black arrows indicate that the shots were blocked by other players or were off target. Even though Messi's Argentina were knocked out by the Champions France in the Round of 16, Messi's numbers in terms of shots and passes were impressive. He almost had the similar number of shots compared to Hazard in- spite of playing three less games. He was unlucky as he could score and convert only one of those shots into a goal and his team got knocked out too early compared to Hazard's Belgium which finished third in the tournament.

The heat maps and the passes plot(see Fig. 4) revealed some important information about the positions. Hazard's heat map revealed that he stuck to his position for the majority of the tournament which means he executed the plans that were laid out to him, we can see that by the even distribution of red and blue arrows which represent the first half and the second half touches respectively. For Messi the position that he played in was not quite clear as we can see from his tournament heat map that he kept switching between the positions of a Left Winger and a Central attacking midfielder. We can also see that the majority of the red arrows are in the center, and blue arrows are on the left. This indicates that he played as a left winger in the first half and moved in the center in the second half. This type of analysis could only be possible with the help of location data that was recorded by StatsBomb [9].

For the Central Attacking midfield position Kevin De Bruyne who had an incredible season with Manchester City also played well in the world cup and could have been picked in the World Cup XI if Belgium had won the world cup. Unfortunately for him, Antoine Griezmann beat him to that place as he scored four important goals in the tournament including a goal in the final. Griezmann also had the second highest number of shots in the tournament. Cristiano Ronaldo had a good world cup where he scored four goals in hour games including a hat-trick against Spain. But as Portugal got knocked out in the Round of 16, Kylian Mbappe who also scored four goals including a goal in the final was preferred over him in the Team of the tournament. Harry Kane from England ended up as the top scorer of the World cup with 6 goals did not make this list as three of his goals came against

Panama which is not a strong team. Also, three of his six goals were penalties. Kylian Mbappe who is a teenager knocked out Argentina with his two goals in that match which proved to be a solid reason for picking him in the team of the tournament.

B. Creating a xG model

For creating a xG model, one requires good data. This kind of data is available at some cost from the companies like Opta, StatsBomb, etc. that specialize in maintaining detailed high quality. StatsBomb [9] recently made public match data from a number of leagues and tournaments for free. The World Cup 2018 data that I am working on was obtained from StatsBomb open source GitHub repository. This data set consisted of over 1700 shots. Data files obtained from StastBomb [9] were in the .json format.

1) Data Cleaning and Preparation:

The data containing the World Cup 2018 data was extracted from the local copy of the repository. Competition "43" mapped to the World Cup 2018 and using this information, a list of all matches belonging to this tournament was obtained. After obtaining the information of all the match files, event data was extracted for each match using the match numbers. Shot specific information like shot location, timestamp of the shot, outcome, body part, technique, play pattern, xG prediction of StatsBomb, player name, key pass, preceding event player and team information, cross, cutback, etc. was extracted from the events directory for each match and parsed to its corresponding lists. A variable called pack density[6] was introduced, it calculates how many players are in between the goal and the shot location. This is calculated using a Barycentric Technique [14] in which a triangle is created by shot location and edges of the goal and the location of each player is checked if they lie in the triangle. Total number of players in the triangle will give the pack density value for that particular shot. Euclidean distance is used to calculate the distance between the shot location and the goal and shot angle is calculated using the cosine rule to find the angle between shot location and goal edges. All this information is stored into a data frame. Then, all the columns are checked for the number of missing values. We found that the columns key pass type and key pass pattern have 526 missing values and the preceding event player has 13 missing values. Let us look into these columns in depth.

Preceding event players can be missing in events like a direct free kick. On looking into the preceding event information, we can see events like tactical shift, half start and camera on which were not really important with respect to the shots. So, it makes sense to drop these rows. Moving on to key pass pattern, on looking at the values this column holds, we can find that there are 520 Regular play key passes, followed by passes from Throw in, Corner and Free kick which are above 160. All the other key passes are less than 50. So, we associate all the shots with missing key pass information(key pass type and key pass pattern) to have the value "None". After this is done, we pickle our data frame so that it can be used in other parts of the project.

2) Exploratory Analysis:

The pickle file stored in the previous section is loaded. This data frame contains 1693 entries of shots with 27 columns. Packages like matplotlib, pandas and seaborn will be used to construct visualisations for exploration. Just to get an idea of all the shots we plot all the shot locations on a football pitch. We can observe from the Fig. 5 that more goals are scored from shots that are taken closer to the goal. Shots in red dots indicate goals. We then add 'isGoal' and 'isGoalBool' columns to our dataframe which store the information if a shot resulted in goal.

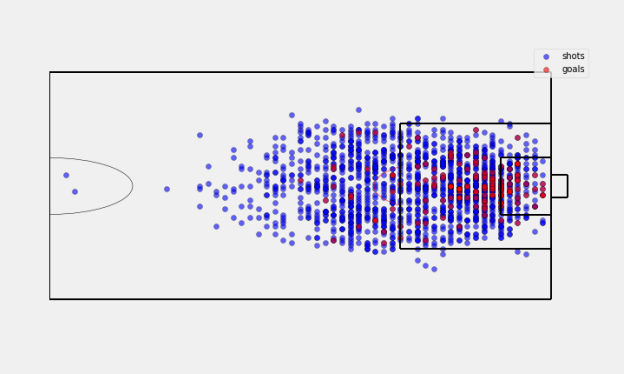

Fig. 5. Location of all the shots in World Cup 2018

Next, we browse through various columns of this data frame and compare them with respect to our newly generated columns. In the column type name, we go through the shot types that were from open plays. Within this we try to look at the body part involved in open play. From the Fig. 7, we observe that right footed shots clearly dominate the headers and left footed shots. The goals from headed shots are quite less than the footed shots. On evaluating the percentage conversion of headed and footed shots, it was observed that headed shots have a higher conversion rate, i.e. 10% than the footed shots which is 8%.

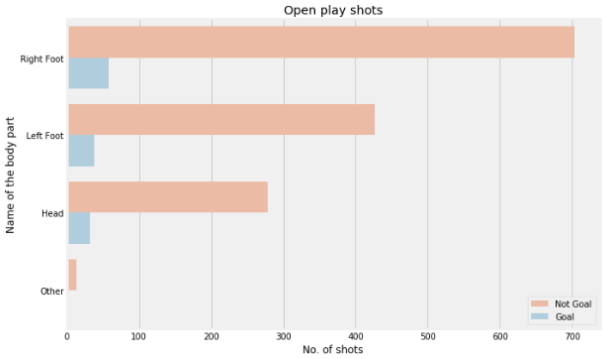

Fig. 6. Body Part involved in the Open Play Shots

This is a surprising result with respect to the plot we just observed. Looking further, into the body part involved and also considering distance of the shots from goal this time we see that more footed shots are taken from distances further from

the goal and these shots do not result in goal and the footed shots that are taken within the penalty box(within 18 yards) have higher chance of resulting in a goal. We can observe clearly from the Fig. 7 that almost all the headed shots(leaving some outliers) are taken from within the penalty box. This biases the comparison of headed and footed shots that we had earlier. To carry out a fair comparison between the two, we must compare only the shots that were taken from within the penalty area. The shot conversion percentage for footed shots within the penalty area is 15.8% compared to 10.6% of the headed shots. Hence, we see that the footed shots are more effective than headed ones when taken from within the penalty box. When we tried to perform the same kind of analysis for the non open play shots we got a division by zero error. This was because one cannot directly head the ball in a free kick or corner. It has to have a preceding pass.

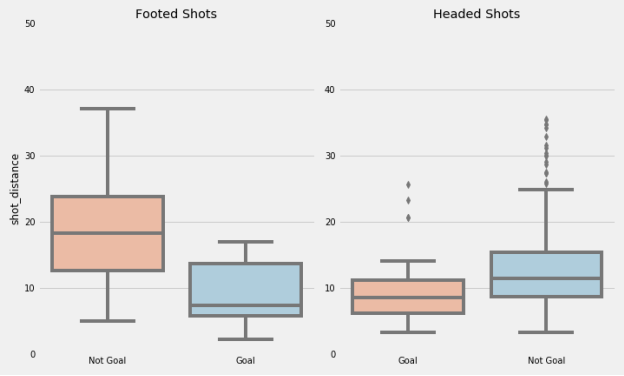

Fig. 7. Box Plots of Headed and Footed shots based on their distance from the goal

Next, we explore the pass play pattern and preceding passes. Looking at the most common types of Preceding passes, we observe from Fig. 8 that Ground pass is the most common type of preceding pass followed by High Pass. The term Low Pass indicates that the pass is below shoulder level, High pass indicates that the pass is above shoulder level and ground pass indicates that the ball travels around the ground. Let us look at the conversion rates of each of these pass types.

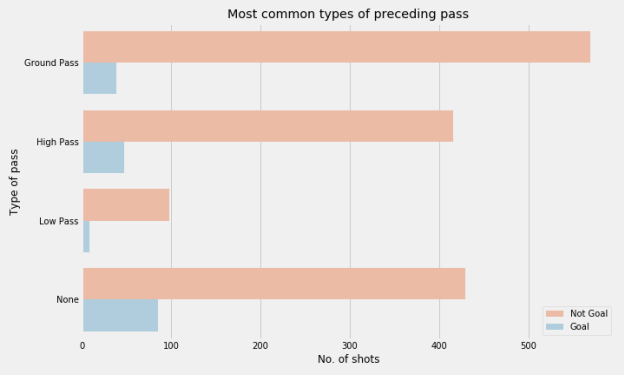

Fig. 8. Common types of preceding passes

For low passes the conversion rate is 8%, for high passes it is 10% and for ground passes it is 6%. This is again a surprising result as the least number of goals were observed from low passes, but the conversion percentage for them is higher than the ground passes. On examining further using box plots based on their distance from the goal, see Fig. 9. From these box plots we observe that High passes are made closest to the goal and ground passes are made away from the goal. This makes sense, as high passes are usually meant for headers and we have seen above that headers are taken from within the penalty box. Ground Passes are fed at the feet of the players and we have seen that footed shots can be taken both within the penalty box and outside it, which was seen in the plots above.

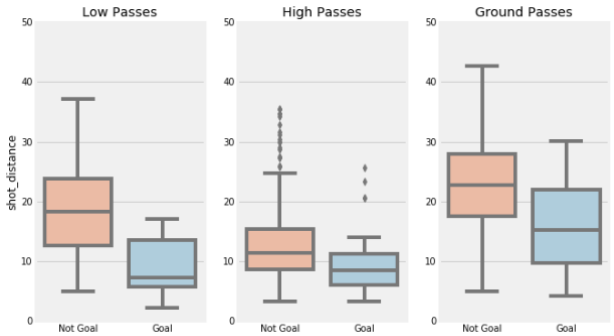

Fig. 9. Common types of preceding passes

Next we try to find the important types of key passes in our dataset. Looking at the key passes pattern and the result of the shot we evaluate the probability of a shot resulting in a goal for each key pass. From Fig. 10, it is clear that the best key passes result from Counter, Keeper, Corners, Throw Ins and Regular play, Among these, it is nearly impossible to directly score from a keepers position, corner or a throw-in but these can prove to be the best types of key passes.

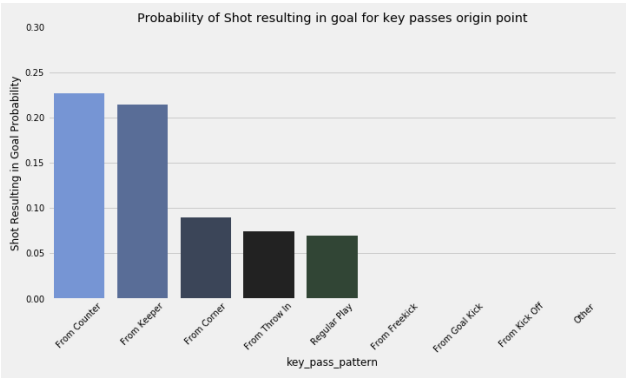

Fig. 10. Probability of shot resulting in goal for an origin point of a key pass

Using the first time feature of StatsBomb, we analyze if a shot that is the end result of a carry is more successful than a shot that is taken first time by a player. A first time shot is usually a shot that is taken without controlling. It cannot

be a result of a player's carry. We observe that the shots that are taken the first time have a slightly higher probability of resulting in a goal. On observing the distance from where they are taken, it was seen that most shots that were taken within the 15 yards mark were likely to end up in a goal and the first time shots that were taken further than this distance were less likely to be successful.

Looking into the shot angles variable from our dataset, it was observed from a box plot that the shots with low angles result in lesser goals and higher the shot angle, higher are the chances of that shot ending up in a goal. This result can be linked with the distance result where in lower angle indicates that the shot was from a larger distance from the goal and higher angle indicates that the shot was from a shorter distance from the goal.

Next, we looked into the success factors of the open play shots, penalties and free kicks and found out that the shot conversion rate for open play shots is 8%, for penalties it is 72% and for the free kicks it is 7%. Penalties correspond to a very high probability of scoring, free kicks on the other hand are not as successful, open play shots are easier to score than the free kicks.

Pack density also affects the success factor of a shot. It is observed that unsuccessful shots have more players in between the shot and the goal. Next considering the Non-Play factors like minute of the shot and Home team or away team, we observed that the chances of shot ending up in a goal is equal for these variables and can conclude that these variables are insignificant in terms of determining if a shot taken is successful or not.

3) Building the Model:

After performing the exploratory analysis, the next step is building the model. For that we need to select the features which could prove to be important. Columns key pass type, isCross and iSCutback are combined into 1 feature in order to make it easier to encode as a categorical feature. Since there are only 6 cutbacks in the dataset and not even one of them result in a goal, it makes sense to drop the rows with key pass type as "cutback". The cross_cutback_forward feature now consists of 863 forward passes, 310 crosses and the other 514 are named 'none'. Modifying the key_pass_pattern feature similarly, we group the samples into Regular Play, Non Regular Play and None. There are 656 shots in Non Regular Play, this is a combination of all the pass patterns other than the ones named 'None' and 'Regular Play'. There are 517 shots from Regular Play and 514 shots from None type.

Next, looking at the type name variable, there are 1543 shots from open play, 80 shots from free kicks and 64 shots from penalties. Looking at the sample size of each type it makes sense to ignore the Free kicks and penalties and focus only on shots that are not direct ones. As we have already seen, the variables timestamp and home_or_away are insignificant in determining the outcome of a shot. Thus, it makes sense to drop these columns. As it was difficult to clearly predict the outcome for the first_time feature, we drop this column as well. Looking at the body_part_name feature and the number of occurrences, we see that the occurrences of the body parts other than head and foot are just 14 compared to 1219 shots from Foot and 310 shots from Head. Hence, we drop the shots that are from body parts other than head or foot.

After, modifying all these features we check if there are missing values in any of these features. As there are numerical and categorical columns in our data frame, we will be using 'ColumnTansformer' from sklearn to encode the categorical columns and scale the numerical columns. Next we split our data into Training and Test data sets. The split percentage is 70-30 %. As the column, 'statsbomb_xg' is not a predictor feature it is removed from test and train data sets, it is extracted and kept aside for comparisons. Next, we try out different models and compare our results with the StatsBomb model.

C. Creating a model to predict the goals scored

Using the shots information from our dataset, I plan to create a simple model that predicts the number of goals scored from the shots that are present in the data set. In this model, I plan to make use of the insights that I observed from the xG model that was explained in the previous section while selecting the features.

1) Data Collection and Preprocessing:

The competitions/ matches/ lineups/ events data that is made public by StatsBomb is obtained using the StatsBomb package[16] in python. This package converts the json formatted data into an easy to use format. Competitions data is extracted and stored into a data frame. From this data frame, we will pick the data for World Cup 2018, that corresponds to the competition_id '43' and season_id '3'. Using these, we obtain extract data of all the matches from the World cup and store them in wc_matches data frame. As the main aim of this model is to calculate the number of goals, we extract the shot information from each match using the events data files and store this information in the shots_df. We have 1706 shots in our shots_df.

2) Data Exploration:

In our shots_df, we remove the shots that are penalties as they might mislead our analysis. Also, in the previous model we discovered that conversion of penalties is 72% compared to 8% from open play shots and 7% from free kicks. So, it makes sense to remove the penalties from our data.

After removal of penalties, we create a goal column which will store the outcome of a shot. 0 indicates that the shot did not result in a goal and 1 indicates that the shot was a goal. We see that the average shot conversion rate is 8.24%. From Fig. 11, we can observe that the amount of shots that do not end up as goals clearly dominate the shots that end up as goals.

3) Selecting features for the Model:

We create a few new variables out of the data that is available to us, in order to improve our model. We create a new variable called distance, by using the distance formula and angle from center of the goal(at (120, 40)) to the location of the shot. We replace the 'Right Foot' and 'Left Foot' shots as 'foot' based shots. 'play_pattern', 'under_pressure',

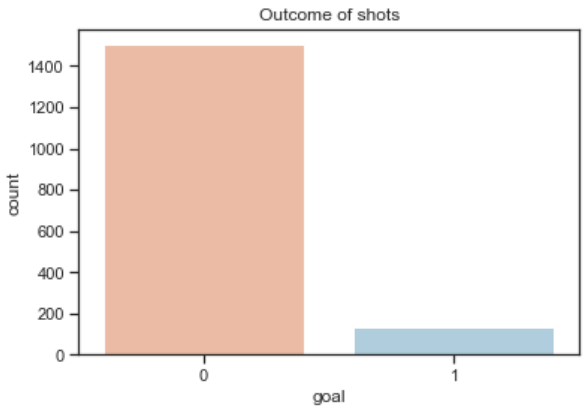

Fig. 11. Outcome of all shots from World Cup 2018

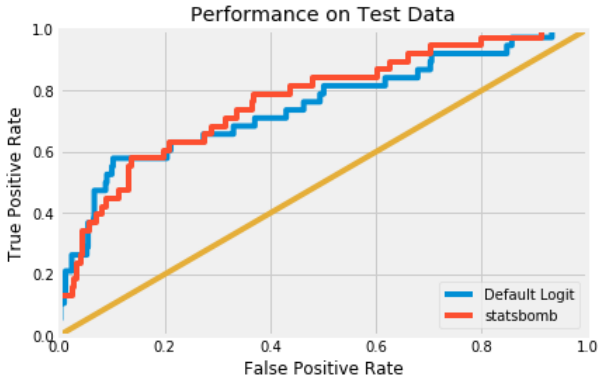

Fig. 12. ROC Curve to compare performance of Logistic Regression Model and StatsBomb Model

'body_part', 'technique', 'first_time', 'follows_dribble', 'redirect', 'one_on_one', 'open_goal', 'deflected', and 'distance' are the columns that we select as our features. We store all the data of these feature columns in the features data frame and their outcomes in the labels data frame. Next, we fill all the NA values with 0.

4) Model Creation:

Now, we have all the important information we need. We look at the categorical features and perform label encoding to make all categorical features to discrete values. This is done so that we can work with a consistent dataset. We split our model features and labels into training and testing sets. The split is 80-20 %. We are using a Decision Tree Classifier from sklearn to train our model and make predictions.

IV. Results

I was able to form a list of Best XI based on their performances in the 2017/18 season. The Players that made it to the Best XI and World Cup XI are listed in table 1. Next, I compared players from both the teams and analyzed how their world cup was using advanced visualizations like radar charts, heat maps, pass maps and shot maps.

TABLE II
Summary of results obtained for each Model

Model Name	AUC Score - Train	AUC Score - Train
StatsBomb	0.81	0.77
Logistic Regression	0.81	0.751
Gradient Boost	0.96	0.733
XGBoost	0.99	0.732

I computed the xG using Logistic Regression, Gradient Boost models and XGBoost and used the Area under the curve(AUC) score as my evaluation metric[3]. The results of all these models are compared with the StastBomb xG prediction. Hyper- parameter tuning was performed for each of these classifier models. Results are summarized in the table II.

We observe that the Logistic regression model performs the best as the AUC score was 0.751 on the test data, which was closest to the AUC score of the StatsBomb predicted model. From the Fig. 12, we observe that the curve for both models is almost similar as they almost overlap each other. Since, 1543 shots are a less number, a dataset consisting of more shots data would yield interesting results. In order to understand which features were significant for our models, we observed from the Fig 13 that shot distance, shot angle and pack density. The most important categorical features were body part name and the type of pass that preceded the shot. We also observe that XGBoost and Gradient Boost do not consider some of the variables, we can see that Logistic regression gives small coefficients to these variables.

Using the best Model, I extract the goals, xg_sum and xg_diff information for individual players, to evaluate which players had the best and worst World Cup in the terms of their xg values. Fig. 14 shows a list of players who had the best and worst world cups in terms of their xG difference which is the difference between the total xG and the goals scored. We see that there is Denis Cheryshev from Russia who was the standout goalscorer in this dataset with the highest xg difference.

Yerry Mina who is a defender also shows up in the list. He scored 3 goals and was a prominent threat from set pieces throughout the world Cup. We also see that the strikers Mbappe, Lukaku and Kane who had a great world cup show up in the list of players with best xG difference. Looking at the worst xG differences Sterling, Giroud and Neymar and Marcus Berg who are all forwards had a negative xG difference and under-performed with respect to their cumulative xG, which indicates that they should have scored more compared to the number of shots they took. They were just unlucky. Also, these shots do not include penalties and free kicks.

Next, let us look at the results that we obtained for our Decision Tree Classifier that predicts the number of goals based on our dataset. From fig. 15 we can see that the model predicts 25 goals and the actual goals in the test data set are 23. Next, we observe that the difference between the xG value predicted by model and the actual goals is 3.25. Precision here indicates that out of all the goals the modelled claimed to be

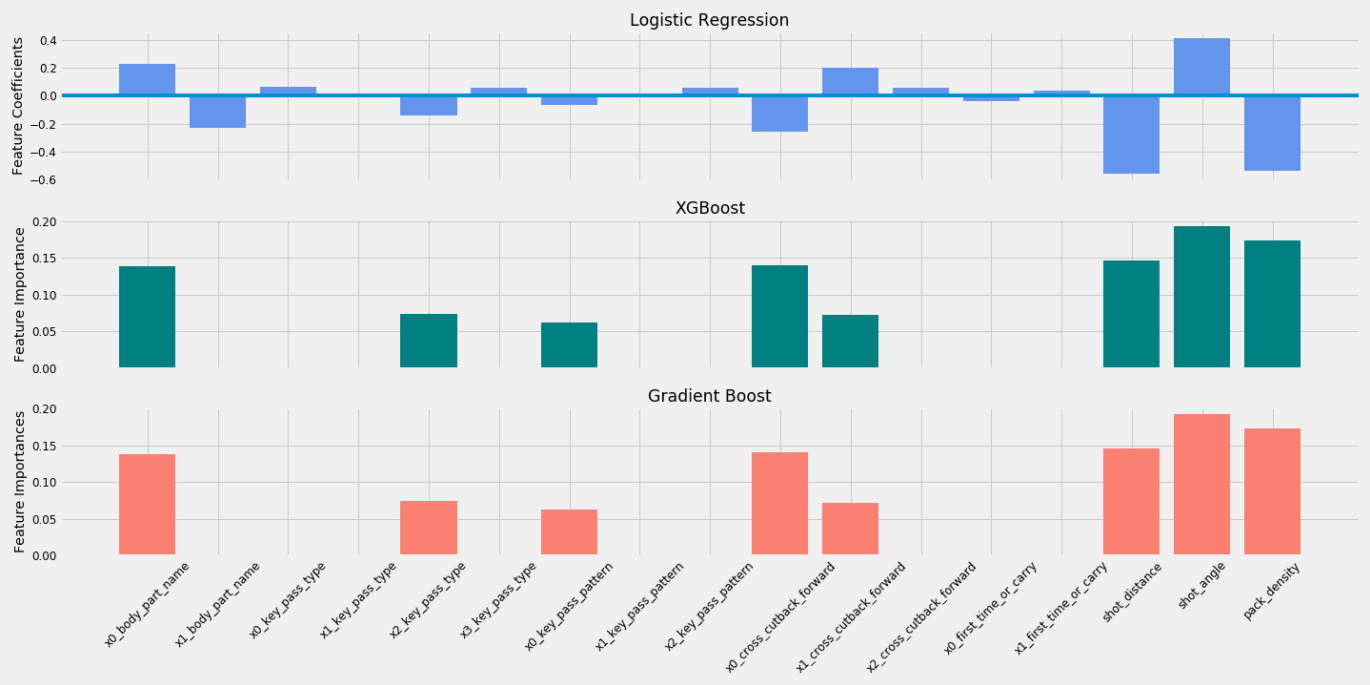

Fig. 13. Important features for each model

shot_player	goals	xG_sum	xG_Diff
Denis Cheryshev	4	1.1	2.9
Yerry Fernando Mina GonzÃ¡lez	3	0.5	2.5
Kylian MbappÃ© Lottin	4	1.7	2.3
Harry Kane	3	1.1	1.9
Romelu Lukaku Menama	4	2.3	1.7

shot_player	goals	xG_sum	xG_Diff
Mats Hummels	0	0.7	-0.7
Olivier Giroud	0	1.3	-1.3
Raheem Shaquille Sterling	0	1.3	-1.3
Neymar da Silva Santos Junior	2	3.4	-1.4
Marcus Berg	0	2.5	-2.5

Fig. 14. List of players with Best and worst xg difference

```
Predicted goals from test data: 25
xG: 26.25
Actual goals (from test set): 23
Difference from my xG value and actual goals =  3.25

              precision    recall  f1-score   support

           0       0.94      0.94      0.94       305
           1       0.24      0.26      0.25        23

    accuracy                           0.89       328
   macro avg       0.59      0.60      0.60       328
weighted avg       0.89      0.89      0.89       328
```

Fig. 15. Results obtained for the Decision Tree Classifier

a goal, how many were actually goals? Recall indicates that out of all the actual goals, how many did the model predict to be a goal? The accuracy of our model is determined by F1 score, which is the average of Precision and Recall. Higher the F1 score, the better that model is.

Fig. 16 indicates the most important features from the model. The distance feature, which was formed by combining the variables shot_distance and shot_angle is the most important Feature in this model. Other important features are

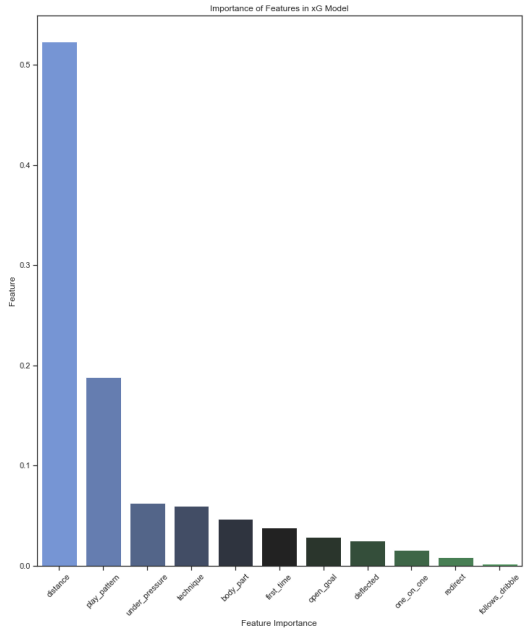

Fig. 16. Importance of features in the Decision Tree model

play_pattern, under_pressure and technique.

V. Future Work and Conclusion

At this stage, I can say that I was successfully able to create a Best XI for the season based on their performances over

the 2017/18 season. I was then able to compare the list of players from my Best XI with the World Cup XI and analyze how the world cup went for each of them using advanced visualizations. Then I created 3 models that predicted the xG for a shot and observed that the Logistic Regression model generated the best results with AUC score of 0.77 for the test data, StatsBomb xG model had an AUC score of 0.77 on the same test data. I was then able to use this model to extract the individual player information to have a look at the players who had the best and worst shot conversions(xG difference). Next, I created a Decision Tree classifier model which predicted the number of goals scored based on the shots in the data set. I used an F1 score to evaluate the accuracy of this model and achieved an F1 score of 0.89.

As a future scope, I plan to carry ahead the work from this project into the other datasets available in the open StatsBomb[9] repository. I want to make use of advanced visualization like I did in this project to carry out in depth player analysis and carry forward this work for team analysis[2] as well. I want to evaluate different strategies for various formations and use my insights to pick the team for Fantasy Premier League every game week. I also plan to simulate a soccer game where I can pre define the player attributes and behaviours in advance and carry out an actual game. This concept is used in RoboCup soccer games[10], where robots are used as players and a soccer game is carried out.

ACKNOWLEDGMENT

I would like to thank my advisor Prof. Carol Romanowski who guided me throughout this project. It was due to her constant support and weekly feedback that I was able to successfully complete this project. I would also like to thank the Computer Science Department at Rochester Institute of Technology, this project has only been possible due to the knowledge that I have acquired while studying at RIT. Last but not the least, I would like to thank StatsBomb for providing the data and resources that were really helpful in successfully completing my capstone project.

REFERENCES

[1] 25 johan cruyff quotes that will change the way you think about football - paste. https://www.pastemagazine.com/soccer/football/25-johan-cruyff-quotes/.

[2] Advanced sports visualization with python, matplotlib and seaborn. https://towardsdatascience.com/advanced-sports-visualization-with-pandas-matplotlib-and-seaborn-9c16df80a81b.

[3] Classification: Roc curve and auc — machine learning crash course. https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

[4] A data driven goalkeeper evaluation framework. http://www.sloansportsconference.com/wp-content/uploads/2019/02/Data-Driven-Goalkeeper-Evaluation-Framework.pdf.

[5] Dyche's long ball tactics should be lauded - football news, views & transfer rumours — football whispers. https://www.footballwhispers.com/blog/dyches-long-ball-tactics-lauded/.

[6] Expected goals — how i combined learning data science with my soccer obsession. https://medium.com/analytics-vidhya/expected-goals-how-i-combined-learning-data-science-with-my-soccer-obsession-f81d721432c7.

[7] Fifa world cup 2018: Official team of the tournament is out, four france players feature. https://www.timesnownews.com/fifa-football-world-cup-2018-russia/article/fifa-world-cup-2018-official-team-of-the-tournament-is-out-four-france-players-feature/255889.

[8] Football formation creator - footballuser.com. http://www.footballuser.com/post.

[9] Free football data from statsbomb. https://github.com/statsbomb/open-data.

[10] grsim – robocup small size robot soccer simulator — springerlink. https://link.springer.com/chapter/10.1007/978-3-642-32060-6_38.

[11] How liverpool fc is using data science to dominate british premier league. https://www.zmescience.com/science/how-liverpool-fc-is-using-data-science-to-dominate-british-premier-league/.

[12] The numbers game: Why everything you know about soccer is wrong - christopher anderson, david sally - google books. https://books.google.com/books/about/The_Numbers_Game.html?id=0LWKDQAAQBAJ&source=kp_book_description.

[13] Player stats database fifa 18 fifa index. https://www.fifaindex.com/players/fifa18/?name=.

[14] Point in triangle test. https://blackpawn.com/texts/pointinpoly/.

[15] Soccer teams— espn. https://www.espn.com/soccer/teams/_/league/FIFA.WORLD/fifa-world-cup.

[16] statsbomb · pypi. https://pypi.org/project/statsbomb/.

[17] xg explaination. https://fbref.com/en/expected-goals-model-explained/.

[18] J. F. Luke Bornn, Dan Cervone. Soccer analytics: Unravelling the complexity of "the beautiful game" - bornn - 2018 - significance - wiley online library. https://rss.onlinelibrary.wiley.com/doi/full/10.1111/j.1740-9713.2018.01146.x, May 2018.