Predicting The Outcome Of A Shot at Goal

1.0 Problem Statement

- To predict whether a shot in a game of football will result in a goal or not ?

Machine learning and statistical analysis in this field is a relatively new concept. By carrying out this investigation we can gain a greater understanding of its capabilities and potential applications giving football clubs the capacitive edge.

2.0 Data

The data collected was obtained using Statsbomb's free demo dataset found here. It contains data recovered from over 800 games across 6 different leagues and is provided as JSON files exported from StatsBomb's official Data API. Each individual game is stored as a list of dictionaries where each dictionary represents an event within the game. Figure 1 gives an event instance which in this example was a miscontrol.


```
{'id': '2c0070b5-6e05-438c-a26a-e7a87e840823',
 'index': 11,
 'period': 1,
 'timestamp': '00:00:06.740',
 'minute': 0,
 'second': 6,
 'type': {'id': 38, 'name': 'Miscontrol'},
 'possession': 2,
 'possession_team': {'id': 971, 'name': 'Chelsea LFC'},
 'play_pattern': {'id': 9, 'name': 'From Kick Off'},
 'off_camera': False,
 'team': {'id': 971, 'name': 'Chelsea LFC'},
 'player': {'id': 4634, 'name': 'Crystal Dunn'},
 'position': {'id': 16, 'name': 'Left Midfield'},
 'location': [108.0, 10.0]}
```

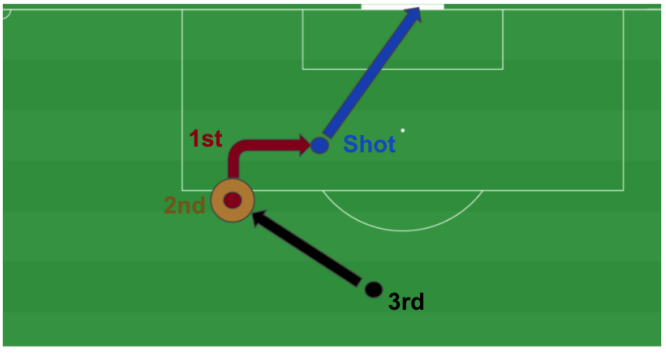

Figure 1 - Event JSON

3.0 Strategy

As well as evaluating the shot characteristics I also analysed the events which occurred leading up to it. All events were considered ranging from a pass or ball receipt to tactics or formation changes. Events as early as 3 prior to the shot were collected into individual data frames and concatted together on their matching game id's. Json_normalize was utilised to form the dataframes and creates a column for every eventuality present across the 800 games for each of the 4 instances. Figure 2 below gives a visual representation of the approach used. All shots and events are manipulated so they're in the same direction for comparison.

Instance	Event	Timestamp
Shot	Shot at Goal	15:06:01
1st	Ball Carry	15:00:41
2nd	Ball Receipt	14:59:16
3rd	Pass	14:57:11

Figure 2 - Event Instances



4.0 Additional EDA

4.1 Area of Goal

Figure 3 below is a KDE plot of all the shots at goal resulting in goals. The goal was split into 6 equal sections and the shot was categorised based on its end location. Any shot outside of this was recorded as OFF_TARGET. The lighter the colour the more goals scored in this region. As expected considerably less goals were scored in the middle region. Bottom Left had the most successful shots and may be down to most players being right footed and therefore the left would be their strongest side to shoot at.

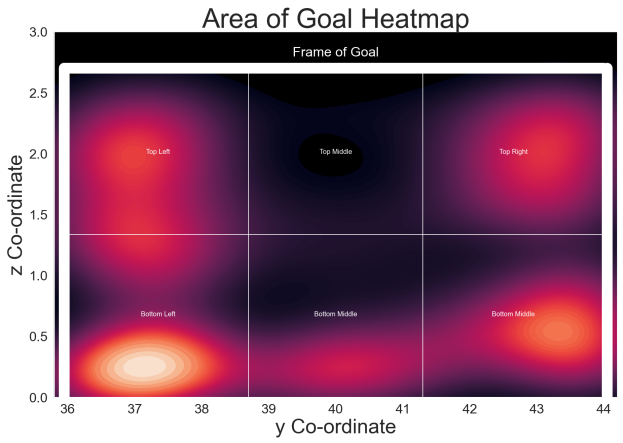


Figure 3 - KDE Plot of Successful Shots in Goal

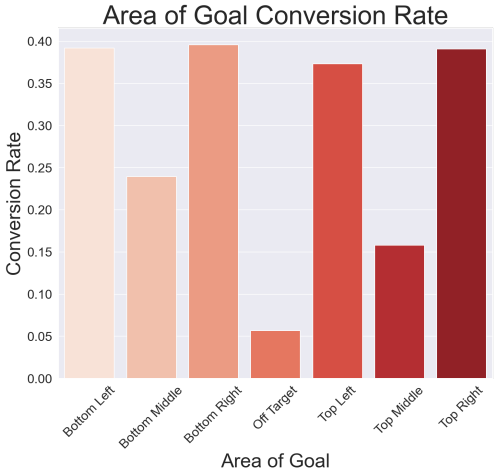


Figure 4 - Bar Plot of Area of Goal Conversion Rates

Figure 4 shows the conversion rates of the 6 regions of the goal. Aiming for either of the corners can be seen to achieve a considerably higher conversion rate. Interestingly shots off target have managed to result in a conversion rate of just over 5% and may be down to goals going in off the frame or deflections. Overall there doesn't seem to be much fluctuation in conversion rates between the 4 corners of the goal.

4.2 Distance From Goal

Figure 5 gives a physical representation of the distance from goal. It is given by distance between the shot location and the centre of goal at (120,40).

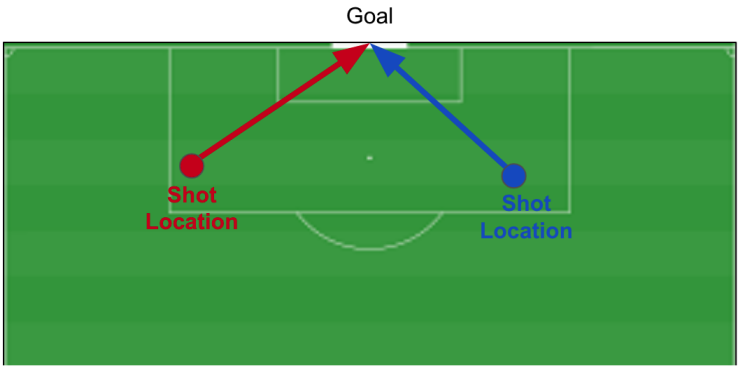


Figure 5 - Distance From Goal

Figure 6 can be seen to be a bar plot of the respective distances seen in figure 7. Distances were split into bins of width 10 for comparison. In general as the distance from goal increases, the conversion rate decreases. Apart from a significant drop seen at 20<30. 0<10 has a significantly larger conversion rate in comparison and may be down to the fact this region will

include penalties.

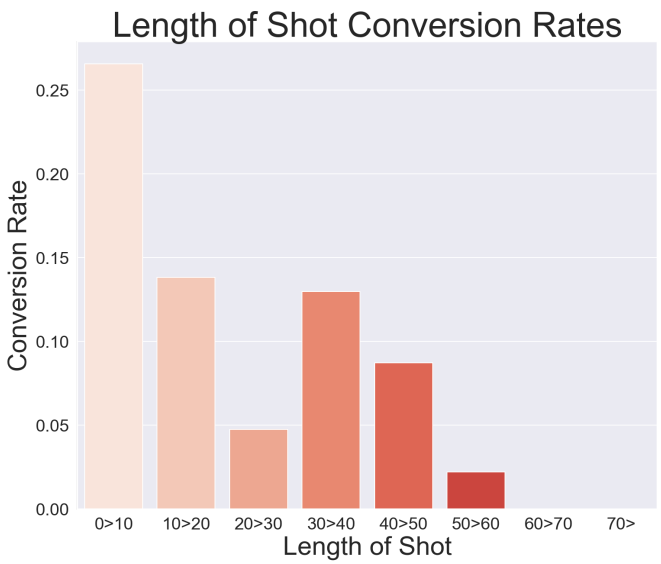


Figure 6 - Distance From Goal Conversion Rates

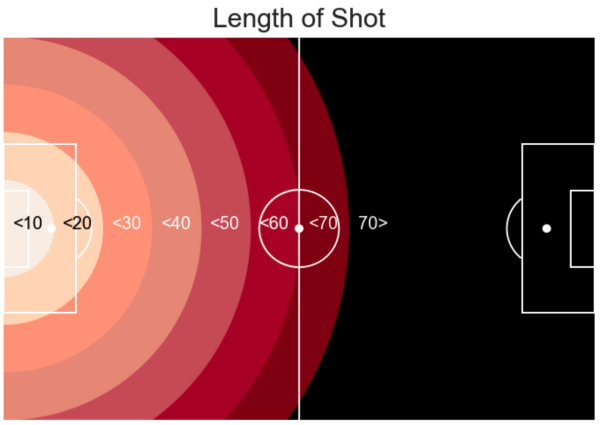


Figure 7 - Binned Distances From Goal

4.3 Distance From Goal

Figure 8 gives a physical representation of the angle to goal. It is the angle between the black centre line and the blue or red connecting line from the shot location to the centre of the goal.

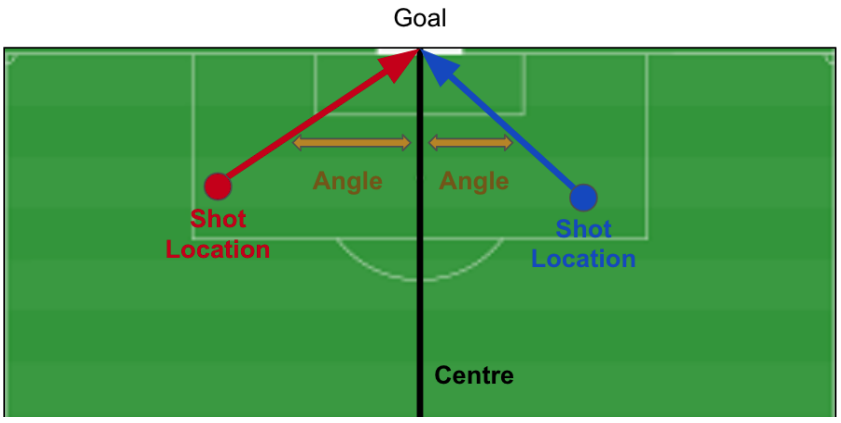


Figure 8 - Angle To Goal

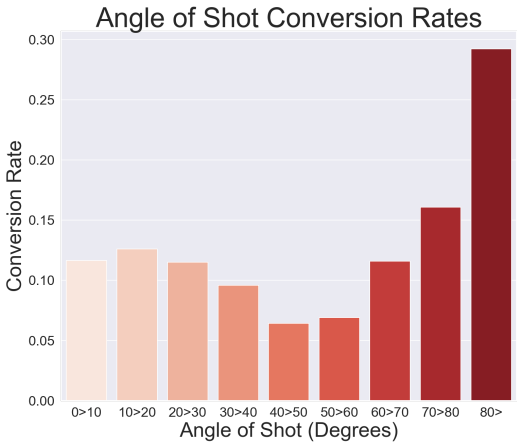


Figure 9 - Angle To Goal Conversion Rates

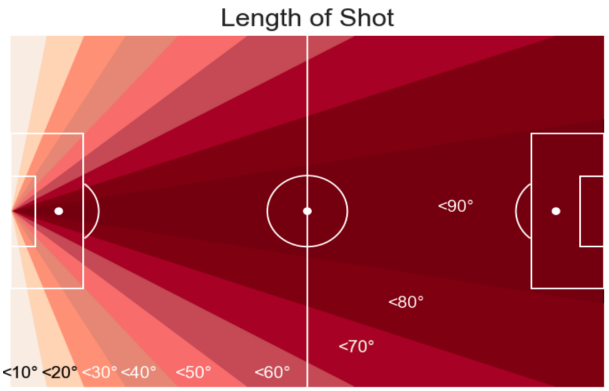


Figure 10 - Binned Angles of Width 10

Figure 9 gives a bar plot of the conversion rates of the binned shot angles seen in figure 10. There seems to be little correlation between the shot angles and conversion rates. Greater

than 80 degrees is significantly greater and may also be down to the fact this region will include penalties. 40 to 50 is said to have the lowest conversion leaving more scope for further feature variables potentially regarding whether the shot was aimed at the near post or not. Splitting the pitch into a grid would also allow for taking shot distance into consideration as well as the angle. Figure 11 shows a shot location heat map for all the successful shots taken. Thus demonstrating an extremely high correlation between location and number of successful shots. Most goals were scored in between the edge of the six yard box and the penalty spot directly in front of goal. The investigation may be improved with removal of penalties as their conversion rate is much higher than an ordinary shot at goal.

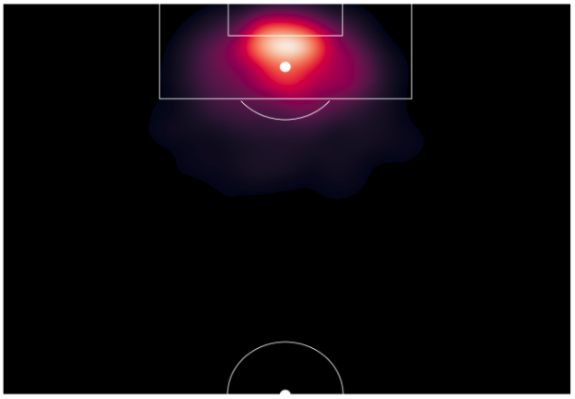


Figure 11 - Successful Shot Location Heatmap

4.4 Defenders In Front of Goal

Figure 12 gives a visual representation of the defenders in front of goal. A triangle is formed between the shot location and each post. All opposition players within this apart from the goalkeeper contribute to value and was carried out using Shapely.

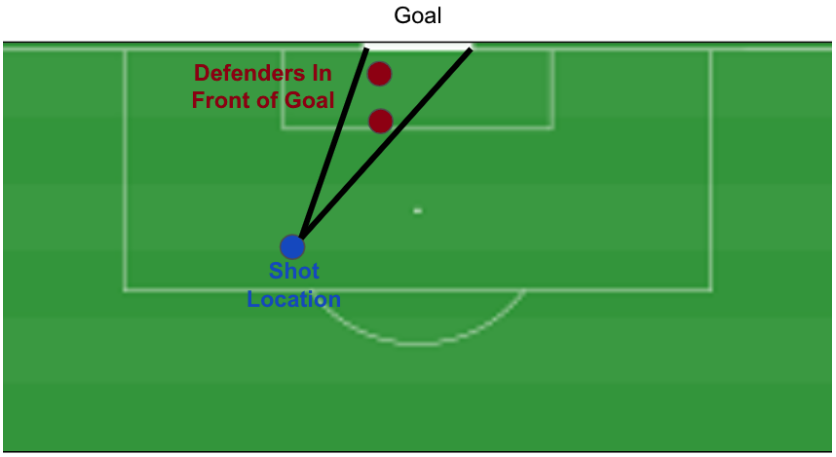


Figure 12 - Defenders In Front of Goal

Figures 13 and 14 give the count and conversion rates and for defenders in front of goal. As expected the fewer defenders in front of goal the more shots were taken. However in terms of conversion rates it can be seen to roughly decrease from 0 defenders to 4. After this however the conversion rate becomes skewed and increases with the number of defenders. This is because after 5 defenders a minimal amount of shots were taken therefore any successful shot within this criteria result with an unexpectedly high conversion rate. For example only one shot was taken with 10 defenders in front of goal which resulted in a goal. Thus giving it a 100% conversion rate.

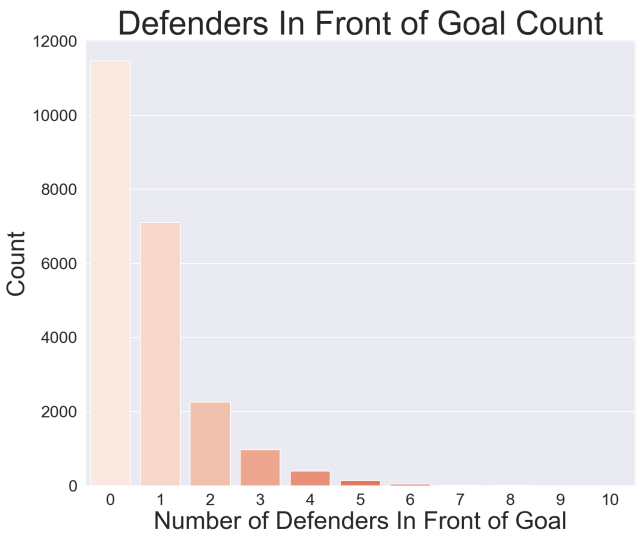


Figure 13 - Bar Plot of Number of Shots for Defenders In Front of Goal

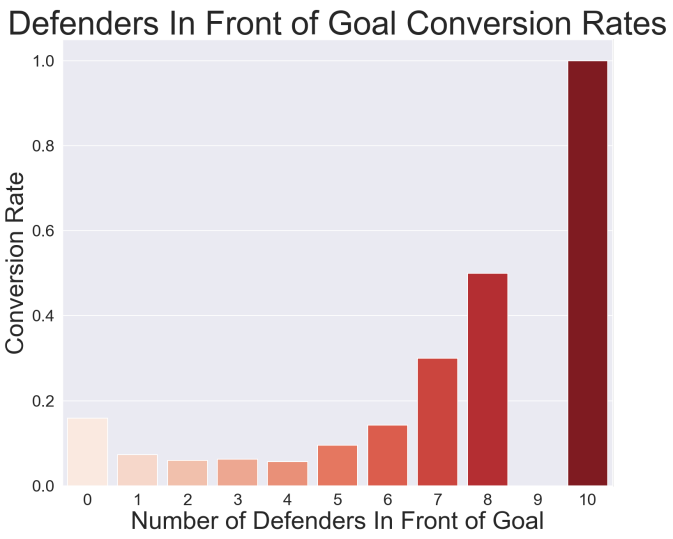


Figure 14 - Bar Plot of Conversion Rates for Defenders In Front of Goal

5.0 Models

Figure 15 below gives a bar plot of a comparison between the mean cross validation scores for all the models tested in this investigation. Out of the 22731 recorded shots only 2582 were successful yielding a baseline accuracy of 0.886, seen in white. From the models seen below logistic regression was seen to achieve the highest mean cross validation score of 0.9529 roughly 0.07 above baseline. Gradient boost was found to achieve the lowest of around 0.91 only 0.03 above baseline

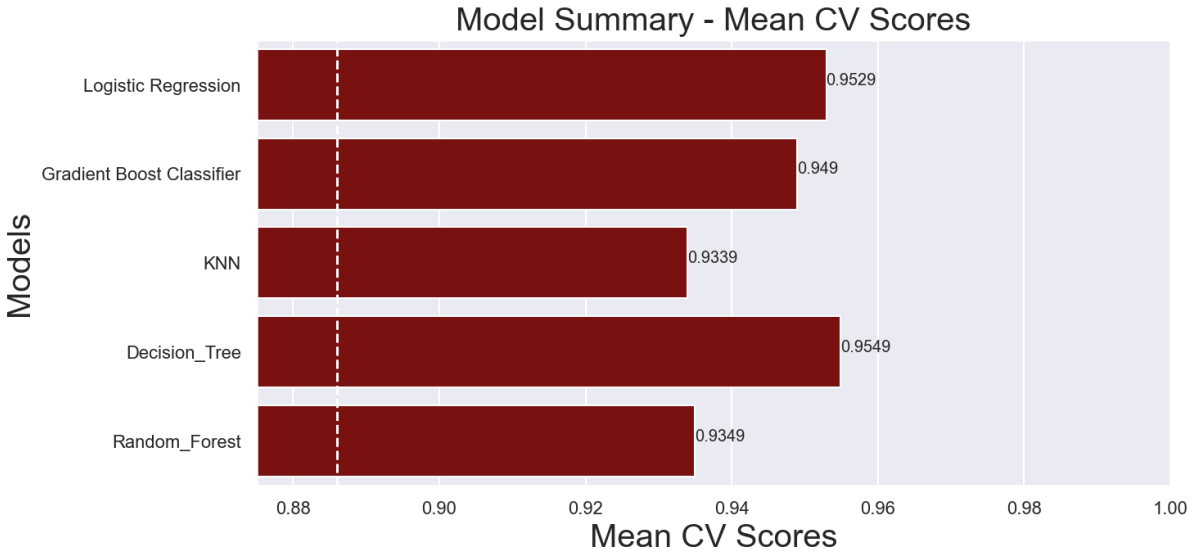


Figure 15 - Mean CV Score Model Comparison

5.1 Decision Tree Classifier

Figure 16 below gives the 15 most important features determined by the decision tree classifier. Full list of feature meanings can be found below from most to least important.

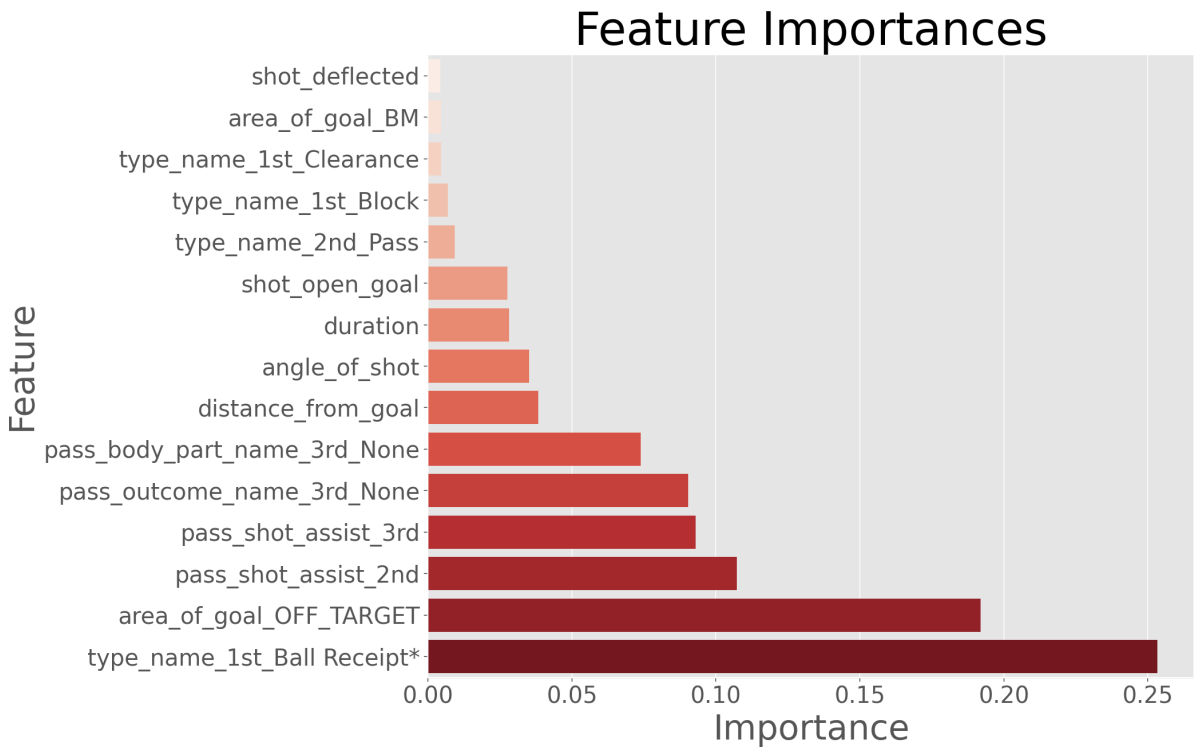


Figure 15 - Mean CV Score Model Comparison

Position	Feature	Description
1	**type_name_1st_Ball_Receipt**	First instance was a ball receipt
2	**area_of_goal_OFF_TARGET**	Shot was off target
3	**pass_shot_assist_2nd**	Second instance was an assist to a shot
4	**pass_shot_assist_3rd**	Third instance was an assist to a shot
5	**pass_outcome_name_3rd_None**	Third instance wasn't a pass so no outcome
6	**pass_body_part_name_3rd_None**	Third instance wasn't a pass so no body part
7	**distance_from_goal**	Distance from the centre of goal from shot location
8	**angle_of_shot**	Angle of the shot taken
9	**duration**	Time taken for shot instance
10	**shot_open_goal**	Shot taken at open goal (no goalkeeper)
11	**type_name_2nd_pass**	Second instance is a pass
12	**type_name_1st_block**	First instance is a block
13	**type_name_1st_clearance**	First instance is a clearance
14	**area_of_goal_BM**	Shot at bottom middle section of goal

15	**shot_deflected**	Deflected shot

6.0 Conclusion

By carrying out this investigation it has allowed me to further understand the capabilities of statistical analytics and machine learning in the beautiful game. It was determined all areas of the goal apart from the middle had relatively similar conversion rates despite the bottom left corner being the most populated with successful shots. In terms of distance from goal as expected there was a relatively clear inversely proportional relationship between distance from goal and conversion rate. The same couldn't be said for angle to goal. Shots taken in the 40 < 50 and 50 < 60 degree region had considerably lower conversion rates than the other 10 degree width regions. This leaves scope for further work on developing a more accurate feature variable which better represents location on the pitch. Potentially splitting the pitch into a hexagonal grid as a circle would greater represent a players location. This would leave spaces between when in a grid. As expected the number of defenders in front of goal dramatically decreased the number of shots taken. However the conversion rates for this were skewed. Few shots were taken (as little as 1) for large numbers of defenders in front of goal. Therefore 1 un-expected converted goal has a negative effect on the validity of the results.

In this investigation 5 different classification models were tested and tuned using cross validation grid searches. It was determined that Decision Tree Classifier had the greatest predicting power yielding a mean score of 0.9549, 0.07 greater than baseline. Thus highlighting the potential of machine learning when improving teams attacking outputs. The first instance being a ball receipt was found to have the greatest feature importance followed by a shot being off target and the second instance being a pass for a shot assist.