II FOOTBALL DATA INTELLIGENCE COURSE

EVALUATION AND ANALYSIS OF EXPECTED ASSIST PATTERN ON CORNER KICK DEVELOPMENTS

SPEAKERS

Michele Crestani

Aldo Comi

Antonio Gagliardi

Roberto Angioni

Francesco Bordin

CANDIDATE

Salvatore Nizza

26 September - 28 October 2022

INDEX

FOREWORD 3

1. INTRODUCTION 3

2. STATISTICAL DATA 6

3. DATASETS USED 8

4. UNFOLDING 8

4.1 STEP 1 - DATA PREPARATION 8

4.2 STEP 2 - CLASSIFICATION 9

4.2.1 GROUP A 9

4.2.2 GROUP B 9

4.2.3 GROUP C 10

4.2.4 GROUP D 10

4.3 STEP 3 - MODEL IMPLEMENTATION 12

5. DATA ANALYSIS 13

5.1 MODEL EVALUATION 13

5.2 RESULTS 19

6. COMPARISON 20

7. FINAL CONSIDERATIONS 24

FOREWORD

The purpose of this study is to evaluate and analyse an expected assists model obtained from a logistic regression model looking at the 2021-2022 Serie A season in order to understand in which corner kick situations there is an offensive benefit.

1. INTRODUCTION

As stated in the Regolamento del Giuoco del Calcio accompanied by the FIGC Official Decisions and the AIA Practical Guide, *"a corner kick is awarded when the ball, last touched by a player of the defending team, has completely crossed the goal line, either on the ground or in the air, without a goal having been scored"*.

The corner kick is a key game situation within a football match that can allow a goal to be scored from a dead ball and have an important influence on the outcome of the match.
Just to mention a few examples:

- Marco Materazzi, 2006 World Cup Final, Italy - France, equalising goal on an assist from Andrea Pirlo to take the team to penalty kicks and win their fourth World Cup title

- Sergio Ramos, Champions League Final 2013-2014, Real Madrid - Atletico Madrid, equalising goal on an assist from Luka Modric to take the team into extra time and win the *'tenth'*

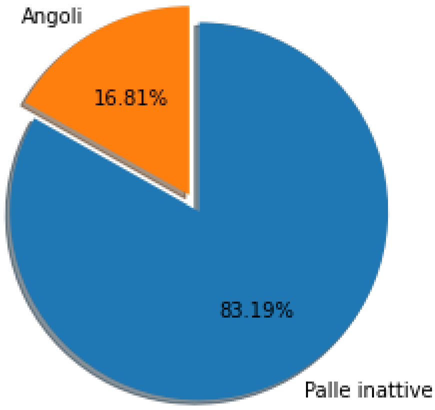
- Kostas Manolas, Champions League 2017-2018 quarter-final return leg, Roma - Barcelona, goal 3-0 on an assist from Cengiz Under and passage of the round

- Matias Vecino, Week 38 of Serie A 2017-2018, Lazio - Inter, winning goal on an assist from Marcelo Brozovic and Inter's return to the Champions League after six years

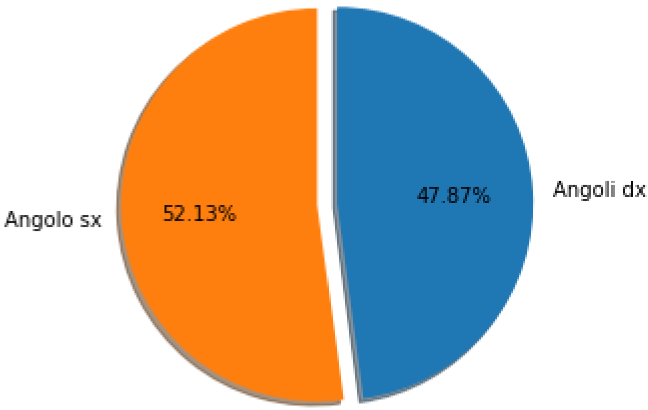
- Divock Origi, 2018-2019 Champions League Semi-Finals Return leg, Liverpool - Barcelona, 4 - 0 goal on an assist from Trent Alexander-Arnold and passage of the round

2. STATISTICAL DATA

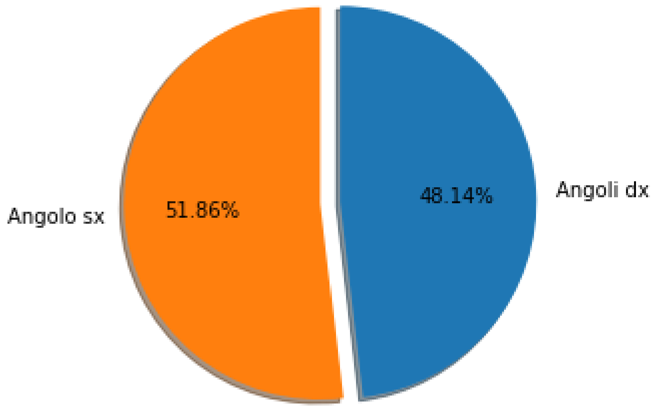
During the 2021-22 Serie A season, events from corner kick situations accounted for 16.81 % (3,654) of the free kicks in favour (18,079) that occurred in the league.



52.13 % (1,905) of the corners occurred from the left side with respect to the direction of attack while 47.87 % (1,749) occurred from the right side with respect to the direction of attack.



On the other hand, considering only the passes that ended up inside the opponent's penalty area, 51.86% (1,547) of the events were from left corner kicks and 48.14% (1,436) from right corner kicks.



Having understood the frequency of corner kicks, the drafting continued.

3. DATASETS USED

Data from the following datasets were used to develop the topic:

- *SICS* - positional data from corner kicks (*SICS-Serie A 2021-22_events*)
- **Transfermarkt** - data on preferred foot (*database_foot_serieA_2021-22*)
- database of player identification numbers (*player_id_serieA_2021-22*)

4. UNFOLDING

The study was divided into the following phases:

- STEP 1: Data preparation
- STEP 2: Data Classification
- STEP 3: Model implementation

4.1 STEP 1 - DATA PREPARATION

Prior to the classification of corner kicks according to the area of the field from which they are taken (right or left side) and according to the foot of the kicker (right or left), which will be examined later in Phase 2, the data were prepared, in particular the union of the dataset concerning the positional data of ***SICS SICS-Serie A 2021-22_events*** and the dataset concerning the information on the preferred foot of each player ***database_foot_serieA_2021-22*** obtained from a scraping operation of the data derived from Transfermarkt.

As the two datasets do not have any key in common through which to perform the union, a third dataset was used concerning the different identification numbers of each player (SICS, Soccerment, Skillcorner) ***player_id_serieA_2021-22***.

It can be seen that the datasets of the identification numbers and Transfermarkt have in common the name of the players to be used as a merging key.

After merging the latter two datasets, a further merger was performed with the dataset '**SICS-Series A 2021-22_events**' using the SICS identification number as the merging key.
Having completed this operation and not considering the ambidextrous players, for whom it was not possible to associate the foot with which the corner kicks were taken, the final dataset was obtained with which to perform the data analysis.

4.2 STEP 2 - CLASSIFICATION

Having obtained the dataset useful for the described study, we moved on to classify the corner kicks according to the area of the field from which they are taken and according to the foot of the batter.
Four groups were classified:

- **GROUP A**: right corner, right foot striker (cross with outward effect)
- **GROUP B**: right corner, left foot striker (cross with return effect)
- **GROUP C**: left corner, left foot striker (cross with outward effect)
- **GROUP D**: left corner, right foot strike (cross with return effect)

4.2.1 GROUP A (crosses with exit effect from the right)

The data were filtered according to:

1. the starting co-ordinates (P0 norm $x > 103$, P0 norm $y < 2$) of the right corner kick pass
2. the arrival coordinates of the pass inside the penalty area ($88.5 \leq$ P1 norm $x \leq 105$, $14 \leq$ P1 norm $y \leq 54$) from a right corner kick
3. right-footed hitters

4.2.2 GROUP B (crosses with return effect from the right)

The data were filtered according to:

1. the starting co-ordinates (P0 norm $x > 103$, P0 norm $y < 2$) of the right corner kick pass
2. the arrival coordinates of the pass inside the penalty area ($88.5 \leq$ P1 norm $x \leq 105$, $14 \leq$ P1 norm $y \leq 54$) from a right corner kick
3. left-footed hitters

4.2.3 GROUP C (crosses with exit effect from the left)

The data were filtered according to:

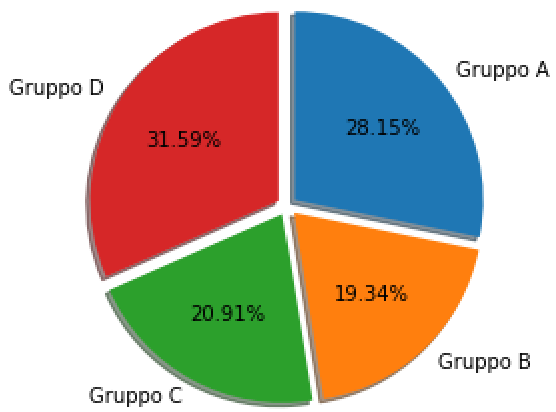
1. the starting co-ordinates (P0 norm x > 103, P0 norm y > 66) of the left corner pass
2. the arrival coordinates of the pass inside the penalty area (88.5 ≤ P1 norm x ≤ 105, 14 ≤ P1 norm y ≤ 54) from a left corner kick
3. right-footed hitters

4.2.4 GROUP D (crosses with return effect from the left)

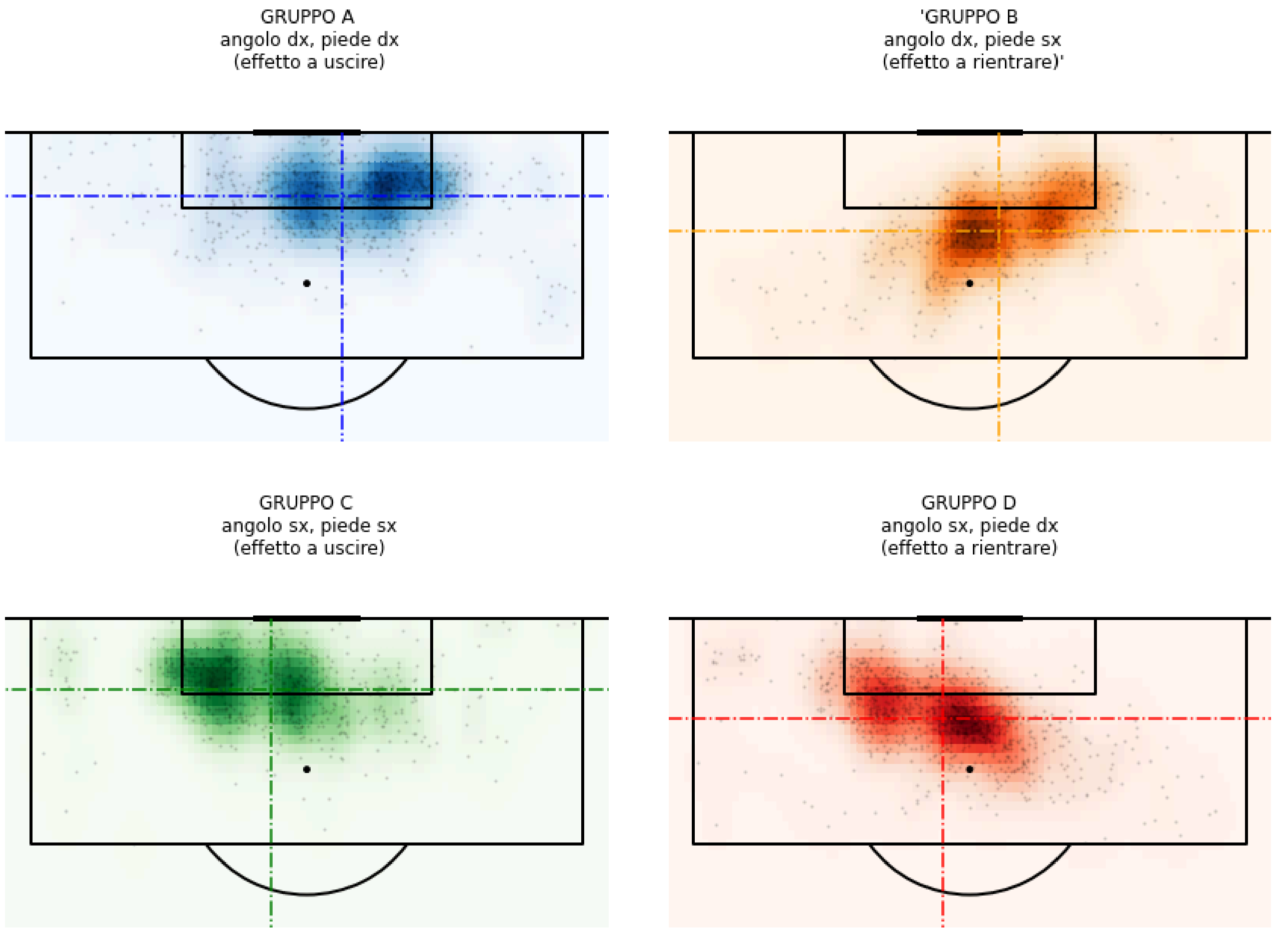
The data were filtered according to:

1. the starting co-ordinates (P0 norm x > 103, P0 norm y > 66) of the left corner pass
2. the arrival coordinates of the pass inside the penalty area (88.5 ≤ P1 norm x ≤ 105, 14 ≤ P1 norm y ≤ 54) from a left corner kick
3. left-footed hitters

As shown in the graph below, Group A accounted for 28.15% (770) of corner kick events, Group B for 19.34% (529), Group C for 20.91% (572) and Group D for 31.59% (864).



Subsequently, the arrival coordinates of the passes made for each of the 4 groups were plotted on a scatterplot in order to visualise through a heatmap which portion of the penalty area has a higher density of balls.



With regard to group A and group C (crosses with an outward effect), most of the passes fall at the apex of the goal area closest to the point where the corner kick is taken and at the longitudinal line of the goal area in a central position.

With regard to group B and group D (crosses with return effect), most passes fall in the central area between the longitudinal line of the goal area and the penalty spot

The horizontal and vertical dotted lines represent respectively the average height of the coordinates of all passes within the penalty area in relation to the long side of the pitch and the average height of the coordinates of all passes within the penalty area in relation to the short side of the pitch.

From a tactical point of view, one might think that such solutions are adopted to prevent the opponent's goalkeeper from intercepting the ball before it reaches the area of potential goalkeeping.

4.3 STEP 3 - MODEL IMPLEMENTATION

After describing the different corner kick solutions by classifying them into four groups, we moved on to the implementation of a simplified logistic regression model of expected assist xA with the aim of estimating which of the different adoptable solutions brings the most benefit in terms of assists.

In creating the model, the assumption was made that the metric is *shot-centric*, i.e. that the assist depends solely on whether the hitter actually received the pass.

In light of the above consideration, it can be assumed that the expected assist model is comparable to the expected goal model. Therefore, assumptions that are valid for an expected goal model will be adopted when creating the expected assist model (the starting co-ordinates of the passes will not be considered within the model, as all the events taken into account derive from corner kicks, a common point for all passes).

To implement the model, a target variable and predictor variables were first identified.

As a target variable, a new column was created for assists to which, for each situation, a value of 1 was associated in the case of a pass that reached the hitter or a value of 0 in the opposite case (if no shot was recorded).

Instead, the distance to the centre of the door, i.e. the hypotenuse of the triangle formed by the coordinates of a point, and the door angle, i.e. the angle implied by the width of the door for each point, were considered as predictor variables.

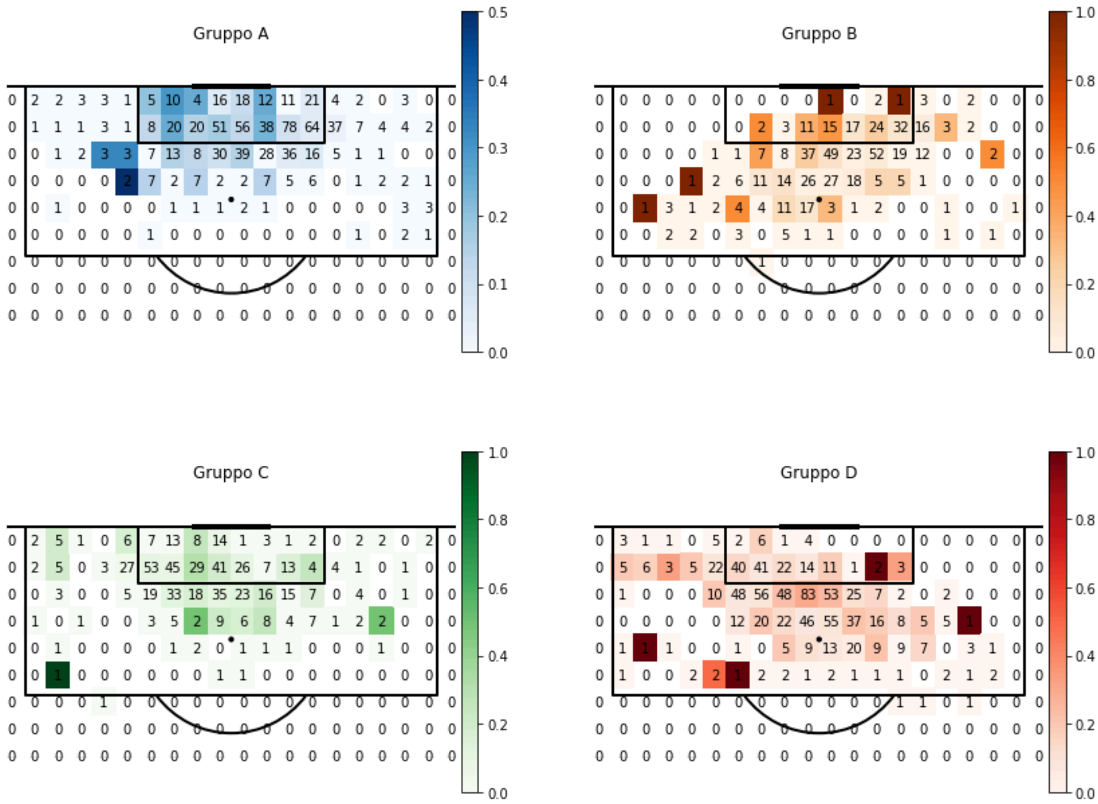
The two variables are linked to the probability of a shot becoming a goal and therefore simultaneously to the probability of a pass becoming an assist. The shorter the distance from the goal (minus the portion of the area near the goal line manned by the goalkeeper) and the greater the angle of the goal, the greater the probability of an assist occurring, and vice versa.

Having defined the target variable and the predictor variables, the statistical model for expected assists was implemented by means of a logistic regression and finally trained with the available data.

5. DATA ANALYSIS

5.1 MODEL EVALUATION

Looking at the conversion rate of the target variable, one can see that there are extreme values due to the low number of assists in those areas of the penalty area.



Therefore, solving this situation, a logistic regression statistical model was adopted to find a correlation between the target variable (xA) and the predictor variables (distance and goal angle).

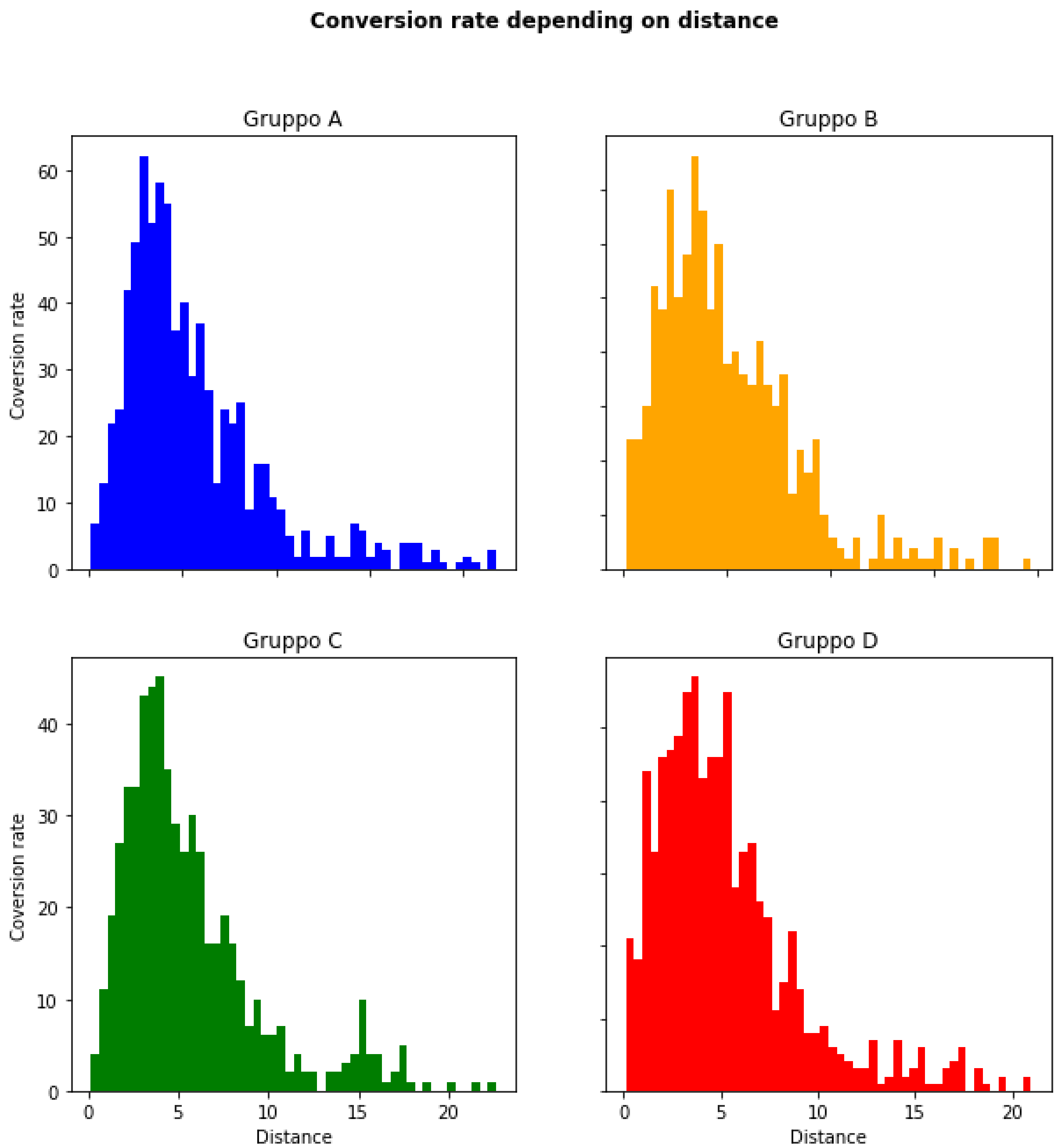
As mentioned above, one should expect the following:

1. the shorter the distance from the goal, the greater the likelihood of an assist occurring (unless the portion of the area close to the goal line is manned by the goalkeeper)
2. the greater the angle of the goal, the greater the probability of an assist occurring

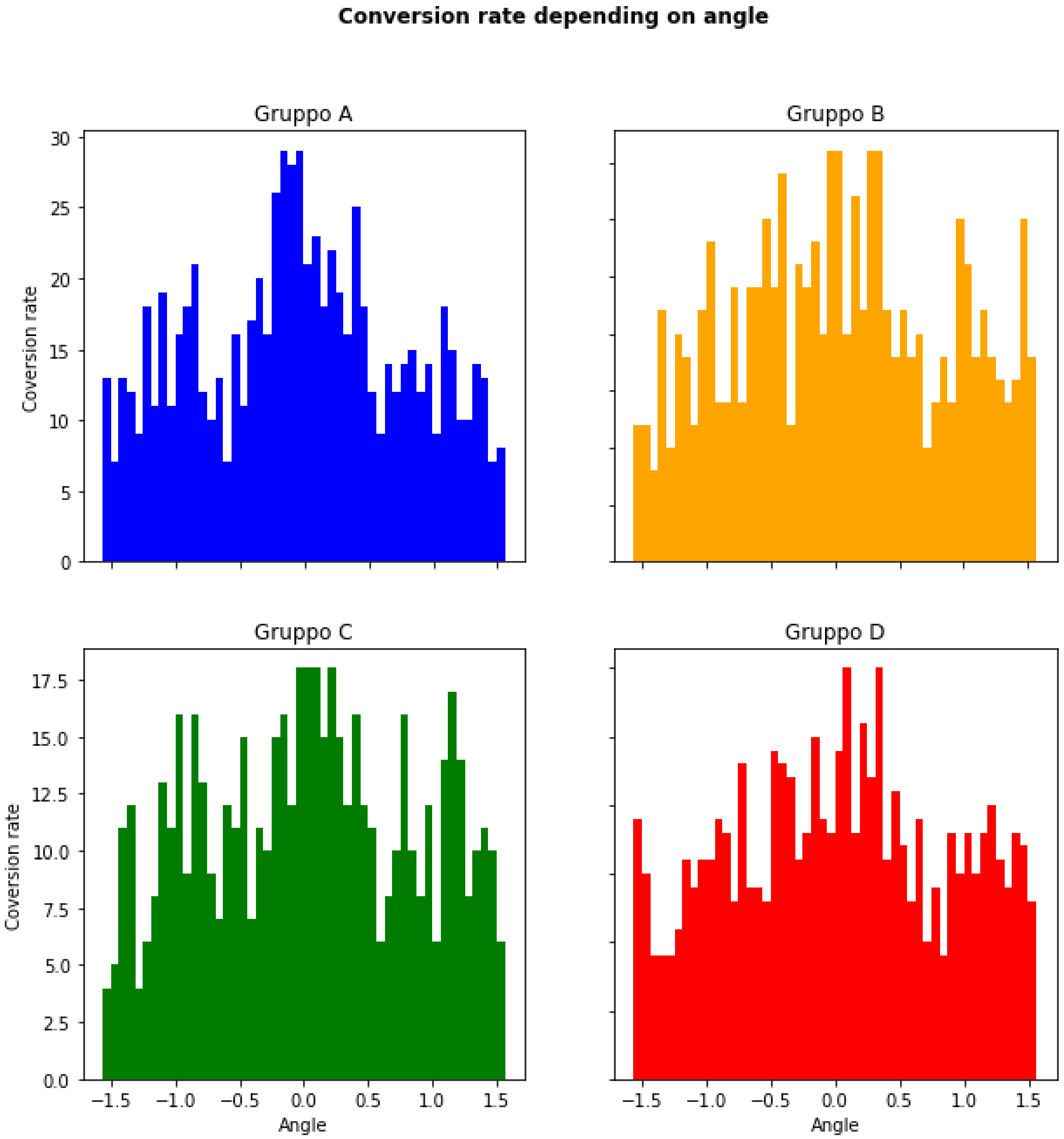
and vice versa

3. the greater the distance from the goal, the lower the probability of an assist occurring
4. the smaller the angle of the goal, the less likely it is that an assist will occur

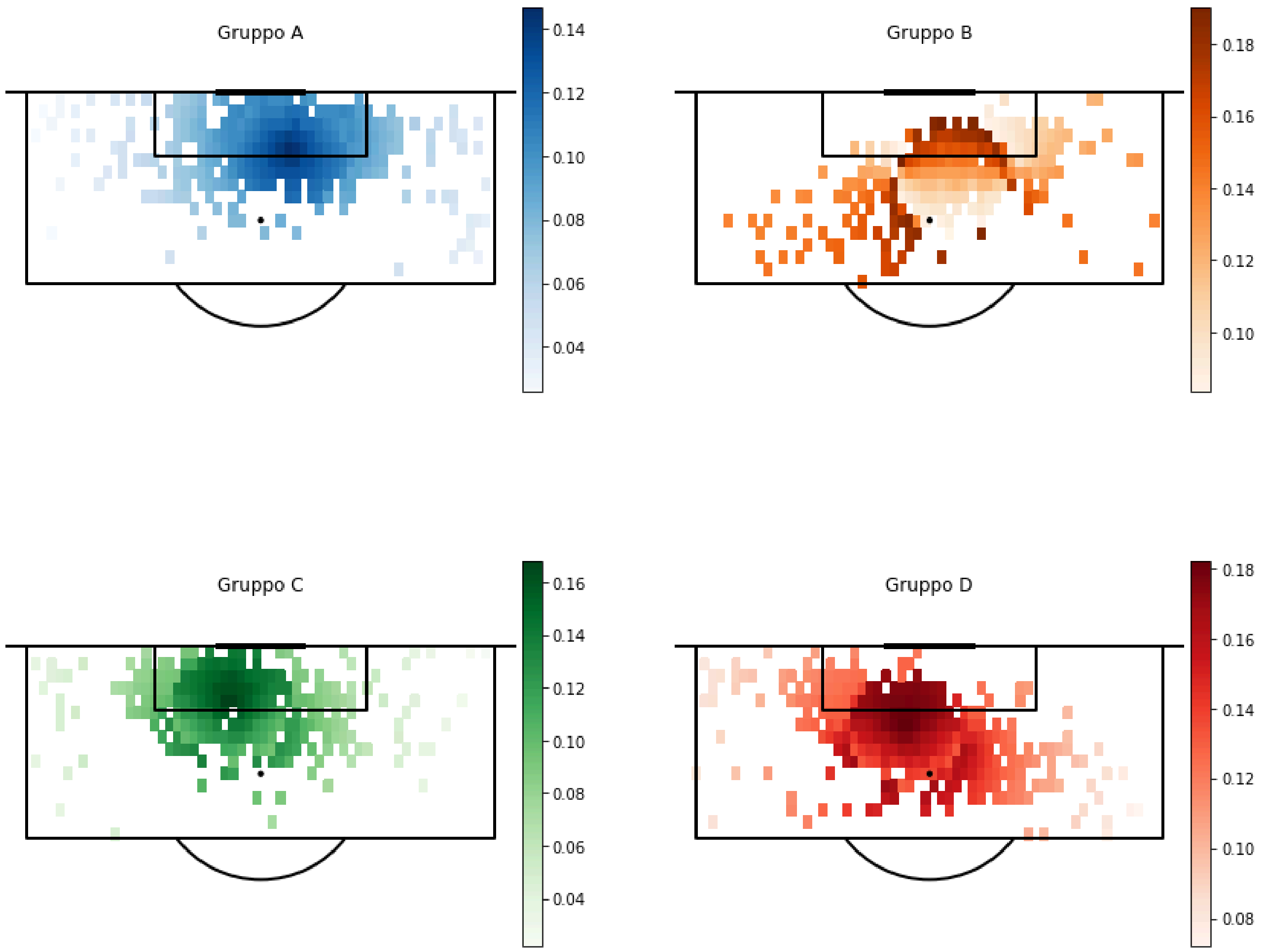
By relating the conversion rate and the distance to the goal, it can be observed that in the four cases examined, the conversion rate decreases as the distance to the goal increases (minus the portion of the area close to the goal line) as assumed when the model was created.



Relating the conversion rate and the goal angle, however, shows that in the four cases examined, the conversion rate decreases as the goal angle decreases, although not as clearly as assumed during the creation of the model.

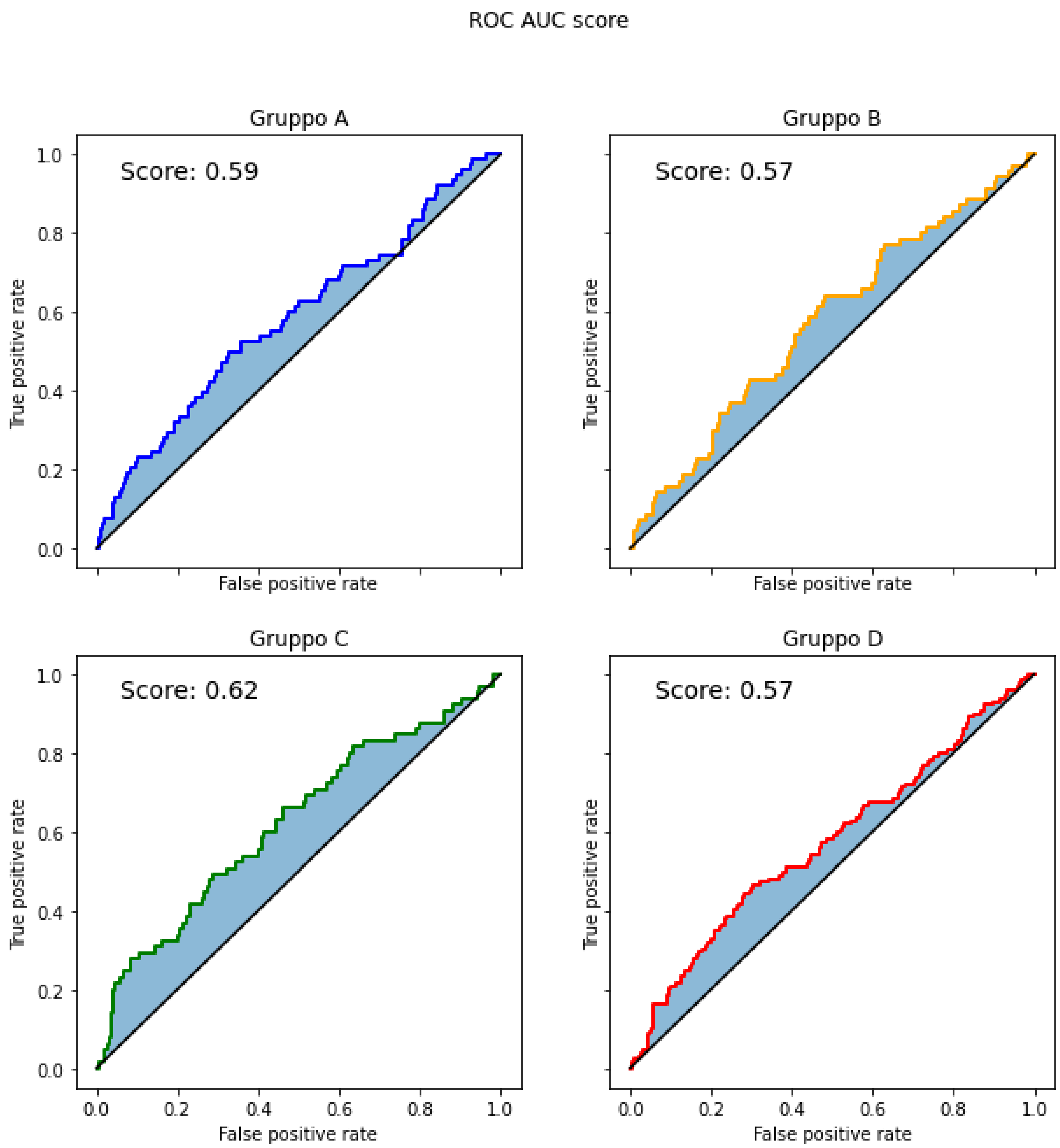


Subsequently, the statistical model was trained and the expected assist values were plotted in the graphs below:



It would appear that in three out of four cases, the model returned values in line with the assumptions made.

To assess the goodness of the model from a quantitative point of view, the ROC AUC score was used, which relates the fraction of true positives and the fraction of false positives.

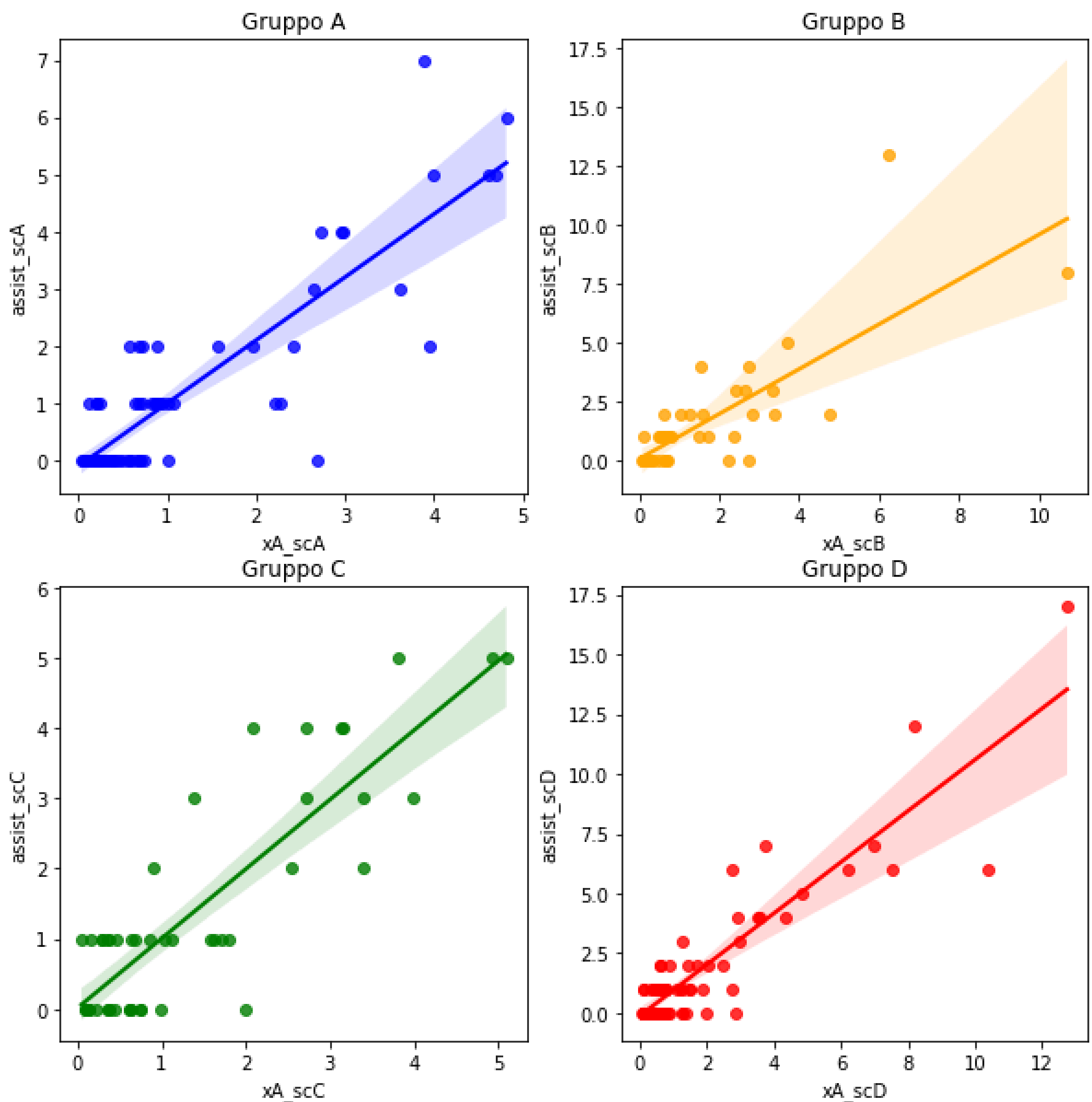


As can be seen from the graphs, the ROC AUC value for the four cases examined, the model is quite reliable.

Instead, LogLoss was used to assess the goodness of the model from the point of view of the probabilistic accuracy of the individual values.

From the values obtained (0.32 Group A, 0.39 Group B, 0.35 Group C, 0.42 Group D), it can be seen that the model is quite accurate.

Finally, the expected assists derived from the statistical model and the recorded assists grouped by player derived from the data in the dataset examined were correlated in order to test whether there is a linear correlation between the metrics.



It can be seen that for low xA values (below 3) there is some linear correlation, whereas for higher xA values there is no good correlation.

In conclusion, in light of the considerations made, the model, albeit with some limitations, can be used to summarily describe the expected assists.

In order to improve the model, it may be necessary:

1. build the model on a much broader base of data from previous seasons and/or other championships;
2. consider other parameters such as the position of the defending team's players and the type of defence adopted (man-to-man or zone) or *headers*.

5.2 RESULTS

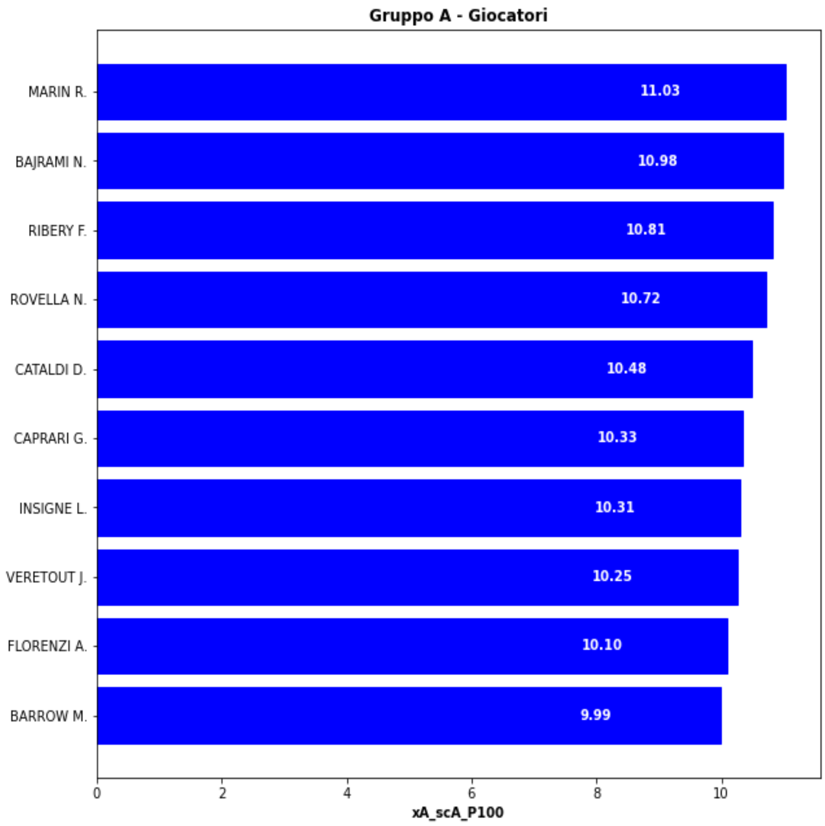
Having said that the model can be used to describe expected assists, the total expected assists of each group were compared in order to assess in which situation an offensive benefit is obtained.

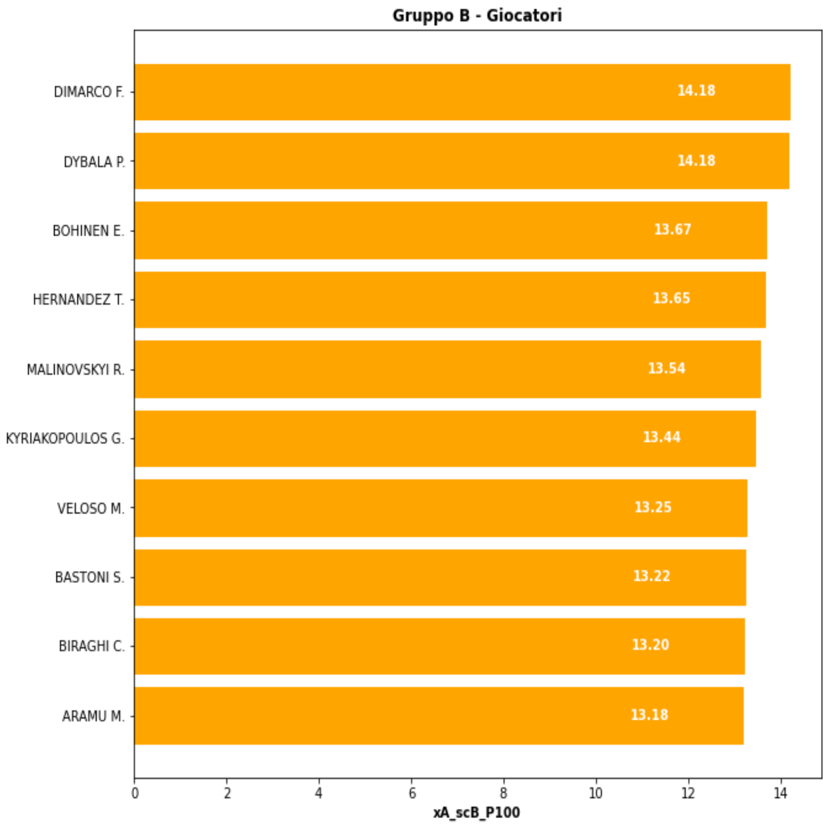
The data were normalised for 100 corner kicks in order to have a common basis on which to compare given the different amount of events for each situation.

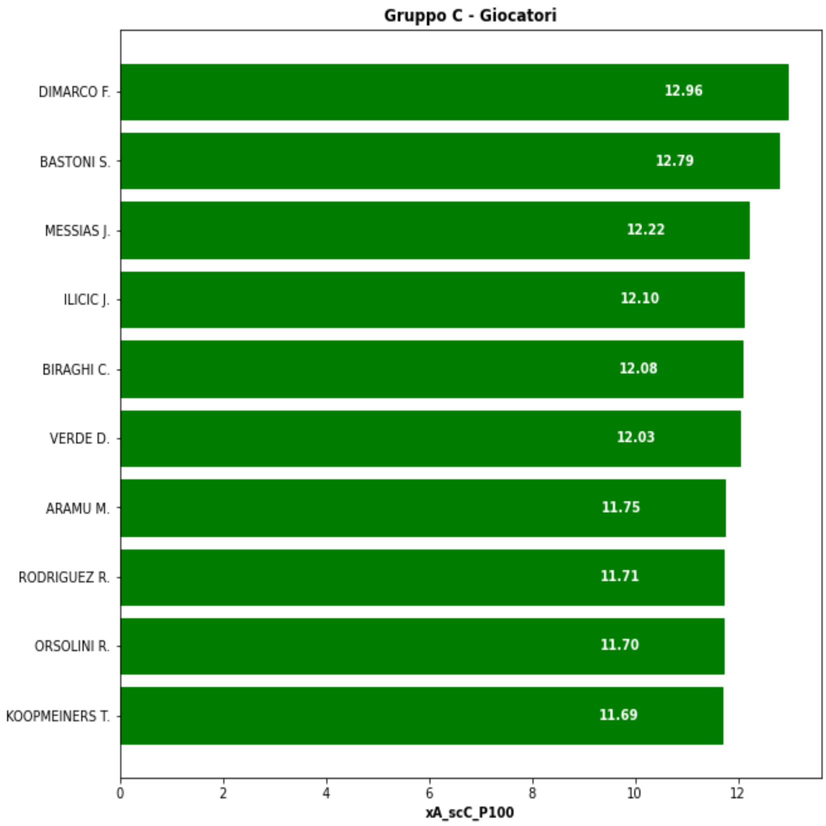
The values show that groups B [13.23 xA] and D [14.93 xA] (crosses from corner kicks with a return effect) bring a greater offensive benefit in terms of assists than groups A [10.13 xA] and C [11.36 xA] (crosses from corner kicks with a return effect)

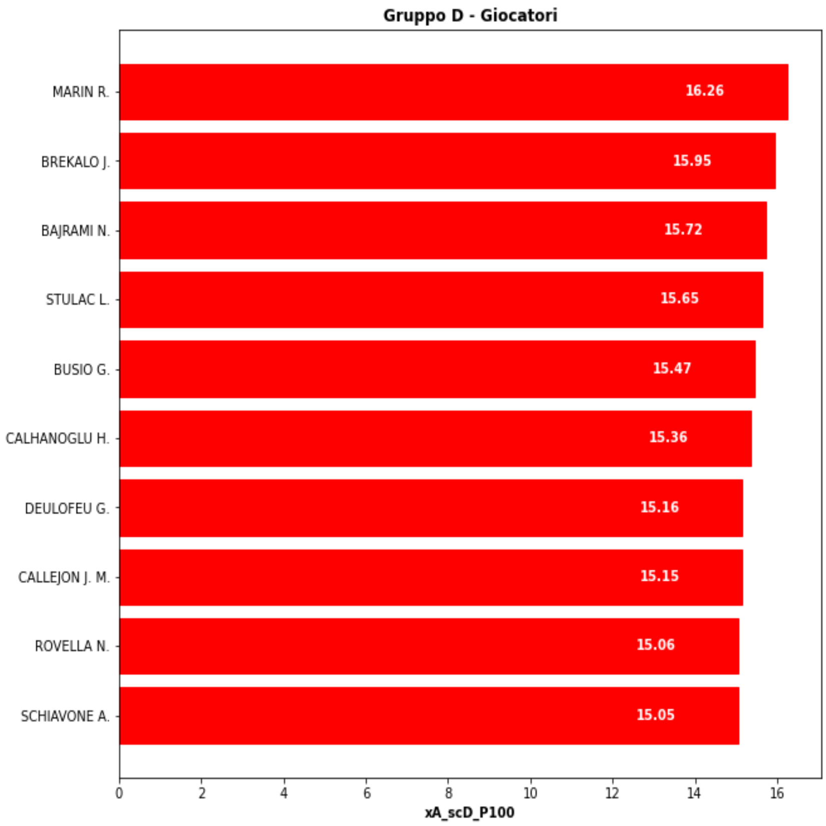
6. COMPARISON

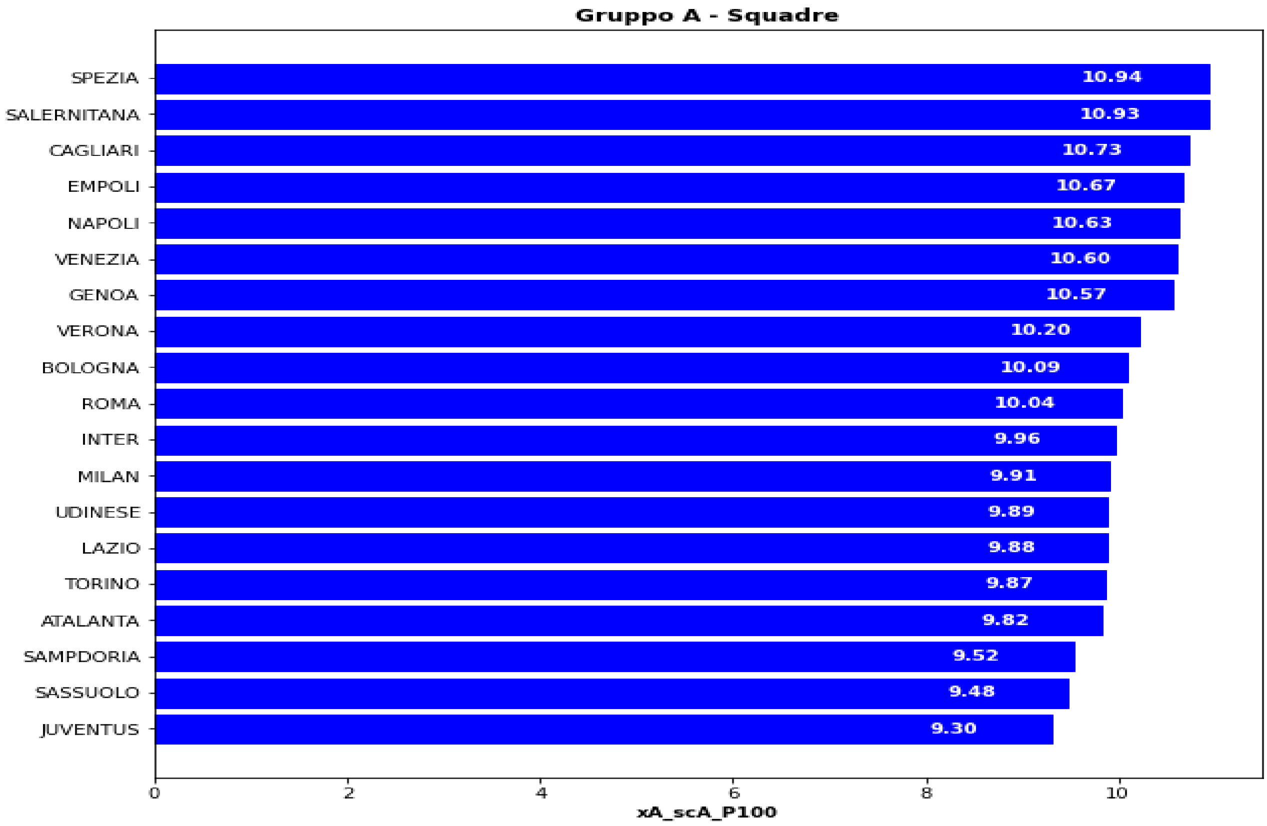
In conclusion, the data on hitters (with at least 10 corner kicks taken) and the teams that generate the most expected assists in the different situations are shown

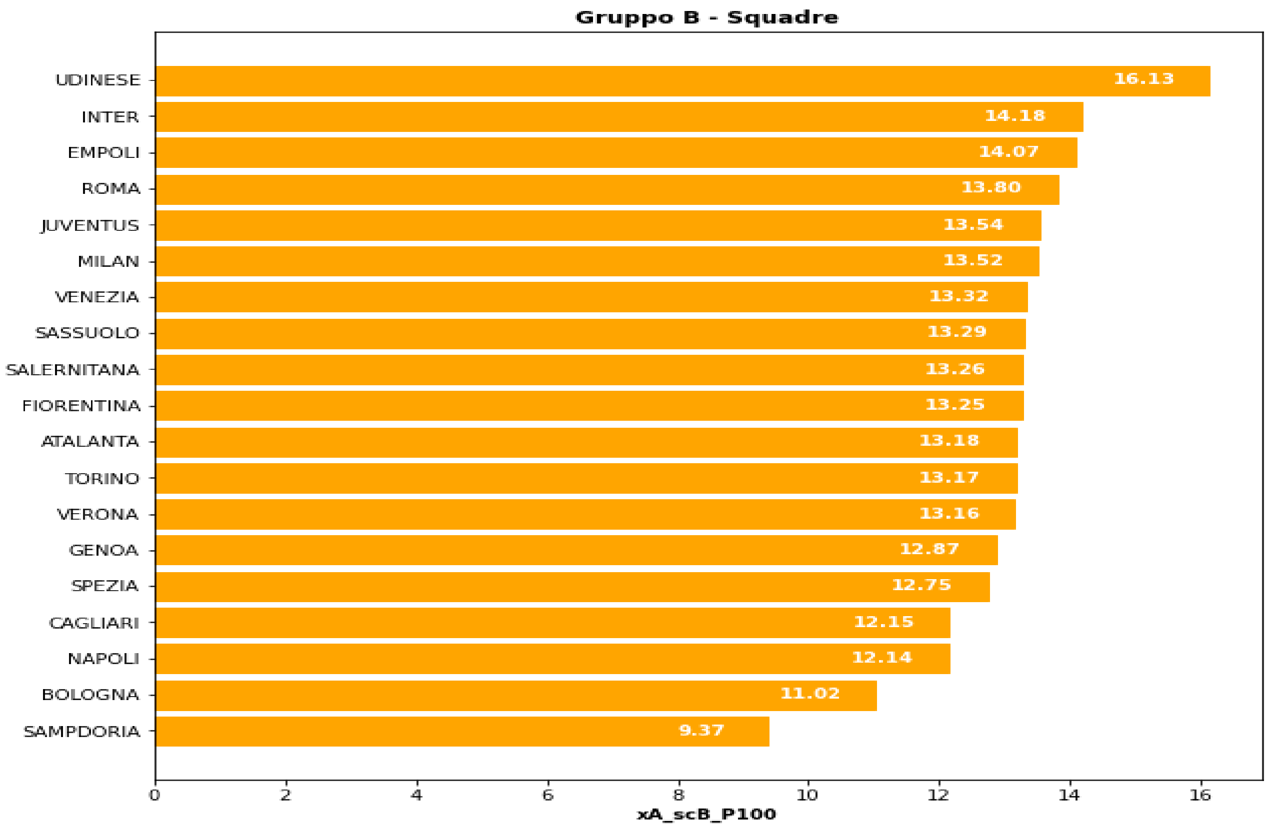


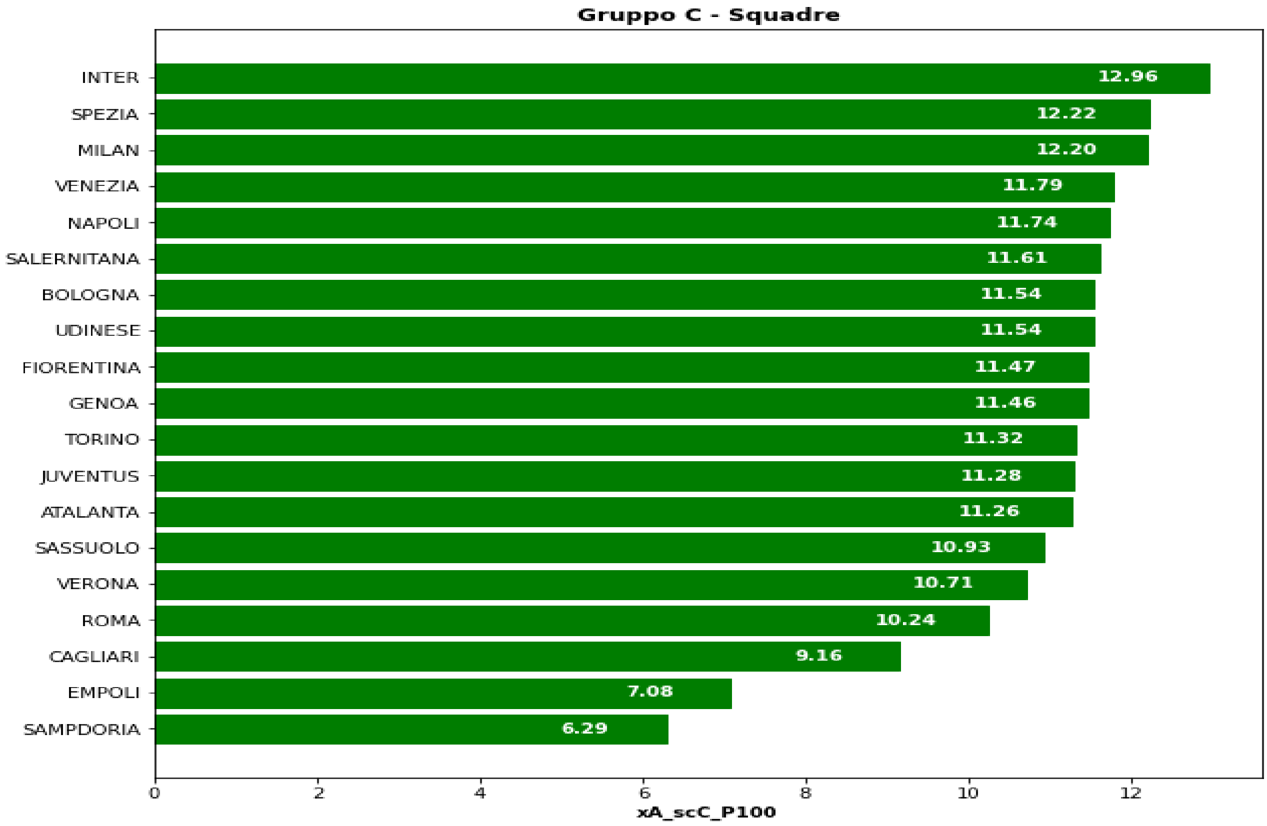


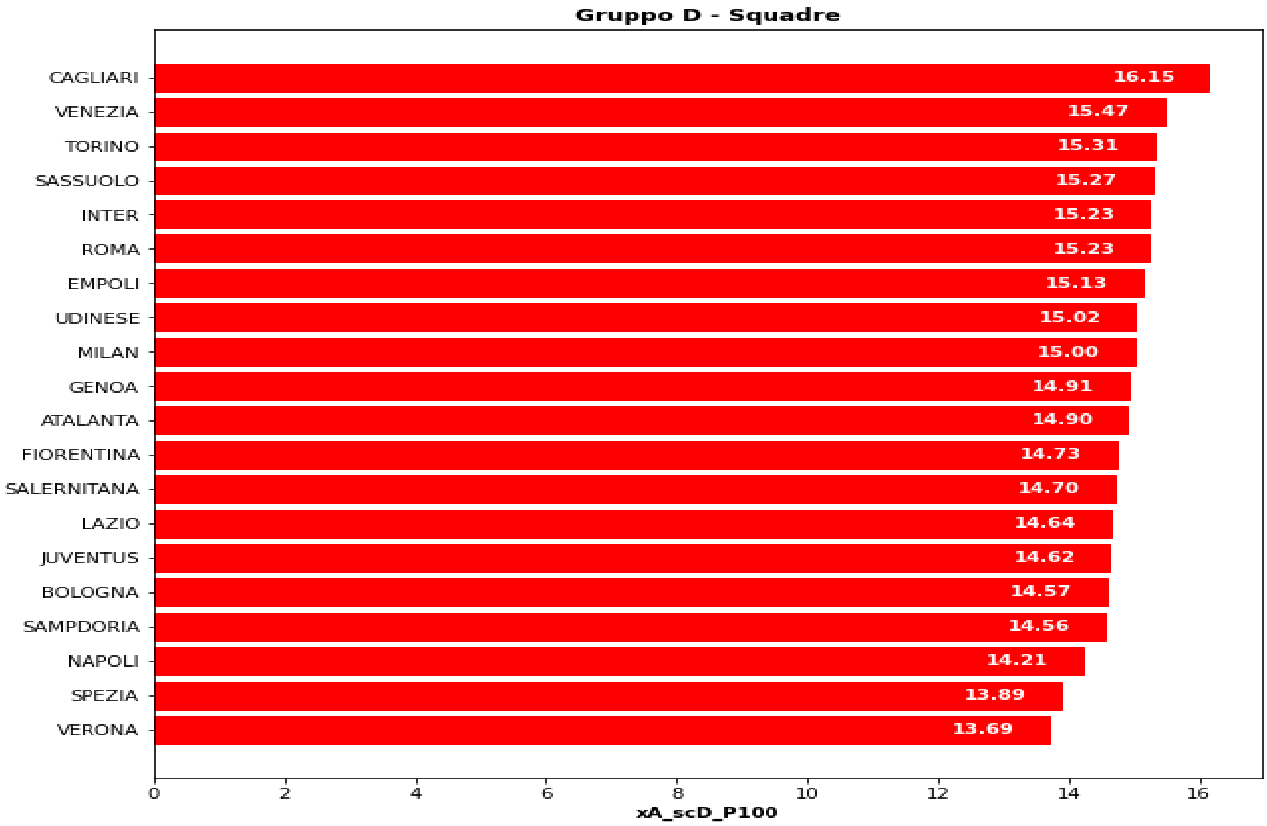

Gruppo C - Giocatori
DIMARCO F.
12.96
BASTONI S.
12.79
MESSIAS J.
12.22
ILICIC J.
12.10
BIRAGHI C.
12.08
VERDE D.
12.03
ARAMU M.
11.75
RODRIGUEZ R.
11.71
ORSOLINI R.
11.70
KOOPMEINERS T.
11.69
0
2
4
6
8
10
12
xA_scC_P100


Gruppo D - Giocatori
MARIN R.
16.26
BREKALO J.
15.95
BAJRAMI N.
15.72
STULAC L.
15.65
BUSIO G.
15.47
CALHANOGLU H.
15.36
DEULOFEU G.
15.16
CALLEJON J. M.
15.15
ROVELLA N.
15.06
SCHIAVONE A.
15.05
0
2
4
6
8
10
12
14
16
xA_scD_P100


Gruppo A - Squadre
SPEZIA 10.94
SALERNITANA 10.93
CAGLIARI 10.73
EMPOLI 10.67
NAPOLI 10.63
VENEZIA 10.60
GENOA 10.57
VERONA 10.20
BOLOGNA 10.09
ROMA 10.04
INTER 9.96
MILAN 9.91
UDINESE 9.89
LAZIO 9.88
TORINO 9.87
ATALANTA 9.82
SAMPDORIA 9.52
SASSUOLO 9.48
JUVENTUS 9.30
0 2 4 6 8 10
xA_scA_P100


Gruppo B - Squadre
UDINESE 16.13
INTER 14.18
EMPOLI 14.07
ROMA 13.80
JUVENTUS 13.54
MILAN 13.52
VENEZIA 13.32
SASSUOLO 13.29
SALERNITANA 13.26
FIORENTINA 13.25
ATALANTA 13.18
TORINO 13.17
VERONA 13.16
GENOA 12.87
SPEZIA 12.75
CAGLIARI 12.15
NAPOLI 12.14
BOLOGNA 11.02
SAMPDORIA 9.37
0 2 4 6 8 10 12 14 16
xA_scB_P100


Gruppo C - Squadre
INTER
SPEZIA
MILAN
VENEZIA
NAPOLI
SALERNITANA
BOLOGNA
UDINESE
FIORENTINA
GENOA
TORINO
JUVENTUS
ATALANTA
SASSUOLO
VERONA
ROMA
CAGLIARI
EMPOLI
SAMPDORIA
12.96
12.22
12.20
11.79
11.74
11.61
11.54
11.54
11.47
11.46
11.32
11.28
11.26
10.93
10.71
10.24
9.16
7.08
6.29
0
2
4
6
8
10
12
xA_scC_P100


Gruppo D - Squadre
CAGLIARI
VENEZIA
TORINO
SASSUOLO
INTER
ROMA
EMPOLI
UDINESE
MILAN
GENOA
ATALANTA
FIORENTINA
SALERNITANA
LAZIO
JUVENTUS
BOLOGNA
SAMPDORIA
NAPOLI
SPEZIA
VERONA
16.15
15.47
15.31
15.27
15.23
15.23
15.13
15.02
15.00
14.91
14.90
14.73
14.70
14.64
14.62
14.57
14.56
14.21
13.89
13.69
0
2
4
6
8
10
12
14
16
xA_scD_P100

7. FINAL CONSIDERATIONS

To recapitulate:

- The corner kick within a match can be an offensive weapon to allow a team to be dangerous and be able to unlock its goalscoring;

- 4 types of corner situations: right corner, right foot batter (cross with return effect), right corner, left foot batter (cross with return effect), left corner, left foot batter (cross with return effect), left corner, right foot batter (cross with return effect);

- the logistic regression statistical model for expected assists, trained on the available data, is quantitatively quite reliable;

- the statistical logistic regression model for expected assists, trained on the available data, is at the level of probabilistic accuracy quite accurate;

- there is a certain linear correlation between expected assist model values and assists from recorded events;

- need to improve the statistical model by training it on a broader base of data from previous seasons and/or other leagues and considering additional dependency parameters;

- Corner kicks with crosses with a return effect bring a greater offensive benefit in terms of assists than corner kicks with crosses with an exit effect.