Statistical Analysis on Classification and Forecasting Models in Football

Lorenzo Leoni

***Abstract*—In this study will be shown the odds model on a dataset of Serie A matches of 2021/2022 season played until now (27^{th} matchweek) for classifying the outcomes of a football event and it'll be compared to ordered logistic models (*OLM*) based on the value of the home and away team (refers to the entire squad or only to the players in the lineups during the match) their respectively performances in the previous two matches and the goal difference. With some generalizations of the independent variables of the *OLM*, it can be observed that the model with the value of the players in the lineups is the best, according to several metrics such as AIC, BIC and precision in addition to having passed statistical tests as Lipsitz and the ordinal Hosmer-Lemeshow. Afterwards it'll be selected a subset of matches played by AC Milan for forecasting their no penalty expectation goals (*npxG_Expected*) using a generalized linear model with a gamma distribution and a logarithmic link function rather then the canonical inverse. Using forward selection and an initial distribution analysis of residuals, it finds that the gamma regression model without forward selection and without outliers removal is the best one, since the deviance on the test dataset is the lowest value.**

I. Introduction

The study has two main goals: the first one is to try to understand in which way the odds model can be improve, using different variables and combinations of them. The second one is to find a model that, using some variables created for the first goal, can predict the number of expected goal without penalty kicks.

The difficult part for the first part is to find such variables that can improve both models. For the ordered model was important that the variables used are more "general" as possible. This term means that variables such as assist, number of yellow cards or number of shots during the match can't be used. For example, our model can't known before the match how many shots the home or away team attempt. Therefore it's necessary to use variables known before every games of each team.

The analysis on AC Milan is slightly different from this point of view. What it is tried to do is to use variables known after the games to train the model. Afterwards it is used to predict the next game expected goal using the mean of each independent variable instead of the real value that, like said before, can't be known before the game. This because the idea behind this model is that can be used to understand which value of expected goal can be reach by the team if they and the opponent have certain statistics (used like regressors).

Goddard & Asimakopoulos [2] analyzed a model that takes into account results of the past matches, the importance of them, the travel distance, the participation in cup matches and concluded that these factors are all significant influence on the outcome of a match. A generalization can be done using ELO model. The ELO is a value that refers to the strength of the team in analysis. It changes based on the outcome of matches played and can be a more detail method to evaluate the team performance.

Dixon & Coles [5] confirmed that factors such as team performance, different strength of teams and home advantage are important factors. The models showed in this study takes into account some of these factors mentioned before.

II. Methodology

In this section it'll be shown step by step the models created, how data was manipulated and which feature have been chosen and why.

A. Odds Model

The first model analyzed was the odds model. Odds were taken from sport website *statistichesulcalcio* [10]. The dataset is the set of every match until the 27^{th} matchweek, with the date, the two teams (home and away), the number of goals for each team and odds for home win, draw, and away win. It's necessary to define a variable named *FTR*, which represents the team that win the match, H (Home), A (Away), D (Draw). It'll be fundamental to create the confusion matrix, in order to calculate the precision of the model.

The next step is to transform odds in probability. In this dataset the odds are in the format decimal/european, then they conversion should be $\frac{1}{Odds}$. This is true when the odds are from a fair book, that is when the sum of all the probabilities are equal to 1. Obviously the odds from bookmakers are set in order to return a certain profit. As shown by Cortis [4], if the implied probability of any outcome is lower or equal to than the actual probability, then the bookmaker is expected to have a profit. The actual probability is given by the formula before, therefore bookmakers set lower odds and this cause a higher probability. Adding them together the value is higher then 1. Subtracting the fair value 1 to the value obtained it gets the percentage named *overround*. It represent the bookmaker margin of profit.

Known that, to obtain the fair odds of the match it's necessary to calculate the probability as follow:

$$P_w = \frac{\frac{1}{Odds_w}}{\frac{1}{Odds_w} + \frac{1}{Odds_d} + \frac{1}{Odds_l}}$$

This is true also for all the other probability P_d and P_l. Once obtained the probabilities, the next step is to find the higher one for each match and declare the outcome. In the end it can be possible to visualize the confusion matrix of the model and calculate the precision of the classification.

B. Ordered Logit Model with TM Value Costant

Once built the model with odds, the next step is to build other ones in order to compare them and find the best classification possible. Before to build the model it's important to do some data cleaning of the new dataset, which contains all the statistics from all the matches played until the 27^{th} matchweek, which can be named as *SerieA*. It belongs from *FBref* [11] and was extracted via *worldfootballR*, a R library [12]. An important step to do is to edit the format of the match dates. To do this was used Excel, and the date was formatted in *"YYYY-MM-DD"*. This procedure will be fundamental for the next model. Even here, a variable similar at *FTR* was created, in this case it'll be named *Win*.

Afterwards, it's loaded the *Transfermarkt* value of the teams updated at $01/03/22$ [13]. it's assumed that the value of the teams at the beginning of the season is the same as on March 1st. Since value of teams are measured in millions of euro and huge differences can be present, the ratio between team values was done to reduce them. The distribution of the ratio is in Figure 1.

Logarithmic transformations are a convenient means of transforming a highly skewed variable into one that is more approximately normal [3], this can be seen in Figure 2. The presence of highly skewed variables can influence the distribution of residuals making them non-normal.

Using Excel it was possible to insert the variable *Home/Away_Cup* C. It was modeled like this:

$$\begin{cases} C_{i,j} = 1, & i^{th} \text{ team played in cup} \\ C_{i,j} = 0, & \text{Otherwise} \end{cases}$$

As can be seen, this variable takes into account the influence of a match during the week, like did by Goddard & Asimakopoulos.

SerieA has two rows for each match, this because one row is for the home team statistics and the other one is for the away team statistics. Their will be fundamental for the model about AC Milan. Since the idea for the ordered model wasn't to use these statistics but general variables, such as team value and cup match variable, it were deleted the duplicated rows. Using the previous variable *Win*, it can be possible to define another one important variable, *Winvalue* V. It'll be the dependent variable for the model and is defined as follow:

$$\begin{cases} V_{i,j} = 2, & i^{th} \text{ team W } j^{th} \text{ match} \\ V_{i,j} = 1, & i^{th} \text{ team D } j^{th} \text{ match} \\ V_{i,j} = 0, & i^{th} \text{ team L } j^{th} \text{ match} \end{cases}$$

Before to build the model there are other variables useful to define, such as the *Home/Away_Performance* P and the *Home/Away_Diff_Goals_lag* D. The first one keep track of the performance of the team in analysis during the last two games, using a score similar to a boost/penalty. More precisely:

$$\begin{cases} P_{i,j} = P_{i,j-1} + 1, & i^{th} \text{ team W } j^{th} \text{ match} \\ P_{i,j} = P_{i,j-1} - 0.5, & i^{th} \text{ team D } j^{th} \text{ match} \\ P_{i,j} = P_{i,j-1} - 1, & i^{th} \text{ team L } j^{th} \text{ match} \end{cases}$$

The maximum and minimum values admissible are 2 and -2. Hence, if a team has a performance score equal to -1.5 or less and they lose the next game, the score can't go below -2. The idea to set a penalty for teams that draw it's because a streak of draws can't be irrelevant. If it was set at 0 a team that has win the last two matches still maintain their top form. It's important to note that some matches was postponed, then the teams that haven't played their match maintain their score performance unchanged.

The last variable is the difference of goals for each team. In the j^{th} match was fixed the difference until the match before (this is the motivation of its name "lag"). This assumption is fundamental because if it could know the number of goals made in the match in analysis then it would always be able to classify correctly the final result. This approach can't be interpreted as prediction of the outcome of the match.

Once created all these variables, a subset of the original dataset was saved and named *Games*. It contains features that can be useful for the ordered logit model, which is particular type of logistic regression model that applies to dichotomous dependent variables. Where logistic regression assigns probabilities that a

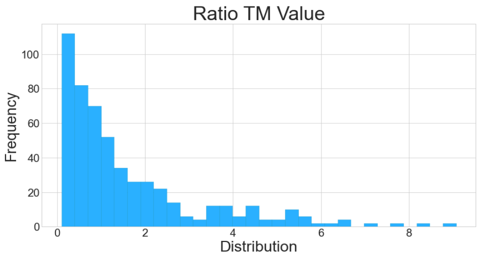


Fig. 1: Distribution of the ratio between teams value in Serie A

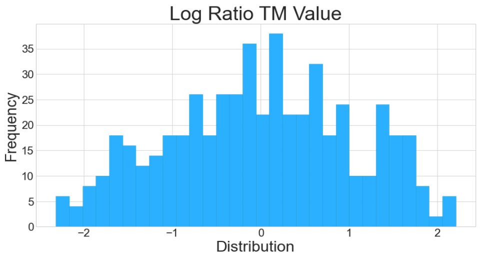


Fig. 2: Distribution of the log ratio between team values in Serie A

variable will take on a specific value, ordered logit assigns probabilities that values will fall below a certain threshold. The model is based on the cumulative probabilities and can be modeled as follow:

$$logit(g_{c_i}) = log\Big(\frac{g_{c_i}}{1-g_{c_i}}\Big) = \alpha_c - \beta' x_i$$

where g_{c_i} is the cumulative probability to be in a certain category c_i, $\forall i \in I$ with I the set of all possible outcomes. α_c are the thresholds [8]. These ones are the bounds of the categories where the outcome can be. The vector x are the regressors. In this specific case the model is given by:

$$\begin{cases} log(\frac{\mathbb{P}_l}{1-\mathbb{P}_l}) = \alpha_{c_{l,d}} - \beta' x_i & \text{Home L} \\ log(\frac{\mathbb{P}_l+\mathbb{P}_d}{1-\mathbb{P}_l-\mathbb{P}_d}) = \alpha_{c_{d,w}} - \beta' x_i & \text{Home L or D} \end{cases}$$

$\mathbb{P}_i$ is the probability that home team has the outcome i. Fitting the ordered logit model (using the *bevel* package [14]) it is obtained the two thresholds for the three possible outcomes and the slopes for each regressor. Probabilities are calculated using the equations:

$$\begin{cases} \mathbb{P}_l = \frac{1}{e^{-(\alpha_{c_{l,d}} - \beta' x)}} \\ \mathbb{P}_d = \frac{1}{e^{-(\alpha_{c_{d,w}} - \beta' x)}} - \mathbb{P}_l \\ \mathbb{P}_w = 1 - \mathbb{P}_l - \mathbb{P}_d \end{cases}$$

Now feature selection. After several attempts the main regressors found are *log_ratio_Value* and *Home/Away_Performance*. These variables will compose the main model that will be used as comparing one for doing feature selection. In order to avoid to much covariates, and therefore the overfitting, several simulations were done using first *Home/Away_Diff_Goals* and then *Cup_Home/Away*. Adding the main three variables with the two pairs, the results are reported in Table I. P-value is over the 5%, the fixed significant level, for all the variables and the null value is within the confidential intervals, hence these features aren't statistically significant [9].

The statistical significant can be the only consideration to take into account when it have to select features. An important phase is the inferential one, where models are tested in order to find the best set of regressors. The *Games* database was saved in csv from Python and loaded in R, where it can be possible to do the analysis in a more statistic way. Two tests were carried out: Lipsitz (L) and the ordinal Hosmer-Lemeshow (HL). They are best suited to detect lack of fit associated with continuous covariates [7]. The results reported in Table II suggest that it has a correct specifications for the model with *Cup_Home/Away*, while the model with *Home/Away_Diff_Goals* doesn't confirm the null hypothesis of the correct specifications for Hosmer-Lemeshow test.

Another comparison is on the AIC and BIC. They are metrics used for model selection, the main difference is that the BIC penalizes more for additional parameters than AIC. The results are reported on Table III, they, unlike before, suggest that the model with *Home/Away_Diff_Goals* variables has better values. The last metric that was taken in consideration is the precision and, as it

TABLE I

		Cup_Home/Away			
Feature	Beta	Std. Err.	P-value	0.025	0.975
CH	-0.2092	0.3657	0.5673	-0.9260	0.5076
CA	0.0292	0.3653	0.9363	-0.6868	0.7452
		Home/Away_Diff_Goals			
Feature	Beta	Std. Err.	P-value	0.025	0.975
GH	-0.0154	0.0149	0.3011	-0.0447	0.0138
GA	0.0178	0.0155	0.2520	-0.0127	0.0483

Parameters from OLM with Winvalue as dependent variable and log_ratio_Value, Home/Away_Performance as fixed regressors

TABLE II

	Cup_Home/Away		
Test	LR/χ^2	df	P-value
L	9.7297	9	0.3728
HL	22.455	17	0.1678
	Home/Away_Diff_Goals		
Test	LR/χ^2	df	P-value
L	13.253	9	0.1515
HL	31.913	17	0.01542

Parameters from OLM with Winvalue as dependent variable and log_ratio_Value, Home/Away_Performance as fixed regressors

TABLE III

Cup_Home/Away	
AIC	BIC
518.92	543.98
Home/Away_Diff_Goals	
AIC	BIC
517.32	542.38

Parameters from OLM with Winvalue as dependent variable and log_ratio_Value, Home/Away_Performance as fixed regressors

TABLE IV

Cup_Home/Away		
Win	A	H
A	66	27
D	33	38
H	23	78
Home/Away_Diff_Goals		
Win	A	H
A	62	31
D	36	35
H	24	77

Parameters from OLM with Winvalue as dependent variable and log_ratio_Value, Home/Away_Performance as fixed regressors

can see in the Table IV, the best model is the one with the regressors *Cup_Home/Away*, since the precision is about 53.96%. The model with *Home/Away_Diff_Goals* has a precision about 52.45%.

After these analysis, the main model (with the three base covariances) is the best one in all the metrics takes in consideration. The analysis of the models stops here, as it has been chosen not to continue with the calculation of the residuals and the Shapiro test due to the calculation method used. *resids()* (from the *sure* R library) uses random sampling from a continuous distribution [15], so it doesn't get identical residuals in all runs. This implies that the Shapiro test is sometimes positive and sometimes negative. Now it can be possible to proceed to the next model.

C. Ordered Logit Model with TM Value Lineups

in order to generalize the previous model, the formations that the teams have lined up on the field have been taken into consideration rather than the full value of their squad. In this way it is possible to model the injuries factor, since the value of the teams depends on the players in the pitch. To do this, it was loaded a database created in R called *ln_TM*, where was merged the lineups for each team in each game (from *FBref* [11]) and the value of each player that played the match (from *Transfermarkt* [16]). Since player names, in some cases, are written in different ways from these two websites, the merge was made using a txt file [17] where each player has the url link for *FBref* and *Transfermarkt* in order to avoid missing players in the lineups caused on the different written form of their names. The database was extracted via *worldfootballR*, a R library [12]. There are several missing values, which represent players in bench with 0 minutes player, that were drop from the database.

Once the dataset *ln_TM* is loaded in Python, an important reflection has to be made. How much a player contributes in the team value in a given match? Think that a player, which market value is for example 80 millions of euro, contributes for his full value in the team cost even if he has played for only 5 minutes is not true. From this consideration it was used the following condition:

$$TMR_{i,j} = \frac{TM_{i,j} \cdot MIN_{i,j}}{90}$$

where the i^{th} player in the j^{th} match with *TM* his market value and *MIN* his minutes played in that match, gives a real contribution in the team value equal to *TMR* for the match in analysis.

Thanks to formatting of the dates in the database before, it can be possible to compared the dates of the matches in the databases *Games* and *ln_TM*, in order to find for each team all the players in the pitch for each match. Thanks to that it are created two columns in the main dataset *Games* that represent the value of the lineups of the teams, called *ln_TM_Home/Away*.

As before, it was calculated the logarithmic of the ratio of the teams value in each match. How it was represented in Figure 3, the variable given by only the ratio was highly skewed. The logarithmic transformation help to turn its distribution in a

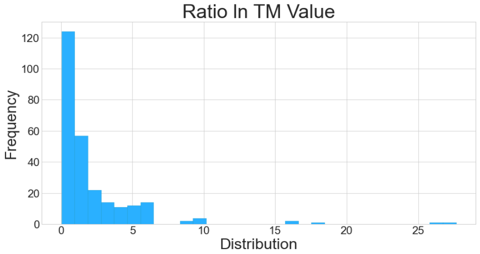


Fig. 3: Distribution of the ratio between the value of the lineups of the Serie A teams

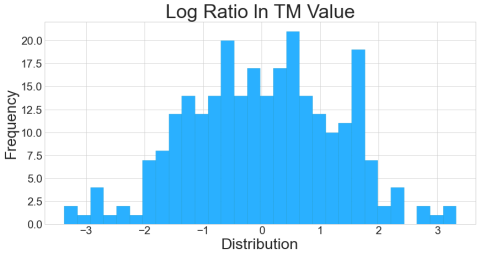


Fig. 4: Distribution of the log ratio between the value of the lineups of the Serie A teams

normal one (see Figure 4). The converted variable is called *log_ratio_ln_TM*.

As before, it can proceed with the feature selection. After some attempts the main regressors found are the same as the previous OLM, that are *log_ratio_ln_TM*, *Home/Away_Performance*, but now there are other two that resulted relevant, *Home/Away_Diff_Goals*. In this case the first three independent variables alone aren't the best set of feature, then the main model that will be used as comparing one for doing feature selection is the one with these 5 regressors. In order to avoid to much covariates, and therefore the overfitting, several simulations were done using first *log_ratio_ln_TM*, *Home/Away_Performance* alone and then *Cup_Home/Away* without the two new entry *Home/Away_Diff_Goals*. The results are reported in Table V. For the *Cup_Home/Away* P-value is over the 5%, the fixed significant level, and the null value is within the confidential intervals, hence these features aren't statistically significant [9]. While the main regressors confirm to be statistically relevant for both the P-value and CI.

The statistical significant can be the only consideration to take into account when it have to select features. An important phase is the inferential one, where models are tested in order to find the best set of regressors. The Games database was saved in csv from Python and loaded in R, where it can be possible to do the analysis in a more statistic way. Two tests were carried out: Lipsitz (L) and the ordinal Hosmer-Lemeshow (HL). They are best suited to detect lack of fit associated with continuous covariates [7]. The results reported in Table VI suggest that it has a correct specifications for both the model.

Another comparison is on the AIC and BIC. They are metrics used for model selection, the main difference is that the BIC penalizes more for additional parameters than AIC. The results are reported on Table VII and suggest that the

TABLE V

		Cup_Home/Away			
Feature	Beta	Std. Err.	P-value	0.025	0.975
CH	−0.1470	0.3633	0.6857	−0.8590	0.5650
CA	−0.0294	0.3691	0.9364	−0.7528	0.6939
		Main Regressors Only			
Feature	Beta	Std. Err.	P-value	0.025	0.975
LN	1.2203	0.1663	0.0000	0.8943	1.5463
PH	−0.2860	0.1063	0.0071	−0.4943	−0.0776
PA	0.2736	0.1079	0.0112	0.0621	0.4852

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM and Home/Away_Performance as fixed regressors

TABLE VI

	Cup_Home/Away		
Test	LR|χ^2	df	P-value
L	1.8485	9	0.9936
HL	12.081	17	0.7952
	Main Regressors Only		
Test	LR|χ^2	df	P-value
L	2.3961	9	0.9835
HL	11.811	17	0.8114

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM and Home/Away_Performance as fixed regressors

TABLE VII

Cup_Home/Away	
AIC	BIC
506.65	531.71
Main Regressors Only	
AIC	BIC
502.88	520.78

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM and Home/Away_Performance as fixed regressors

model with the main regressors has better values.

The last metric that was taken in consideration is the precision and, as it can see in the Table VIII, the best model is the one with the regressors *Cup_Home/Away*, since the precision is about 53.96%. The model with the main regressors has a precision about 53.58%.

Like it said before, the model with the three main covariances and the *Home/Away_Diff_Goals* is the best one in all the metrics take in consideration. The analysis of the models stops here, as it has been chosen not to continue with the calculation of the residuals and the Shapiro test due to the calculation method used. *resids()* (from the *sure* R library) uses random sampling from a continuous distribution [15], so it doesn't get identical residuals in all runs. This implies that the Shapiro test is sometimes positive and sometimes negative.

D. Ordered Logit Model with Elo

The last OLM uses a new feature, the Elo. A club's Elo rating is an estimation of its strength based on past results allowing predictions for the future [18]. Thanks to this variable, the new model can have an historical memory about the strength and its changes have more sensibility than the *Home/Away_Performance*. Once saved and formatted the data in Excel, the database named *Elo* is loaded in Python. Using dates it was possible to set for each team its Elo, creating two new variables in the main dataset *Games*, called *Elo_Home/Away*.

TABLE VIII

Cup_Home/Away			
Win	A	D	H
A	63	4	26
D	30	2	39
H	20	3	78
Main Regressors Only			
Win	A	D	H
A	64	3	26
D	28	1	42
H	22	2	77

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM and Home/Away_Performance as fixed regressors

The value of this new feature has a min and max equal to 1428.278442 and 1903.163696 respectively. Since this value are relative high compared to the other values in the main database, even in this case the ratio is the best solution. In this way the large difference in Elo values are restricted to lower ones and similar to the other in *Games*. How it can be see from the Figure 5, it's not necessary a logarithmic transformation since the values seem to be normally distributed.

As before, it can proceed with the feature selection. In this model the new variable *ratio_Elo* has to be in the set of the regressors, since it want to see the effect of it, together with the main regressors *log_ratio_ln_TM* and *Home-/Away_Performance*. Like the previous OLM the pair *Home/Away_Diff_Goals* is relevant so they are included too. Then the main model that will be used as comparing one for doing feature selection is the one with these 6 regressors. In order to avoid to much covariates, and therefore the

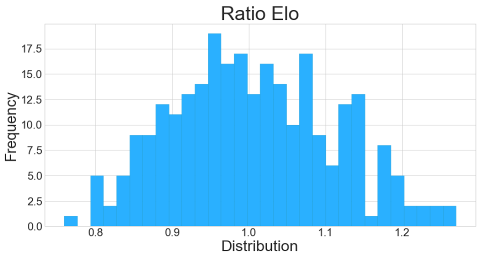


Fig. 5: Distribution of the ratio between the Elo of the Serie A teams

TABLE IX

Cup_Home/Away					
Feature	Beta	Std. Err.	P-value	0.025	0.975
CH	−0.1518	0.3665	0.6787	−0.8702	0.5665
CA	−0.0231	0.3744	0.9509	−0.7569	0.7108
Main Regressors + ratio_Elo					
Feature	Beta	Std. Err.	P-value	0.025	0.975
LN	1.2034	0.2645	0.0000	0.6851	1.7218
PH	−0.2882	0.1099	0.0087	−0.5036	−0.0729
PA	0.2760	0.1116	0.0134	0.0572	0.4947
RE	0.2533	3.0991	0.9348	−5.8207	6.3274

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM, Home/Away_Performance and ratio_Elo as fixed regressors

TABLE XI

Cup_Home/Away			
Test	LR|χ^2	df	P-value
L	2.2443	9	0.987
HL	11.546	17	0.8268
Main Regressors + ratio_Elo			
Test	LR|χ^2	df	P-value
L	4.2436	9	0.8947
HL	14.066	17	0.6625

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM Home/Away_Performance and ratio_Elo as fixed regressors

overfitting, several simulations were done using first *log_ratio_ln_TM*, *Home/Away_Performance* and *ratio_Elo* alone and then *Cup_Home/Away* without *Home/Away_Diff_Goals*. The results are reported in Table IX. For the *Cup_Home/Away* P-value is over the 5%, the fixed significant level, and the null value is within the confidential intervals, hence these features aren't statistically significant [9]. While the main regressors plus the Elo contribute are statistically relevant for both the P-value and CI.

The statistical significant can be the only consideration to take into account when it have to select features. An important phase is the inferential one, where models are tested in order to find the best set of regressors. The Games database was saved in csv from Python and loaded in R, where it can be possible to do the analysis in a more statistic way. Two tests were carried out: Lipsitz (L) and the ordinal Hosmer-Lemeshow (HL). They are best suited to detect lack of fit associated with continuous covariates [7]. The results reported in Table XI suggest that it has a correct specifications for both the model.

Another comparison is on the AIC and BIC. They are metrics used for model selection, the main difference is that the BIC penalizes more for additional parameters than AIC. The results are reported on Table X and suggest that the model with the main regressors and the Elo feature has better values.

The last metric that was taken in consideration is the precision and, as it can see in the Table XII, the best model is the one with the regressors *Cup_Home/Away*, since the precision is about 54.34%. The model with the main regressors and Elo term has a precision about 53.96%.

Like it said before, the model with the three main covariances, the *ratio_Elo* and the *Home/-Away_Diff_Goals* is the best one in all the metrics take in consideration. The analysis of the models

TABLE X

Cup_Home/Away	
AIC	BIC
508.64	537.28
Main Regressors + ratio_Elo	
AIC	BIC
504.87	526.35

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM Home/Away_Performance and ratio_Elo as fixed regressors

TABLE XII

Cup_Home/Away			
Win	A	D	H
A	63	5	25
D	30	3	38
H	20	3	78
Main Regressors + ratio_Elo			
Win	A	D	H
A	64	3	26
D	28	2	41
H	22	2	77

Parameters from OLM with Winvalue as dependent variable, the first case is with log_ratio_ln_TM Home/Away_Performance and ratio_Elo as fixed regressors stops here, as it has been chosen not to continue with the calculation of the residuals and the Shapiro test due to the calculation method used. *resids()* (from the *sure* R library) uses random sampling from a continuous distribution [15], so it doesn't get identical residuals in all runs. This implies that the Shapiro test is sometimes positive and sometimes negative.

E. Standings

All the OLMs were built, plotting the expected points can be useful to understand if the generalization done step by step starting from the basic OLM with constant team value to the model with the Elo factor was a good approach.

The expected points for each match and each team are easy to calculate, once obtained the probabilities fitted by the chosen model, the formula that has to be used is the following:

$$xP_{i,j} = \sum_{j=1}^{n} 3 \cdot (\mathbb{P}_w)_j + (\mathbb{P}_d)_j \qquad \forall i \text{ team}$$

where j are the matches. Since probabilities calculated from the models refer to the home team, when it comes to calculating the expected points for the away team is important to use the probability of the lose $\mathbb{P}_l$ instead of $\mathbb{P}_w$. For this reason is simpler divided the database *Games* in matches played for each team as the home one (dataset *Home_Team*) and matches played as away team (dataset *Away_Team*).

Afterwards, expected points are calculated from these two database, merging them for each team and summing all the points made for the model in analysis. Doing this for all the 4 models, it can be possible to visualize the standings predicted.

The interpretation of the models accuracy with only the standings isn't easy, for this it was defined a score which works as follow:

$$S_k = \sum_{i=1}^{20} 20 - |(R_t)_i - (R_k)_i| \quad \forall k \text{ model}$$

where R_t is the true rank for the i team, while R_k is the predicted one by the model k for the team i.

Note that in the 1st place there are two teams, hence the models that predicted these teams in one of the first two positions are rewarded with 20 points of score. This happens also for the 15th place, where three teams are in that position. In this case the models that predicted these teams in one of the 14th, 15th or 16th positions are rewarded with 20 points of score.

F. GLM npxG_Expected AC Milan

The last model has the goal to forecast the no penalty expected goal of AC Milan using the generalized linear model (GLM). The random component of a GLM consists of a response variable y with independent observations (y_1,...,y_n) from a distribution having probability density y_i belonging from the natural exponential family [1].

In this case the dependent variable is *npxG_Expected*, that has only positive value. Such variables usually have distributions that are right skew, because the boundary at zero limits the left tail of the distribution [6]. As it can be see from Figure 6, this is the case of the response variable in analysis. Hence, there are two possible distribution to chose in order to model it: the gamma and the inverse normal distributions. The second one is used when the responses are even more skewed than suggested by the gamma distribution.

It was modeled four GLMs, in particular three are gamma distributed while the last is the inverse

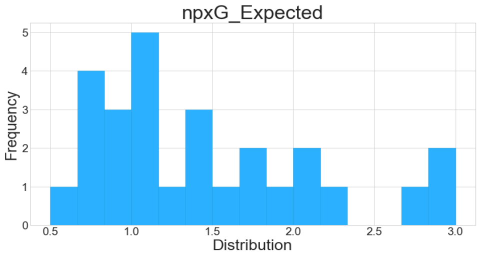


Fig. 6: Distribution of AC Milan no penalty expected goal

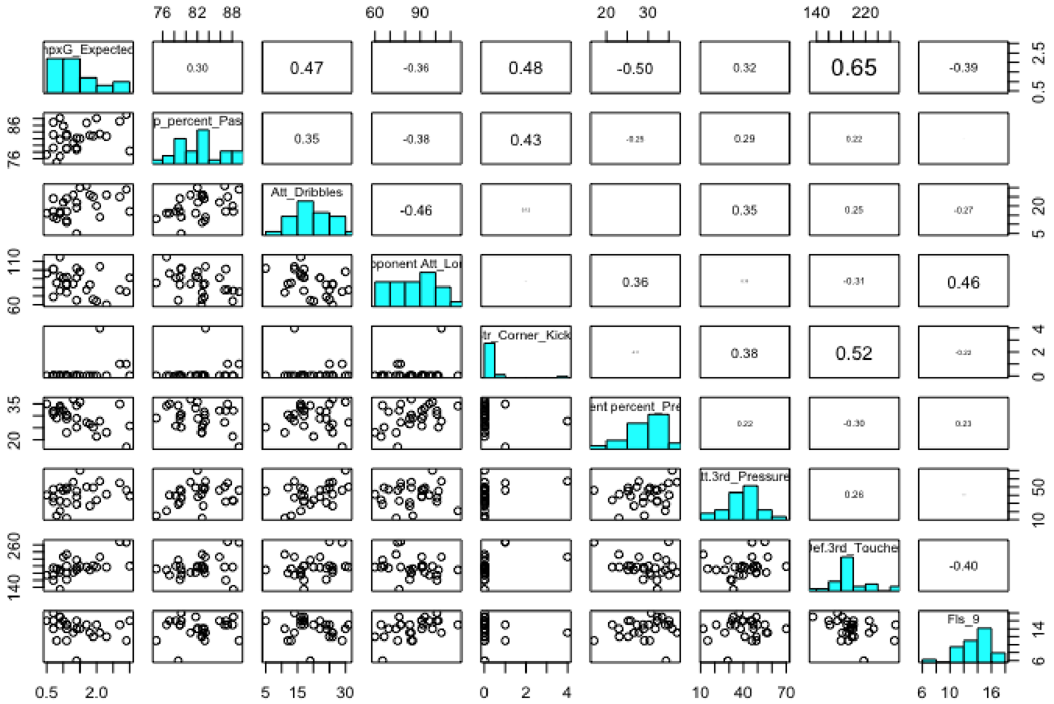


Fig. 7: Plot, distribution and Spearman correlation for the main statistics

normal. The data were partitioned into train and test dataset. The train one is composed by all the matches in the first half of the season (until 19^{th} matchweek) while the test has the remain games. Models are fitted using the logarithmic link function. It's the link function most commonly used for gamma and inverse Gaussian GLMs, to ensure $\mu > 0$ and for interpretation purposes [6]. Before proceed to the study of these models it's necessary explain why some features are selected from the *SerieA* database. Once extracted all the statistics in AC Milan matches, the csv is saved and imported in R to start the analysis. Fixed the response variable *npxG_Expected* and using two loops in the code, it was possible to calculate the Spearman correlation between the main variable and all the other feature. The choice of this correlation rather than Pearson is because to use the data must meet the several requirements such as to be continuous (many are count data) and normally distributed [19]. The selection of feature was made fixing the threshold of correlation at 0.3 in absolute terms and keeping only "general" variables. More precisely, features such as shoot or assist can't be selected since the goal of this model is to be used to do forecasting using variables that can be factors that AC Milan can work on during the training week before each match.

In the following sections will be reported the three least performing models. The statistics results are in Table XIII while metrics value (Cox & Snell R squared, AIC and deviance) are in Table XV

1) Forward Selection: In the first model the forward selection is used, in particular was

TABLE XIII

Feature	Beta	Std. Err.	z	P-value	0.025	0.975
Forward Selection						
In	1.5942	1.528	1.044	0.297	-1.400	4.588
RE	-0.1855	0.950	-0.195	0.845	-2.047	1.676
DT	0.0015	0.003	0.486	0.627	-0.004	0.007
AP	0.0160	0.006	2.825	0.005	0.005	0.027
O%P	-0.0784	0.018	-4.278	0.000	-0.114	-0.042
MP	0.1336	0.089	1.503	0.133	-0.041	0.308
Outlier Removal						
Feature	Beta	Std. Err.	z	P-value	0.025	0.975
In	3.1797	1.795	1.771	0.077	-0.339	6.699
O%P	-0.1369	0.065	-2.109	0.035	-0.264	-0.010
DT	-0.0073	0.008	-0.917	0.359	-0.023	0.008
O%P:DT	0.0003	0.000	0.991	0.321	-0.000	0.001
AP	0.0148	0.004	3.586	0.000	0.007	0.023
MP	0.0935	0.063	1.488	0.137	-0.030	0.217
MC	-0.0297	0.155	-0.192	0.848	-0.334	0.274
Inverse Gaussian						
Feature	Beta	Std. Err.	z	P-value	0.025	0.975
In	4.5243	2.114	2.140	0.032	0.381	8.668
O%P	-0.2059	0.074	-2.794	0.005	-0.350	-0.061
DT	-0.0133	0.009	-1.416	0.157	-0.032	0.005
O%P:DT	0.0006	0.000	1.791	0.073	$-5.85 \cdot 10^{-5}$	0.001
AP	0.0136	0.005	2.716	0.007	0.004	0.023
MP	0.1257	0.076	1.654	0.098	-0.023	0.275
MC	0.2143	0.161	1.331	0.183	-0.101	0.530

Parameters from GLM with npxg_Expected as dependent variable

through the function *SequentialFeatureSelector* with the floating algorithms. It have an additional exclusion or inclusion step to remove features once they were included (or excluded), so that a larger number of feature subset combinations can be sampled. It is important to emphasize that this step is conditional and only occurs if the resulting feature subset is assessed as "better" by the criterion function after removal (or addition) of a particular feature [20].

After having adjust old variables in order to obtain a more clear partition between features about AC Milan and feature about its opponent, the forward selection can start. The max number of covariates fixed is 6, the score to optimize is the mean gamma deviance. The variables found are: *ratio_Elo, Def_Touches, Att_Pressures, Opponent_percent_Pressures, Milan_Performance*. *Def_Touches* is the number of touches in the defensive 1/3 by AC Milan, *Att_Pressures* is the number of times AC Milan applying pressure to opposing player who is receiving, carrying or releasing the ball, in the attacking 1/3. *Opponent_percent_Pressures* is the percentage of time the opponent gained possession withing five seconds of applying pressure.

TABLE XIV

True	Forward Selection	Outlier Removal	Inverse Gaussian
1.4	0.690440	0.718050	0.793174
2.7	1.194978	1.311641	1.756054
1.8	1.305310	1.261842	1.633011
0.7	0.988338	0.960315	0.878712
1.4	1.124090	1.117067	1.089254
1.5	1.210538	1.128428	1.205066
1.3	0.905569	0.852940	0.730239
0.7	1.146882	1.121664	1.072629

Parameters from GLM with npxg_Expected as dependent variable

TABLE XV

CS R^2	AIC	Deviance
Forward Selection		
0.9113	19.5669	0.2762
Outlier Removal		
CS R^2	AIC	Deviance
0.9912	9.6372	0.2613
Inverse Gaussian		
CS R^2	AIC	Deviance
0.9690	18.2462	0.1690

Parameters from GLM with npxg_Expected as dependent variable

From Table XIII, all the features, except for the *Att_Pressures* and *Opponent_percent_Pressures*, are statistically non-significant since P-values are over 5% and the null value is within the CIs. The predictions on the test data are in Table XIV. In Table XV the Cox & Snell R squared is very high and the deviance on the test data seems good. Plotting the residual (Figure 8 it can see that there is an outlier, the same that will be remove in the next model. Doing the Shapiro-Wilk test, in order to test the normality of the residuals, p-value is

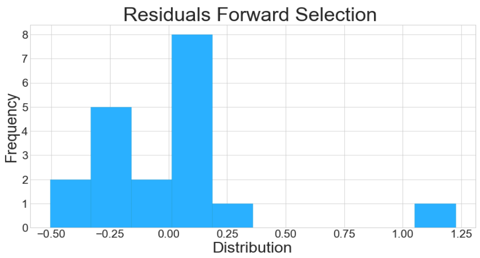


Fig. 8: Distribution of residual of GLM with forward selection, dependent variable npxG_Expected

about 0.0008, therefore the null hypothesis has to be refused.

2) Outlier Removal: Like said before, the outlier detected refers to the match against Atalanta, where AC Milan had a no penalty expected goal equal to 3. In this case it was attempt to remove it in order to see the effect of its absence. There are new features in this model, *Milan_Cup* and the interaction between *Opponent_percent_Pressures* and *Def_Touches*. From Table XIII, all the features, except for the *Att_Pressures* and *Opponent_percent_Pressures*, are statistically non-significant since P-values are over 5% and the null value is within the CIs. The predictions on the test data are in Table XIV. In Table XV the Cox & Snell R squared is higher then the model before and it's almost the max value possible (equal to 1). The deviance on the test data is slightly better, while AIC is much lower. Plotting the residual (Figure 9) it can see that there aren't outliers anymore. Doing the Shapiro-Wilk test, in order to test the normality of the residuals, p-value is about 0.4283, therefore the null hypothesis can't be refused.

3) Inverse Gaussian: The last model is based on a change in the distribution function. Like said before, it is used when the responses are even more skewed than suggested by the gamma distribution. In this case there aren't new features compared to the previous model, since the forward selection process was used to subset the most important variables. They are the start point for this model and the best one, using also different combinations of them. From Table XIII, all the variables, except for the intercept, *Att_Pressures* and *Opponent_percent_Pressures*, are statistically non-significant since P-values are over 5% and the null value is within the CIs. The predictions on the test data are in Table XIV. In Table XV the Cox & Snell R squared is higher then the model with forward selection but lower then the previous one before and it's almost the max value possible (equal to 1). The deviance on the test data is much better then the other two models, while AIC is slightly lower then the model with feature selection but much higher then the previous one. Plotting the residual (Figure 10) it can see that there are two outliers. Doing the Shapiro-Wilk test, in order to test the normality of the residuals, p-value is about 0.04248, therefore the null hypothesis has to be refused.

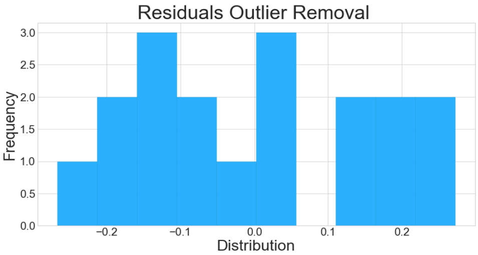


Fig. 9: Distribution of residual of GLM with outlier removal, dependent variable npxG_Expected

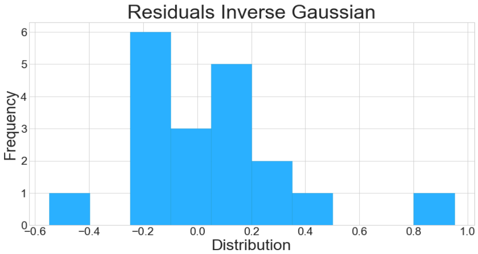


Fig. 10: Distribution of residual of GLM with inverse Gaussian as distribution, dependent variable npxG_Expected

III. Results and Discussions

After that every model described before was important in order to find the best one for each case in analysis. Here it will be reported all the results of the most performing models.

A. Odds Model

In Table XVI is reported the confusion matrix of the model. The precision of the model is about 51.32%, in particular the model predict better home wins (about 78.22%) rather then of the away wins (about 61.29%). Draws aren't predicted since they are a very rare event, this simple model fails to detect them.

TABLE XVI

Odds Model		
Win	A	H
A	57	36
D	31	40
H	22	79

Parameters from Odds model

B. Ordered Logit Model with TM Value Constant

In the section about these models, the best one was the one with only the main regressors (*log_ratio_Value, Home/Away_Performance*). In Table XVII is reported the summary of the OLM. All the features are statistically significant, for both P-value and CIs. This is not the case for the previous models of this section. In Table XVIII both Lipsitz (L) and the ordinal Hosmer-Lemeshow (HL) tests refused the hypothesis of good specification of this model. The competitors of it, instead, pass the test, but for all the other metric this model is the best. Indeed, how it can see in Table XIX, both AIC and BIC are the lower ones. For the BIC it can be predictable since the model has less regressors and this metrics penalized more then AIC, but for AIC no and this is a point in favor of this model. Another metric in its favor is the precision, indeed in Table XX is reported the confusion matrix and it can see that its values is about 54.34%, the higher one compared to the others. Hence, once proofed that this model is the best one, the comparing with the odds model can start. Focusing about the precision again, this model has a much higher value, in particular the precision on away wins is strongly better (about 70.97%) while the precision on home wins is slightly lower (about 76.24%). In the end, this model is better then the odds one.

TABLE XVII

OLM TM Value Costant					
Feature	Beta	Std. Err.	P-value	0.025	0.975
LRV	1.3323	0.1936	0.0000	0.9528	1.7118
PH	-0.2447	0.1063	0.0213	-0.4531	-0.0363
PA	0.2457	0.1088	0.0240	0.0324	0.4590

Parameters from OLM with TM value costant and with Winvalue as dependent variable while log_ratio_Value, Home/Away_Performance as regressors

TABLE XVIII

OLM TM Value Costant			
Test	LR|χ^2	df	P-value
L	18.385	9	0.03096
HL	36.859	17	0.003516

Parameters from OLM with TM value costant and with Winvalue as dependent variable while log_ratio_Value, Home/Away_Performance as regressors

TABLE XIX

OLM TM Value Costant	
AIC	BIC
515.28	533.17

Parameters from OLM with TM value costant and with Winvalue as dependent variable while log_ratio_Value, Home/Away_Performance as regressor

TABLE XX

OLM TM Value Costant		
Win	A	H
A	66	27
D	33	38
H	23	78

Parameters from OLM with TM value costant and with Winvalue as dependent variable while log_ratio_Value, Home/Away_Performance as regressors

C. Ordered Logit Model with TM Value Lineups

In the section about these models, the best one was the one with the main regressors (*log_ratio_ln_TM, Home/Away_Performance*) and the two new entry *Home/Away_Diff_Goals*. In Table XXI is reported the summary of the OLM. All the features are statistically no-significant for both P-value and CIs, except for *log_ratio_ln_TM*. In Table XXII both Lipsitz (L) and the ordinal Hosmer-Lemeshow (HL) tests accept the hypothesis of good specification of this model, as the competitors of it but for all the other metric this model is the best. Indeed, how it can see in Table XXIII, AIC is the lower ones while the BIC doesn't but this was predictable since the model with the main covariates has less regressors and this metrics penalized more models with more feature rather then AIC metric. Another metric in its favor is the precision, indeed in Table XXIV is reported the confusion matrix and it can see that its values is about 56.23%, the higher one compared to the others. Hence, once proofed that this model is the best one, the comparing with the TM value constant model can start. Focusing about the precision again, this model has a higher value, in particular the precision on away wins is better (about 73.11%) while the precision on home wins is slightly lower (about 75.25%). Now this model is able to predict few draws, ideed it has 5 correct prediction on them compared to the 0 predicted by the TM value constant model. Also AIC and BIC are lower, in particular the BIC value is an important signal that this model is better since it has more regressors but a better BIC. About the statistical tests (Lipsitz and the ordinal Hosmer-Lemeshow) this model pass them, while the previous not. In the end, this model is better then the TM value constant one.

TABLE XXI

OLM TM Value Lineups					
Feature	Beta	Std. Err.	P-value	0.025	0.975
LRL	1.4485	0.2054	0.0000	1.0460	1.8510
PH	−0.2482	0.1288	0.0540	−0.5007	0.0043
PA	0.2058	0.1230	0.0942	−0.0352	0.4468
GH	−0.0240	0.0153	0.1164	−0.0540	0.0060
GA	0.0276	0.0160	0.0856	−0.0039	0.0590

Parameters from OLM with TM value of the lineups and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance and Home/Away_Diff_Goals as regressors

TABLE XXII

OLM TM Value Lineups			
Test	LR|χ^2	df	P-value
L	7.9988	9	0.5343
HL	17.631	17	0.4125

Parameters from OLM with TM value of the lineups and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance and Home/Away_Diff_Goals as regressors

TABLE XXIII

OLM TM Value Lineups	
AIC	BIC
502.52	527.57

Parameters from OLM with TM value of the lineups and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance and Home/Away_Diff_Goals as regressors

TABLE XXIV

OLM TM Value Lineups			
Win	A	D	H
A	68	3	22
D	27	5	39
H	23	2	76

Parameters from OLM with TM value of the lineups and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance and Home/Away_Diff_Goals as regressors

D. Ordered Logit Model with Elo

In the section about these models, the best one was again the one with the main regressors (*log_ratio_ln_TM, Home/Away_Performance*) the two new entry *Home/Away_Diff_Goals* and, obviously, the new variable *ratio_Elo*. In Table XXV is reported the summary of the OLM. All the features are statistically no-significant for both P-value and CIs, except for *log_ratio_ln_TM* and *Home_Performance*. In Table XXVI both Lipsitz (L) and the ordinal Hosmer-Lemeshow (HL) tests accept the hypothesis of good specification of this model, as the competitors of it but for all the other metric this model is the best. Indeed, how it can see in Table XXVII, AIC is the lower ones while the BIC doesn't but this was predictable since the model with the main covariates and *ratio_Elo* has less regressors and this metrics penalized more models with more feature rather then AIC metric.

Another metric in its favor is the precision, indeed in Table XXVIII is reported the confusion matrix and it can see that its values is about 56.98%, the higher one compared to the others. Hence, once proofed that this model is the best one, the comparing with the TM value lineups model can start. Focusing about the precision again, this model has a slightly higher value. The precision on away wins is equal (about 73.11%) while the precision on home wins is slightly higher (about 76.24%). About draws the Elo model has a single prediction more (6 in total) then the TM lineups model. AIC and BIC are higher and the statistical tests (Lipsitz and the ordinal Hosmer-Lemeshow) were passed by both, in particular this model has a P-value much lower compared to the TM lineups model. In the end, this model can be judged better then previous one.

TABLE XXV

OLM Elo					
Feature	Beta	Std. Err.	P-value	0.025	0.975
LRL	1.3005	0.2703	0.0000	0.7706	1.8304
PH	−0.2679	0.1313	0.0413	−0.5253	−0.0106
PA	0.2226	0.1249	0.0746	−0.0221	0.4673
RE	2.7676	3.3321	0.4062	−3.7632	9.2983
GH	−0.0275	0.0159	0.0834	−0.0587	0.0036
GA	0.0315	0.0167	0.0596	−0.0013	0.0644

Parameters from OLM with Elo and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance, Home/-Away_Diff_Goals and ratio_Elo as regressors

TABLE XXVI

OLM Elo			
Test	LR|χ^2	df	P-value
L	14.772	9	0.09739
HL	25.323	17	0.08772

Parameters from OLM with Elo and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance, Home/-Away_Diff_Goals and ratio_Elo as regressors

TABLE XXVII

OLM Elo	
AIC	BIC
503.82	532.46

Parameters from OLM with Elo and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance, Home/-Away_Diff_Goals and ratio_Elo as regressors

TABLE XXVIII

OLM Elo			
Win	A	D	H
A	68	5	20
D	25	6	40
H	23	1	77

Parameters from OLM with Elo and with Winvalue as dependent variable while log_ratio_ln_TM, Home/Away_Performance, Home/-Away_Diff_Goals and ratio_Elo as regressors

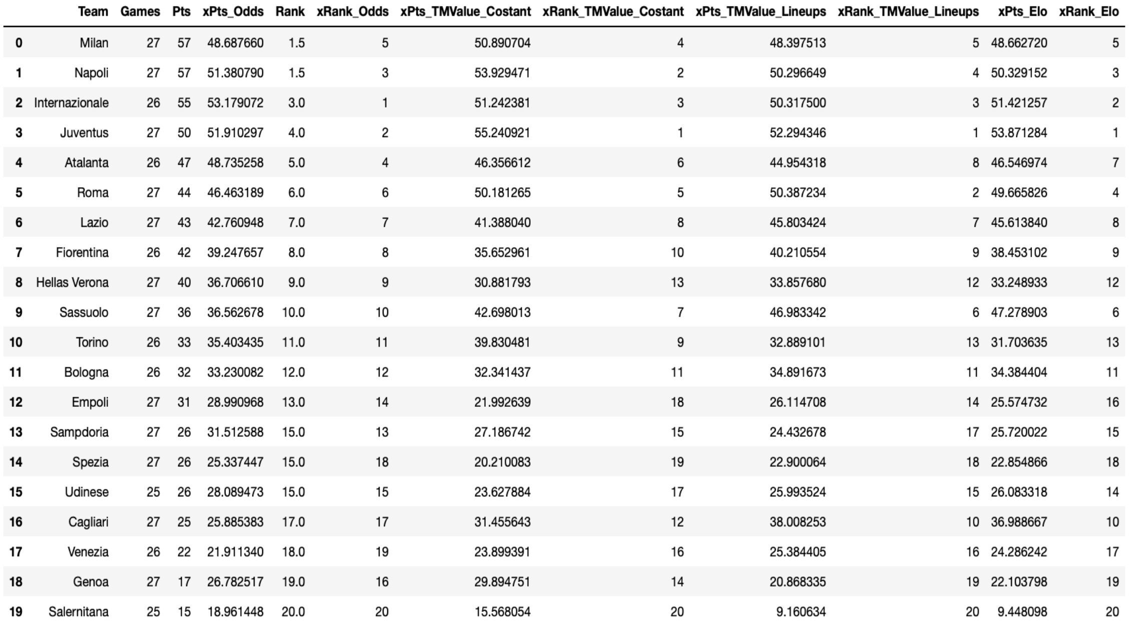
E. Standings

Once shown the best model for each section, the predicted standings can be visualized in Figure 11. Like said in the specific section, it was calculated a score that can give an idea of which of this model is the more accurate in the standing prediction. The results are reported in Table XXIX. The odds model results the best one, followed by Elo, TM Lineups and the last one TM Constant. This not change a lot the judgement about these models, the best one in statistics terms remain the TM Lineups.

TABLE XXIX

Odds	TM Constant	TM Lineups	Elo
380	356	358	361

Score on standings prediction of the best models of each section



	Team	Games	Pts	xPts_Odds	Rank	xRank_Odds	xPts_TMValue_Costant	xRank_TMValue_Costant	xPts_TMValue_Lineups	xRank_TMValue_Lineups	xPts_Elo	xRank_Elo
0	Milan	27	57	48.687660	1.5	5	50.890704	4	48.397513	5	48.662720	5
1	Napoli	27	57	51.380790	1.5	3	53.929471	2	50.296649	4	50.329152	3
2	Internazionale	26	55	53.179072	3.0	1	51.242381	3	50.317500	3	51.421257	2
3	Juventus	27	50	51.910297	4.0	2	55.240921	1	52.294346	1	53.871284	1
4	Atalanta	26	47	48.735258	5.0	4	46.356612	6	44.954318	8	46.546974	7
5	Roma	27	44	46.463189	6.0	6	50.181265	5	50.387234	2	49.665826	4
6	Lazio	27	43	42.760948	7.0	7	41.388040	8	45.803424	7	45.613840	8
7	Fiorentina	26	42	39.247657	8.0	8	35.652961	10	40.210554	9	38.453102	9
8	Hellas Verona	27	40	36.706610	9.0	9	30.881793	13	33.857680	12	33.248933	12
9	Sassuolo	27	36	36.562678	10.0	10	42.698013	7	46.983342	6	47.278903	6
10	Torino	26	33	35.403435	11.0	11	39.830481	9	32.889101	13	31.703635	13
11	Bologna	26	32	33.230082	12.0	12	32.341437	11	34.891673	11	34.384404	11
12	Empoli	27	31	28.990968	13.0	14	21.992639	18	26.114708	14	25.574732	16
13	Sampdoria	27	26	31.512588	15.0	13	27.186742	15	24.432678	17	25.720022	15
14	Spezia	27	26	25.337447	15.0	18	20.210083	19	22.900064	18	22.854866	18
15	Udinese	25	26	28.089473	15.0	15	23.627884	17	25.993524	15	26.083318	14
16	Cagliari	27	25	25.885383	17.0	17	31.455643	12	38.008253	10	36.988667	10
17	Venezia	26	22	21.911340	18.0	19	23.899391	16	25.384405	16	24.286242	17
18	Genoa	27	17	26.782517	19.0	16	29.894751	14	20.868335	19	22.103798	19
19	Salernitana	25	15	18.961448	20.0	20	15.568054	20	9.160634	20	9.448098	20

Fig. 11: Standings of the best models of each section

F. GLM npxG_Expected AC Milan

In the section about these models, the best one was the gamma regression with the following independent variables: *Opponent_percent_Pressures, Def_Touches, Att_Pressures, Milan_Performance, Milan_Cup* and the interaction between the first two. In Table XXX is reported the summary of the GLM. All the features are statistically no-significant for both P-value and CIs, except for the intercept, *Opponenet_percent_Pressures* and *Att_Pressures*. This confirm that the last two features are very important in order to describe the expected goal of AC Milan. Since the link function used is log, then:

$$
\begin{aligned}
log(\mu) = \alpha + \beta_1 \cdot O\%P + \beta_2 \cdot DT + \\
+\beta_3 \cdot O\%P : DT + \beta_4 \cdot AP + \\
+\beta_5 \cdot PM + \beta_6 \cdot CM
\end{aligned}
$$

The beta of O%P means that the an 1% increase of the feature has a decrease in mean of the expected goal equal to $e^{-0.2059} = 0.8139$, this trend is also shown in Figure 12. The beta of AP means that the an 1% increase of the feature has an increase in mean of the expected goal equal to $e^{0.0136} = 1.0137$, this trend is also shown in Figure 13. In this plot it can be see that the relation between the two variables isn't linear but it can be described using a monotonic function. If Spearman correlation was not used, the *Att_Pressures* feature would be discarded even though it was a good explanatory variable for the expected goal.

In Table XXXI are reported all the metrics used to compare models. The Cox & Snell R squared is equal to the model with the inverse Gaussian as distribution, higher then the outlier removal model but lower then the forward selection one. Hence, the model is the second best one as the inverse Gaussian for this metric. Also for the AIC, but in this case the inverse Gaussian has a higher value hence the second place is occupied by this solo model. The best one still is the outlier removal. About the deviance this model is the

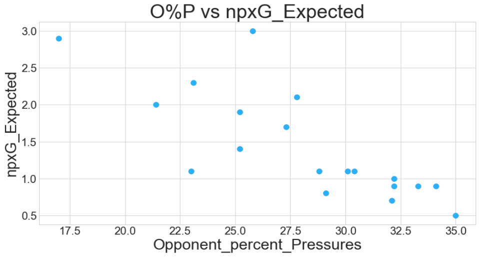


Fig. 12: Distribution observations between Opponent_percent_Pressures and npxG_Expected

TABLE XXX

Gamma Regression						
Feature	Beta	Std. Err.	z	P-value	0.025	0.975
In	4.5243	2.114	2.140	0.032	0.381	8.668
O%P	-0.2059	0.074	-2.794	0.005	-0.350	-0.061
DT	-0.0133	0.009	-1.416	0.157	-0.032	0.005
O%P:DT	0.0006	0.000	1.791	0.073	$-5.85 \cdot 10^{-5}$	0.001
AP	0.0136	0.005	2.716	0.007	0.004	0.023
PM	0.1257	0.076	1.654	0.098	-0.023	0.275
CM	0.2143	0.161	1.331	0.183	-0.101	0.530

Parameters from GLM with npxG_Expected as dependent variable and Opponent_percent_Pressures, Def_Touches, their interaction, Att_Pressures, Milan_Performance and Milan_Cup as regressors

best one by far compared to the forward selection and the outlier removal models, while is slightly better then the inverse Gaussian one.

After these analysis seems that the main models are gamma regression and the outlier removal, with this last that has two metrics on its favor. But the deviance has to be consider the most important one since is a measure of goodness of fit on the test data and not only on the train one as the two other metrics. Indeed, how it can see in Table XXXII, the predicted expected goals are better for the gamma regression model with 6 on 8 results nearer then the predicted ones by the outlier removal. Specially for the second observation, where the true expected goal are very high, like the Atalanta outlier that was removed by the model (the *npxG_Expected was* 3. With the elimination of this outlier, the model loses some variance information about the observations and when it encounters a value as high as the outlier, fails to predict a good expected goal.

The last analysis is on the residuals distribution. In Figure 14 it can be see the outlier that wasn't removed, but except for it seems that residuals ar normally distribuited. Doing the Shapiro-Wilk test, in order to test the normality of the residuals, p-value is about 0.04248, equal to the inverse Guassian model, therefore the null hypothesis has to be refused. The p-value is very close to the fixed significant level (5%) and the histogram are similar to a normal one. Indeed, with a simple removal of the outlier the p-value rises to a much higher value (0.4283) that is the value of the outlier removal model. For this reason the model chosen as the best one is the gamma regression model.

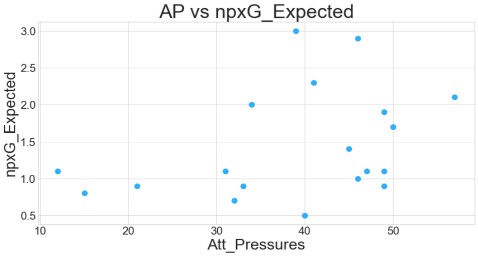


Fig. 13: Distribution observations between Opponent_percent_Pressures and npxG_Expected

TABLE XXXI

Gamma Regression		
CS R^2	AIC	Deviance
0.9690	15.7541	0.1657

Parameters from GLM with npxG_Expected as dependent variable and Opponent_percent_Pressures, Def_Touches, their interaction, Att_Pressures, Milan_Performance and Milan_Cup as regressors

Fixed the best model, it can be possible to forecast of the matches played during the writing of this report. To predict the no penalty expected goal before the matches it was used the means of the value that regressors had during the season. More precisely, the *Opponent_percent_Pressures* used is the mean of the values that the opponent team had in the previous matches, idem for *Att_Pressures, Def_Touches* of AC Milan. The *Milan_Performance* and *Cup_Milan* are always known before a match. The real value of the *npxG_Expected* is taken from FBref after the match, in order to see how much is the distance between the prediction and the true value. After the match, can be also possible to forecast the value using the real statistics of that game, in order to analysis the accuracy of the model but with real data. The results are reported in Table XXXIII. As it can be see, using real data gives a better prediction of the expected goal, as is normal. Only in the last case the mean values return a better prediction.

TABLE XXXII

True	Outlier Removal	Gamma Regression
1.4	0.718050	0.759713
2.7	1.311641	1.821432
1.8	1.261842	1.640999
0.7	0.960315	0.853408
1.4	1.117067	1.074074
1.5	1.128428	1.340863
1.3	0.852940	0.738110
0.7	1.121664	1.073692

Predictions using parameters from GLM with npxg_Expected as dependent variable

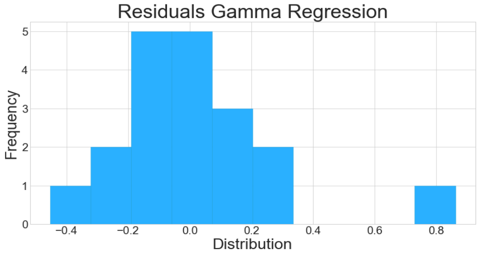


Fig. 14: Distribution of residual of gamma regression with dependent variable npxG_Expected

TABLE XXXIII

True	Means	Real Data
1.1	1.223065	1.149154
0.7	1.470537	1.32834
2.4	1.456363	0.762656

Predictions using parameters from GLM with npxg_Expected as dependent variable

IV. Conclusion

In the end, this study gives back two good models for classification and forecasting. The OLM with TM values of the lineups improved a lot the precision of the base model with the odds. The odds model certainly takes into account other unknown factors at this time, but the results are good just using general variables with good explanatory power. A further generalization could be to take into account events that have occurred inside the locker room, such as some disagreement between players and coach. This could be modeled using a dummy variable with a value equal to 1 in the event of a particular event or 0. The exemption by the coach and the advent of a new coach could also be modeled in this way.

For the gamma regression model, the feature selection was done in a very deep way, using Spearman correlation and the forward selection method. It could try to find other possible interaction variables such as the *Opponent_percent_Pressures* and *Def_Touches*.

References

[1] A. Agresti. *Foundations of Linear and Generalized Linear Models*. Wiley Series in Probability and Statistics. Wiley, 2015.

[2] I. Asimakopoulos and J. Goddard. Forecasting football results and the efficiency of fixed-odds betting. *Journal of Forecasting*, 23:51–66, 01 2004.

[3] K. Benoit. Linear regression models with logarithmic transformations, 2011.

[4] D. Cortis. Expected values and variances in bookmaker payouts: A theoretical approach towards setting limits on odds. *Journal of Prediction Markets*, 9(1):1–14, 2015.

[5] M. Dixon and S. Coles. Modelling association football scores and inefficiencies in the football betting market. *Journal of the Royal Statistical Society: Series C*, 46:265 – 280, 1997.

[6] P. Dunn and G. Smyth. *Generalized Linear Models With Examples in R*. Springer Texts in Statistics. Springer New York, 2018.

[7] M. W. Fagerland and D. W. Hosmer. Tests for goodness of fit in ordinal logistic regression models. *Journal of Statistical Computation and Simulation*, 86(17):3398–3418, 2016.

[8] L. Grilli and C. Rampichini. Ordered logit model. In A. C. Michalos, editor, *Encyclopedia of Quality of Life and Well-Being Research*, pages 4510–4513. Springer Netherlands, Dordrecht, 2014.

[9] S. Gupta. The relevance of confidence interval and p-value in inferential statistics. *Indian journal of pharmacology*, 44:143–4, 02 2012.

[10] https://www.statistichesulcalcio.com/studioquote_seriea2022.php.

[11] https://fbref.com/en/comps/11/schedule/Serie-A-Scores-and-Fixtures.

[12] https://jaseziv.github.io/worldfootballR/index.html.

[13] https://www.transfermarkt.it/serie-a/marktwerteverein/wettbewerb/IT1.

[14] https://github.com/Shopify/bevel.

[15] https://cran.r-project.org/web/packages/sure/sure.pdf.

[16] https://www.transfermarkt.it/serie-a/marktwerte/wettbewerb/IT1.

[17] https://raw.githubusercontent.com/JaseZiv/worldfootballR_data/master/raw-data/fbref-tm-player-mapping/output/fbref_to_tm_mapping.csv.

[18] http://clubelo.com.

[19] https://libguides.library.kent.edu/SPSS/PearsonCorr.

[20] http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/.