Calculating the Probability of a Shot Leading to a Goal in Soccer

Chirag Sreenivasamurthy Panchakshari, Karthik Garimella, Sandeep Alfred

CS 5841 - Group - 27

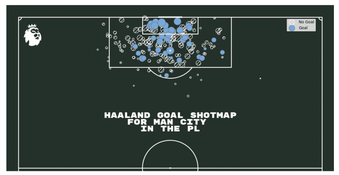
Michigan Technological University

Objective

Not all shots taken by strikers lead to a goal. The conversion rate of the strikers in the top 5 European men's soccer leagues (English Premier League, La Liga, Bundesliga, Serie A and Ligue 1) is barely above 30%.

Every shot taken has a certain probability to end up in goal. That probability ranges from 0 - 1 and it's called **xG (Expected Goals)**. A shot with high xG will be close to 1 and a low xG will be close to 0.

The aim is to calculate the probability of a shot taken by a striker in the top 5 men's European league. Given the X and Y coordinates of the shot taken on the pitch, the body part used for taking that particular shot, in which situation the shot was taken, how did the player taking the shot receive the ball, the result of the shot (i.e. did it result in a goal or not).



Haaland Shotmap: The bigger scatter points indicate a higher xG value.

Dataset

Source: Understat player shot map.

Dimensions of the data: 98,723 records of shots over 8 features.

Target Feature: xG (Expected Goals)

Web Scraped (using selenium) all-time shot history of 586 players positioned forward, in the top 5 European Leagues.

Feature	Description
X	x coordinate of the player taking the shot inside the penalty box. {numeric}
Y	y coordinate of the player taking the shot inside the penalty box. {numeric}
h_a	Whether the game is a home match or an away match for the player. {categorical}
situation	Which situation led to the shot being taken {categorical}
shotType	Which body part was used to take the shot {categorical}
lastAction	What the last action was before the player received the ball {categorical}
result	Outcome, whether it was a goal or not {categorical}
min	The minute the shot was taken. {numerical}
xG	Probability of the player scoring a goal. {numeric}

Preprocessing:

Dropped **10,246** records of missing categorical values from lastAction feature. We classified all the shots not leading to a goal (Missed, Saved, Blocked, Shot On Post) as not a goal. Condensed data values for lastAction feature from 31 to 4 based on similarity of values.

One hot encoded all categorical features. We ignored all the own goals and penalties. All players with more than 40 shots have been considered. The minute column was normalized using StandardScaler.

Methods

Since this is a regression problem, several models were explored and subsequential improvements were considered based on mean squared errors.

Models Considered:

- **Lasso Regression**, **Ridge Regression**, **ElasticNet**, **RandomForestRegressor**, and KNeighbhorsRegressor. These models were run through a 5 fold **GridSearchCV** with several hyperparameters to select the best model. The models were regressed on the xG feature and was evaluated on Mean Squared Error. Accuracy was measured using R2 score.
- We also considered **Artificial Neural Networks** for its pattern recognition. We used 5 hidden layers ranging from 128 to 512 neurons. This model used **LeakyReLU** activation function as traditional ReLU makes all the negative values zero. R2 score and MSE were used as the evaluation metrics. The optimizer used was Adam with varying learning rates.

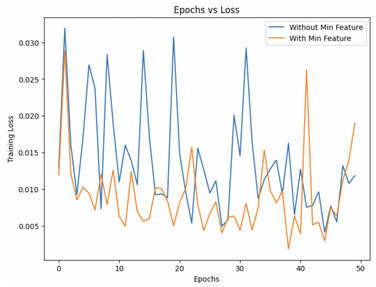
Evaluation

RandomForestRegressor and **Artificial Neural Networks** performed the best compared to **Lasso Regression**, **Ridge Regression**, **ElasticNet**.

Model Training and Testing Performance

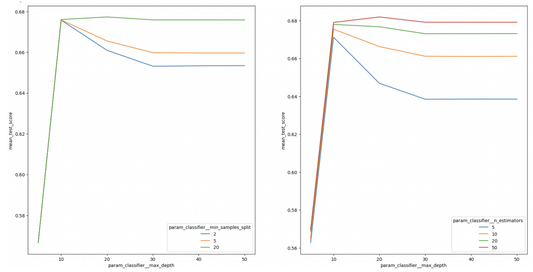
Model	Training R2 Score
Ridge	0.403356
Lasso	0.175933
ElasticNet	0.219582
KNeighborsRegressor	0.509554
RandomForestRegressor	0.685545
ANN	0.690728

Model	Testing R2 Score
RandomForestRegressor	0.691132
ANN	0.679056



ANN loss for each epoch with and without considering the "minute" column.

Random Forest Hyperparameters vs R2 Score



Random Forest was considered with several parameters, the graph above breaks down each parameter's performance to R2 score.

Improving the model:

Initially, Artificial Neural Networks (ANNs) were assessed without including the minute column in the baseline model. However, adding this feature led to a noticeable decrease in loss, signifying improved model performance.

Random Forest Regressor emerged as the best estimator after adding the minute column, through GridSearchCV, its testing performance was even better compared to ANN.

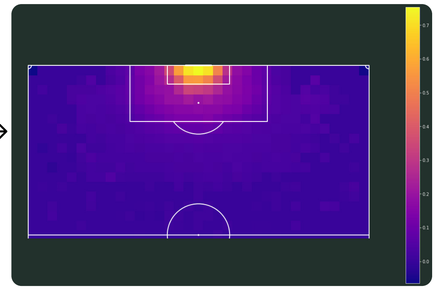
Result

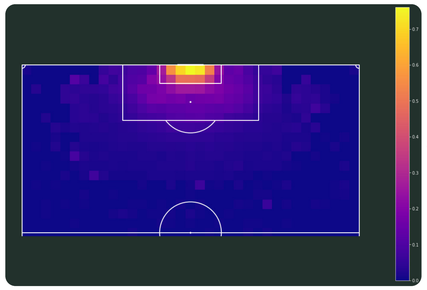
The Model that best explains the data is found to be the **RandomForestRegressor()** with:

- max_depth : **20**, min_samples_split : **20** and n_estimators : **50**

The probabilities estimated by the model were plot using **mplsoccer** library, as seen in the heatmaps, the true values and estimated values are closely similar.

Heatmap of the **estimated xG** values by the model, brighter points indicate a higher xG value.





Heatmap of the **true xG** values from the data, brighter points indicate a higher xG value.

Conclusion

Based on the heatmaps above, the closer a player is to the goal, the better the probability of shot leading to a goal. As shots are taken further away from the goal, the probability is low but not zero.

There are certain points further away from the goal which have better probability than some closer points, this is mostly due to the situations created during the game that lead to a higher xG. Due to lack of publicly available data, modelling every aspect before a player takes a shot to calculate xG is seemingly challenging.

References: Anzer G and Bauer P (2021) A Goal Scoring Probability Model for Shots Based on Synchronized Positional and Event Data in Football (Soccer). Front. Sports Act. Living 3:624475. doi: 10.3389/fspor.2021.624475

Data Scraped from: https://understat.com/