Predicting goals from shots in football using machine learning

UPPSALA UNIVERSITET

Project in Computational Science
Fall Term 2019

Project members:
Sari Shaybany
Chun Wu

Supervisor:
David Sumpter

Course Coordinator:
Maya Neytcheva

In collaboration with:
Twelve
Hammarby football club

Abstract

Models were created using logistic regression and XGBoost, predicting whether a shot from a certain distance or angle to the target will result in a goal. The models were trained with pre-processed event data of matches in Allsvenskan.

The results suggest a higher accuracy in predicting goals for XGBoost, with an AUC of 0.774, compared to LR, for which the AUC was 0.766.

Aim

The aim was to create models using logistic regression and XGBoost that calculates the probability of scoring football goals from various positions of the pitch in order to classify future shots as goals or misses. Also, comparing the accuracy of the two models.

Background

Using machine learning to predict different events in football is valuable due to its good accuracy and application in fields were large amounts of money are involved (i.e. the betting industry, for club managers/owners to formulate strategies to win matches, etc.).

Several models with varying accuracy have been proposed to model the quality of play in football; here the focus was on shots, Raw data was given, describing match events the last three seasons of Allsvenskan.

Methods

From the raw data of detailed match events (Fig.1) the co-ordinates of the shot positions were extracted for events qualified as shots. The OPTA "qualifiers" specified in the data are descriptors of certain details of the events, represented by numbers in the data.


```
[[{"eventId": 1375794403, "eventTypeId": 1, "outcome": true, "min": 0, "sec
"playerId": 192774, "timeStamp": "2019-06-25T18:02:51.41", "qualifiers": [
"value": "17.6"}, {"qualifierId": 56, "value": "Back"}, {"qualifierId": 27
{"qualifierId": 140, "value": "33.9"}, {"qualifierId": 213, "value": "3.4"
"value": "43.3"}], "x": 49.9, "y": 50.1, "description": "Kick-off"}, {"eve
"eventTypeId": 1, "outcome": false, "min": 0, "sec": 3, "teamId": 483, "pl
"timeStamp": "2019-06-25T18:02:53.951", "qualifiers": [{"qualifierId": 213
{"qualifierId": 141, "value": "88.0"}, {"qualifierId": 140, "value": "82.5
"value": "59.0"}, {"qualifierId": 1}, {"qualifierId": 56, "value": "Left"}
"x": 65.4, "y": 57.5, "description": "Failed Pass"}, {"eventId": 140277081
"outcome": false, "min": 0, "sec": 7, "teamId": 195, "playerId": 224856, "
"2019-06-25T18:02:57.966", "qualifiers": [{"qualifierId": 141, "value": "1
{"qualifierId": 56, "value": "Back"}, {"qualifierId": 212, "value": "4.1"
"value": "6.0"}, {"qualifierId": 140, "value": "21.6"}], "x": 17.8, "y": 1
```


Fig. 1. Raw data of match events.

The co-ordinates were used to calculate the distance to the target and the angel between the two goal posts, both of which serving as input arguments (X) to the models, among others. In logistic regression, the probability P of scoring is given by the logistic function:

$$P = \frac{1}{1 + e^{-(a+bX)}}$$

XGBoost was used via open-source software library providing gradient boosig framework.

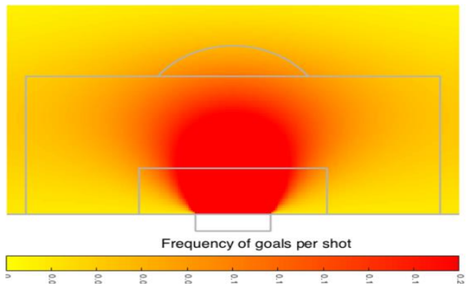

Fig. 2. Typical goal frequencies from various locations around the box, by David Sumpter, Uppsala University, 2017.

Results

The accuracy of the models was determined by the AUC, which is the area under the ROC-curve (Fig.3).

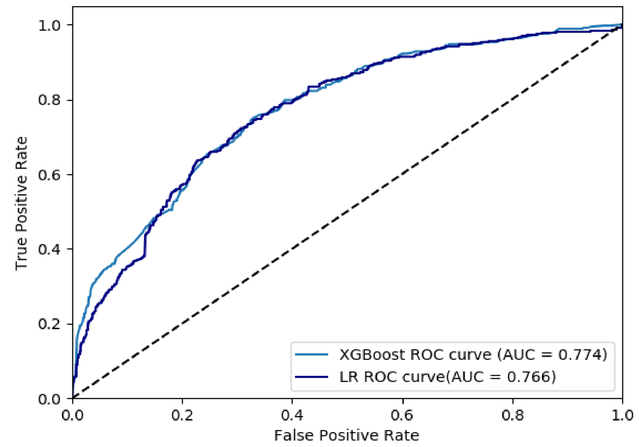

Fig. 3. ROC-curves of logistic regression (AUC = 0.766) and XGBoost (AUC = 0.774).

Discussion & Conclusion

The results (Fig.3) show a bigger AUC for XGBoost than for logistic regression, suggesting that the XGBoost model is more capable of distinguishing between goals and misses. Although, the difference in AUC is relatively small (within the second decimal) .

Literature Cited

Sumpter, D 2017, digital image, David Sumpter, viewed 9 January 2020, <https://medium.com/@Soccermatics/the-geometry-of-shooting-ae7a67fdf760>.