Data Mining 1: Predicting positions of soccer players Project Report

Team No. 9

presented by
Jurgen Amedani (1628022), Cara Maria Damm (1631263), Paul Hesselmann (1371380), Allwyn Menezes (1671634), Martin Mlinac (1364487), Carmen Rannefeld (1631070)

submitted to the
Data and Web Science Group
Prof. Dr. Bizer
University of Mannheim

1 Application Area and Goals

2 Structure and Size of Data

In the context of football analytics, the FIFA 19 dataset is used including information about every player registered database of www.sofifa.com [2]. More than 18207 players are registered in the moment the dataset was created. The FIFA 19 complete dataset includes 86 different attributes. The attributes are listed in the table below.
The key attributes are ID, name, age, nationality, value, potential, value and special which are filled in all the cases. The 86 attributes can be grouped into ten groups:

- Personal information
    - Database ID
    - Name
    - Age
    - Nationality
    - Height
    - Weight
- Contract Information
    - Club
    - Wage
    - Release Clause
    - Jersey Number
    - Joined date
    - Loaned from
    - Contract valid until
- Mentality Statistics
    - Aggression
    - Interceptions
    - Positioning
    - Vision
    - Penalties
    - Composure
- Sports and Football information
    - Value
    - International Reputation
    - Overall Ranking
    - Potential (correlation to overall ranking)
    - Special
    - Position (club)
    - Position (national team)
    - Preferred Foot
    - Week Foot
    - Skill Moves
    - Work Rate Defense
    - Work Rate Attack
    - Body Type
    - Real Face
- Attacking Statistics
    - Crossing
    - Finished
    - Heading Accuracy
    - Short Passing
    - Volleys
- Skill Statistics
    - Dribbling
    - Curve
    - Free Kick Accuracy
    - Long Passing
    - Ball Control
- Defending Statistics
    - Marking
    - Standing Tackle
    - Sliding Tackle
- Movement Statistics
    - Acceleration
    - Sprint Speed
    - Agility
    - Reactions
    - Balance
- Goalkeeping Statistics
    - GK Diving
    - GK Handling
    - GK Kicking
    - GK Positioning
    - GK Reflexes
- Power Statistics
    - Shot Power
    - Jumping
    - Stamina
    - Strength
    - Long Shots

In the dataset one player is assigned to exactly position out of 25 available positions. As players might play on different positions during a season or training the potential for 24 positions (except goalkeeper position) is recognized as well in the dataset. The position is written down as abbreviation meaning:

- Forwards:
  - RF Right forward
  - CF Center forward
  - LF Left forward
  - RS Right Striker
  - ST Striker
  - LS Left Striker
  - RW Right wing
  - LW Left wing
- Midfielders:
  - LAM Left attacking midfield
  - RAM Right attacking midfield
  - CAM Center attacking midfield
  - LM Left midfield
  - CM central midfield
  - LCM Left center midfield
  - RCM Right center midfield
  - RM Right midfield
  - LDM Left defensive midfield
  - CDM center defensive midfield
  - RDM Right defensive midfield
- Defensive:
  - LWB Left Wing Back
  - RWB Right Wing Back
  - LCB Left center back
  - LB Left back
  - RCB Right center back
  - CB Center back
  - RB Right back

These position potentials are not applied for goalkeepers. Not all the attributes have filled values. One reason is that for example not every player is loanded from another club. Other reason is the missing information as some players are young and not every information is yet collected. There are for example 48 players were only key attributes like personal information are filled. A reason might that they are new in the database not better described yet.
Attributes of the statistics groups are continuous data with a range from 0 to 100. Other attributes are categorical data. The dataset includes numerical attributes like wage, release clause, weight and height.
The wage and release clauses starts from a value of 1000 while the specialty attribute is not smaller than 731. Release clause and wage are reported in a format including k (thousand) and m (million). These values have been converted. The height is converted from foot into centimeter. The weight has been recalculated from pound into kilogram. For the attributes height and weight zero values are entered if not information is available and needed to be filtered out.

3 Preprocessing

Not all 83 attributes are necessary to predict the position of a player. We applied the optimize selection operator and analyzed the log reports from Rapidminer to select the important attributes. During one run the attributes Sliding tackle, Skill move, Long passing, LCB (Left center back), Heading accuracy, LAM (Left attacking midfield), Finishing, Crossing, Sprint speed and LB (left back) were rated with a relative importance. With no influence the other playing position results were weighted.

Attributes marked as personal and contract information were excluded from the dataset as they are not relevant to determine players position. The data mining processes started with a higher aggregation of players position which results in four groups: defender, midfielder, strikers and goalkeepers. The following concrete positions were grouped together in an attribute called Position_grouped replacing the original position attribute:

– Strikers
  - ”ST”, ”CF”, ”LF”, ”LS”, ”LW”, ”RF”, ”RS” and ”RW”
– Midfielder
  - ”CAM”, ”CDM”, ”LCM”, ”CM”, ”LAM”, ”LDM”, ”LM”, ”RAM”, ”RCM”, ”RD”, ”RM”
– Defender
  - ”CB”, ”LB”, ”LCB”, ”LWB”, ”RB”, ”RCB”, ”RWB”
– Goalkeeper
  - ”GK”

4 Data Mining

4.1 Result using Gradient Boosted Trees

The gradient boosted trees algorithm is used to create an ensemble of decision trees trough gradually improved estimations. The output is a classification model which can be applied to the test dataset for a prediction of the label attribute position_grouped. One advantage is the written report about the weights of attributes with respect to the label attribute. [1] The number of trees was set to 30. The maximal depth was initially set to 15, while the best result were scored with a maximal depth of 30. The number of bins was set to 30. The process was executed with different settings regarding the maximal depth of trees. In summary, allowing a higher maximal depth resulted in better scores for R^2, recall and precision. Result of this values are shown in table 4.1.

Run	**Max. tree depth**	**Number of folds**	**Recall**	**Precision**	R^2	**Mean squarred error**
1	10	15	88,21 %	89,01 %	64,4 %	0.33733332
2	15	10	0	0	65,3 %	0.3289592
3	25	10	0	0	65,4 %	0.3286427
4	30	15	88.09%	88.73 %	65,4 %	0.328648

Table 1. Performance results of gradient boosted trees

As important attributes the following were highlighted in table 4.1:

References

1. Rapidminer Documentation https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/gradient_boosted_trees.html. Last accessed May 12, 2019
2. Sofifa.com https://sofifa.com/players. Last accessed May, 12 2019

Variable	Relative Importance	Scales Importance	Percentage
SlidingTackle	43887.371094	1.000000	0.179607
Skill Moves	40664.996094	0.926576	0.166419
LongPassing	27751.718750	0.632340	0.113572
LCB	23368.140625	0.532457	0.095633
HeadingAccuracy	22303.111328	0.508190	0.091274
LAM	16302.590820	0.371464	0.066717
Finishing	9715.319336	0.221369	0.039759
Crossing	9004.551758	0.205174	0.036851
SprintSpeed	5180.378906	0.118038	0.021200
ShortPassing	3899.299316	0.088848	0.015958

Table 2. Important Attributes according to gradient boosted trees algorithm