CS-376b

Siyu Chen

Data Quality Plan

* Remove Several betting stats features from the raw data, because they are not relevant to the match result.
* Calculate ratio of features HST/AST (Home/Away shots on target) and HS/AS (Home/Away shots) to get HSTR /ASTR (Home/Away team shots accuracy).
* Calculate STRdf (Shots on target difference) by HSTR – ASTR.
* Calculate Fdf (Fouls committed difference by HF (home fouls) – AF (away fouls)
* Calculate Ydf (Yellow Cards difference by HY (home yellow cards) – AF (away yellow cards)
* Calculate Rdf (Red Cards difference by HR (home Red cards) – AR (away Red cards)
* Calculate Cdf (Corners difference by HC (home Corners) – AC (away Corners)
* Normalize HT (Home Stadium Attendance), because of the big attendance difference among these 17 stadiums.
* Remove all the features used to generate the new ones.
* Remove poorly correlated features (Ydf, Cdf, Fdf).