David Hu

12/29/2019

**Predicting H2H soccer results in the EPL**

**Model:** Supervised learning

**Possible techniques:**

* Logistic regression
* Classification (KNN)
* Ensemble method (hyperparameters may be important – these are to avoid overfitting, ex: flooring # of nodes in a tree):
  + Random Forest
  + Gradient boosting

**Possible Features:**

* Goals scored
* Goals conceded
* Total shots
* Shots on target
* Passes
* Possession
* Recent form
* Home / away

**Coding Resources:**

Data extraction:

* Web scraping in Python:
  + <https://www.datacamp.com/community/tutorials/web-scraping-using-python>

Logistic regression:

* Interpreting logistic regression results in sklearn:
  + <https://stackoverflow.com/questions/51006193/interpreting-logistic-regression-feature-coefficient-values-in-sklearn>
* Interpreting odds ratios in logistic regression:
  + <https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-interpret-odds-ratios-in-logistic-regression/>

**Packages used:**

* Pandas
* Scikit learn

**Topic Resources:**

* Factors influencing match result: <https://sports.stackexchange.com/questions/12233/what-are-the-main-factors-to-predict-which-team-is-going-to-win-a-soccer-match>

**Datasets:**

* EPL results:

<https://datahub.io/sports-data/english-premier-league#resource-english-premier-league_zip>

* EPL results with betting lines:

<http://www.football-data.co.uk/englandm.php>

**EPL H2H dataset:**

**Model outcome variable: Tertiary**

A / D / H (Away win, Draw, Home win)

**Features:**

| Field Name | Order | Type (Format) | Description |
| --- | --- | --- | --- |
| Date | 1 | date (%Y-%m-%d) | Match Date (dd/mm/yy) |
| HomeTeam | 2 | string (default) | Home Team |
| AwayTeam | 3 | string (default) | Away Team |
| FTHG | 4 | integer (default) | Full Time Home Team Goals |
| FTAG | 5 | integer (default) | Full Time Away Team Goals |
| FTR | 6 | string (default) | Full Time Result (H=Home Win, D=Draw, A=Away Win) |
| HTHG | 7 | integer (default) | Half Time Home Team Goals |
| HTAG | 8 | integer (default) | Half Time Away Team Goals |
| HTR | 9 | string (default) | Half Time Result (H=Home Win, D=Draw, A=Away Win) |
| Referee | 10 | string (default) | Match Referee |
| HS | 11 | integer (default) | Home Team Shots |
| AS | 12 | integer (default) | Away Team Shots |
| HST | 13 | integer (default) | Home Team Shots on Target |
| AST | 14 | integer (default) | Away Team Shots on Target |
| HF | 15 | integer (default) | Home Team Fouls Committed |
| AF | 16 | integer (default) | Away Team Fouls Committed |
| HC | 17 | integer (default) | Home Team Corners |
| AC | 18 | integer (default) | Away Team Corners |
| HY | 19 | integer (default) | Home Team Yellow Cards |
| AY | 20 | integer (default) | Away Team Yellow Cards |
| HR | 21 | integer (default) | Home Team Red Cards |
| AR | 22 | integer (default) | Away Team Red Cards |

**Derived fields (cannot use match stats for prediction, need fields that are known before the match):**

* Average Home Goals, Average Away Goals
* Average HT Home Goals, Average HT Away Goals
* Average Home Shots, Average Away Shots
* Average Home SOT, Average Away SOT
* Average Home FC, Average Away FC
* Average Home Corners, Average Away Corners
* Average Home YC, Average Away YC
* Average Home RC, Average Away RC
* Form:
  + Consecutive wins
  + Consecutive losses
  + Draws?
* League position:
  + Challenging for top 5
  + Relegation battle
  + Nothing to play for
* Moving averages (can be used to account for form)

**Other measures to consider:**

* Betting lines: can compare probability from logistic regression to the betting odds to determine favorable/unfavorable bets

**Goal:** Determine which features have a greater effect on winning

**Approach:** Analyze the logistic regression outputs

**Initial model:** Logistic regression using static results (post-match, halftime metrics)

**Predictors:**

* HS
* AS
* HST
* AST
* HF
* AF
* HC
* AC
* HY
* AY
* HR
* AR

**Predictive accuracy: approx. 53-58%**

**Baseline accuracy ~46% (Home team win %)**

* Note this is probably adjusted slightly higher once you factor in opponent strength

**For reference:**

* Home win %: 0.46275880613068027
* Away win %: 0.28824952944339877
* Draw %: 0.24899166442592094

**Model performance (to check for Home/Away bias):**

* Home win %: 66.5%
* Away win %: 31.1%
* Draw %: 2.4%

Static model overpredicts Home wins and underpredicts Draws

**Coefficient outputs:**

* Classes: ['A' 'D' 'H']
* Index(['Home Shots', 'Away Shots', 'Home SOT', 'Away SOT', 'Home Fouls',

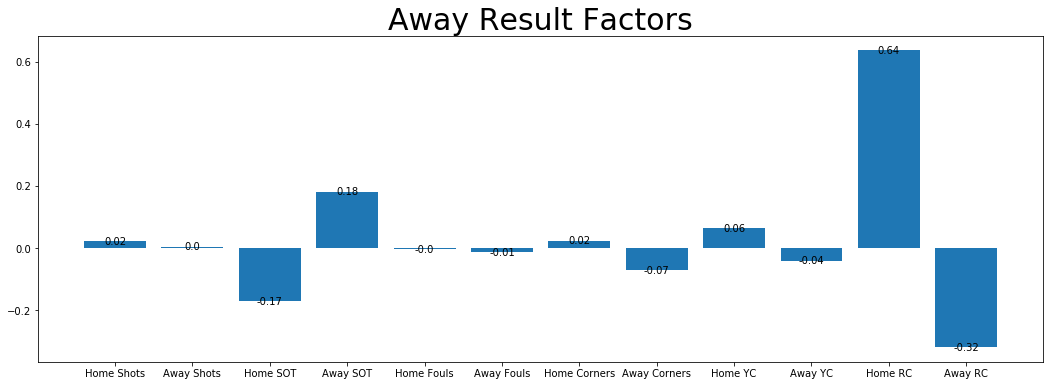
'Away Fouls', 'Home Corners', 'Away Corners', 'Home YC', 'Away YC',

'Home RC', 'Away RC'],

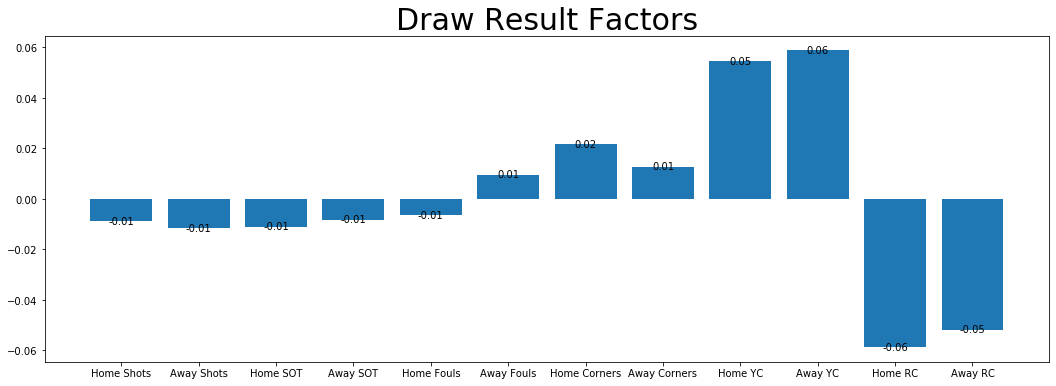
dtype='object')

* 

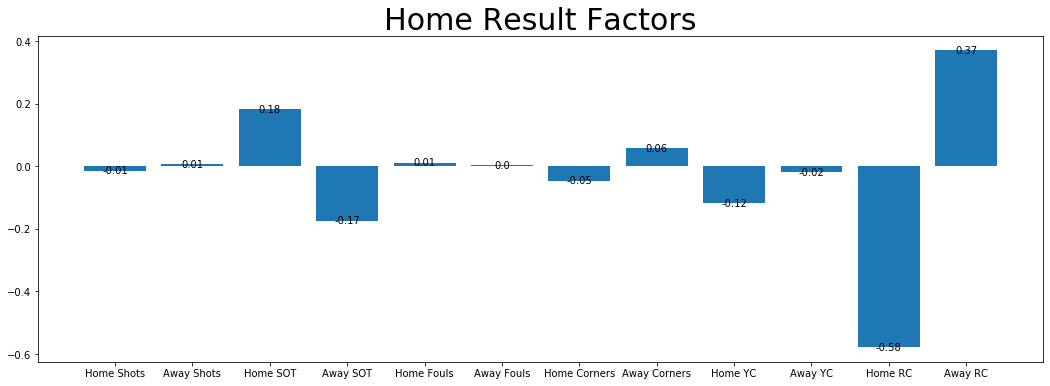
**Away win coefficients analysis:**

****

**Draw coefficients analysis:**

****

**Home win coefficients analysis:**



**Moving average analysis:**

* Decision:
  + Treat Home/Away performances differently for a given team?
    - There are Home/Away splits, but this should be taken into account by the model
  + Look at recent form regardless of Home/Away
    - Define form as points in last 5 games