European Soccer Database Project

Pradyut, Andre, Alan

Description of the data

- 25,305 Matches
- 1,458 teams
- More than 10,000 players

Our goals

- Predict which team will win in a given match
- Make $$$ when betting?
- Have an error rate that is better than random

Downloading and parsing the data

- SQLlite -----> CSV

Columns in the data

- Originally wanted to build a team-wide score based on player attributes, however, data lacks player attributes based on team.
- Team attributes included scores of the teams on a 0 - 100 scale:
  - buildUpPlaySpeed
  - buildUpPlayDribbling
  - buildUpPlayPassing
  - chanceCreationPassing
  - chanceCreationCrossing
  - chanceCreationShooting
  - defencePressure
  - defenceAggression

Formulation of the problem

- Regression or classification?

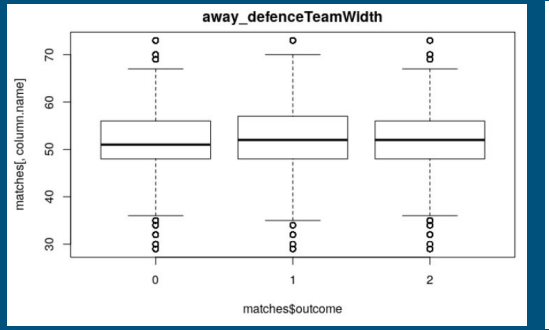
Formulate as a classification problem:

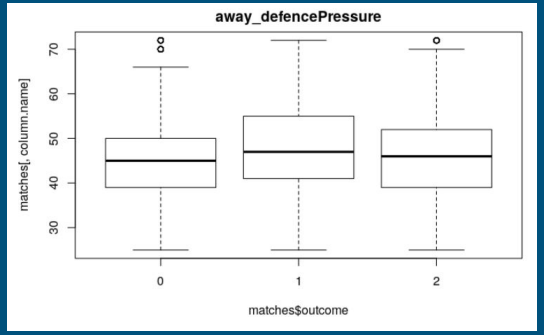
- Three possible outcomes: home team wins, away team wins, draw
- Response variable:
  - New variable, created based on the home team and away team score in the match
- Explanatory variables:
  - Attributes of the teams (which attributes?)

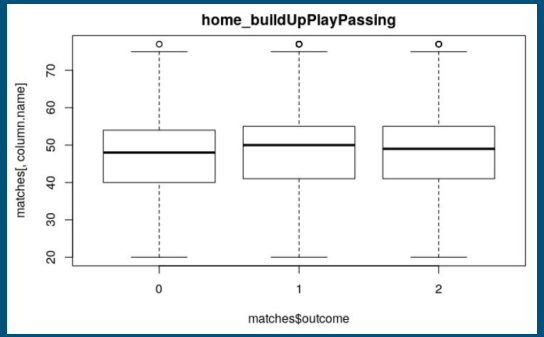
Organization of the data

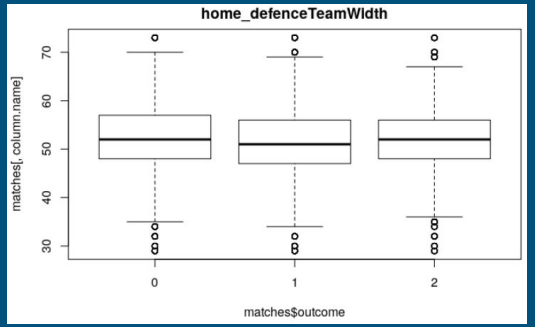
- Organized the data into rows that included the outcome (“0” for home team win, “1” for away team win, “2” for draw)
- Each row contained the match outcome and the attributes of the home and away teams
- Used R’s merge function to merge with the team id’s in the match. Renamed columns after each merge for home / away team.

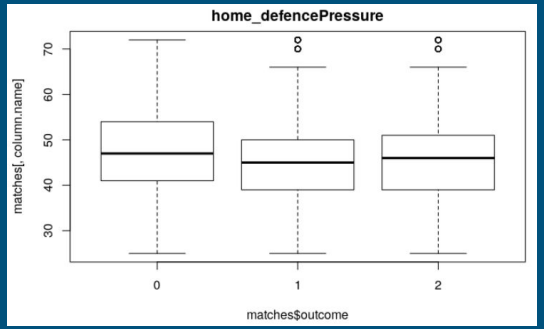
Which explanatory variables to use?

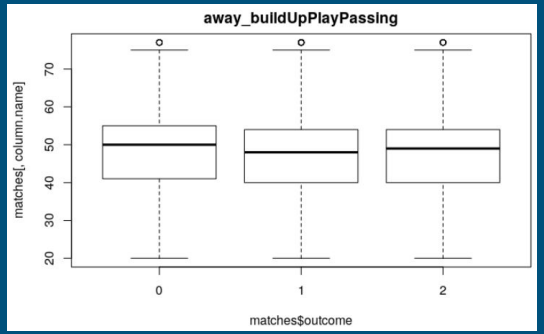












Brainstorming for explanatory variables

- Use only categorical
- Use only numerical
- Take difference of numerical variables
- Change problem so to predict home team win / no win

Models that we tried

- LDA
  - ~53% error rate
  - Predicts no draws
- QDA
  - ~52% error rate
  - Predicts draws
- K-means
  - ~60% error rate
- Logistic
  - 60% error rate

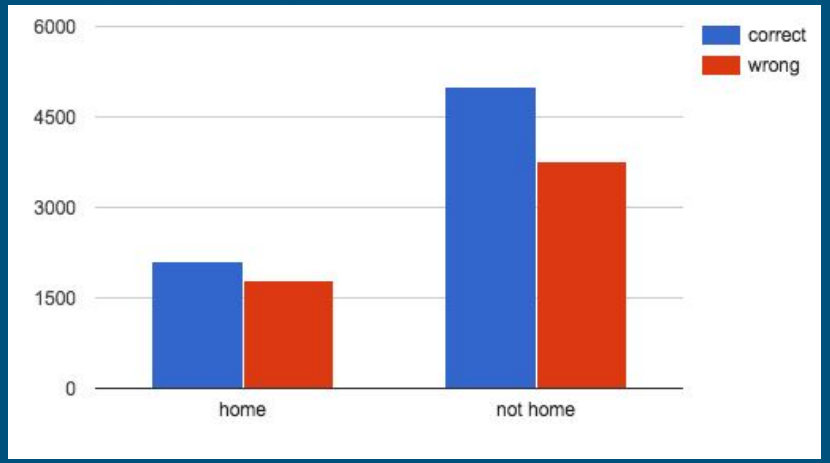
Compare to 66% error rate at random

Our best model (home team win, lose or draw)

- QDA
  - home_defenceAggression+home_chanceCreationCrossing + home_buildUpPlayPassing + home_defencePressure + away_defenceAggression + away_chanceCreationCrossing + away_buildUpPlayPassing + away_defencePressure+away_defenceAggression
  - Error rate of around 52%
  - 66% - 52% ≈ 14% advantage

Our best model (home team no win)

- LDA predicting only home team win / home team lose
  - home_defenceAggression+home_chanceCreationCrossing + home_buildUpPlayPassing + home_defencePressure + away_defenceAggression + away_chanceCreationCrossing + away_buildUpPlayPassing + away_defencePressure+away_defenceAggression
  - Error rate of around 42%!
  - 50% - 42% ≈ 8% advantage



What we learned

- Machine learning is a lot harder without textboot data
- Our final model has odds better than normal
- We could make some $$$ betting but probably not too much so better spend time in a job

Thanks for your attention!