TOP 14 RUGBY

Sakina Lord | STA478

The Data

Matches

Each row is one match of a tournament

- 2005 - 2022
  - Day # of Tournament
  - Date
  - Home & Away Teams
  - Team lineups
  - Whether or not match was delayed
  - Home and Away Scores

Players

Each row is one year of Top14 play

- 1084 different players
  - Birthday
  - Position
  - Team
- Play metrics
  - Matches played
  - Total time played
  - Penalty scores, Tries, Conversions
  - Red/yellow cards

The Data

Joined

Each row is one match/player combination with their metrics for that season and the outcome of the match

- 2005 - 2022
  - Day # of Tournament
  - Date
  - Winning team
  - Score difference
  - Player metrics

Missing stats for 102,854 players

~70% of rows have 11 fields missing

Imputation

Diagnose Missingness → Select method & Impute → Assess imputed data

- MAR
- MNAR
- MCAR

***Should** we impute?*

*What **method** should be used?*

In the *mice* package

- Random Forest
- Predictive mean matching
- Bayesian linear regression

Uses other predictors as predictors for a possible value for what is missing

- Variable distributions
- Bivariate relationships
- Correlation structure

When using mice

- Check chain convergence

Diagnose Missingness

MAR

Missingness depends on observed values, not on the missing values themselves

Controlling for observed values shows missingness is random

MCAR

Missingnes is NOT related to other predictors or the response variable

Dropping missing rows will not bias the data

MNAR

Missingness depends on the missing values themselves

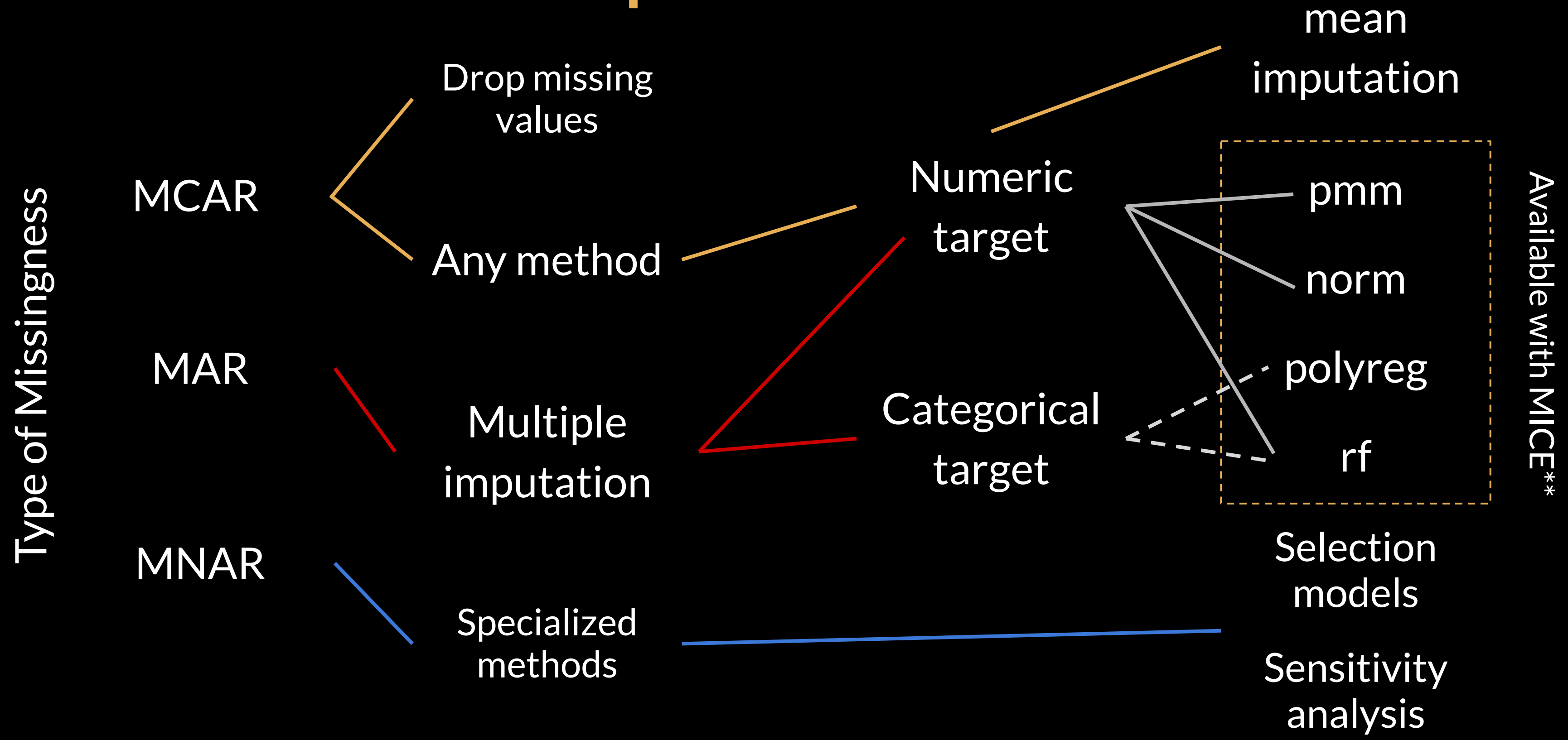
Systemic in how the data is gathered

*Other types of missingness exist, but these are the big considerations when using `mice` package

Select method & Impute

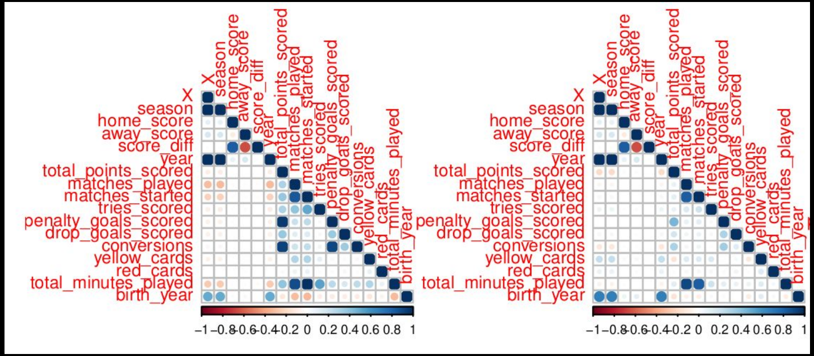
*This is an incomplete list

**MICE - Multivariate Imputation by Chained Equations

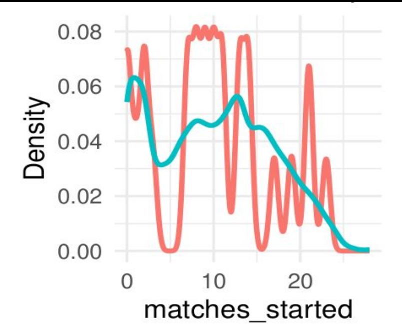


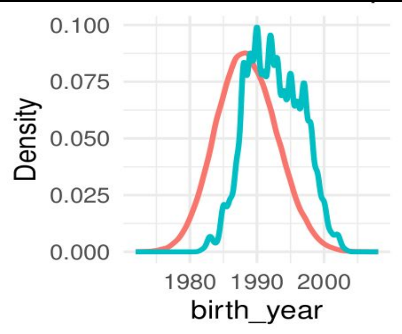
Assess imputed data

Correlation Structure

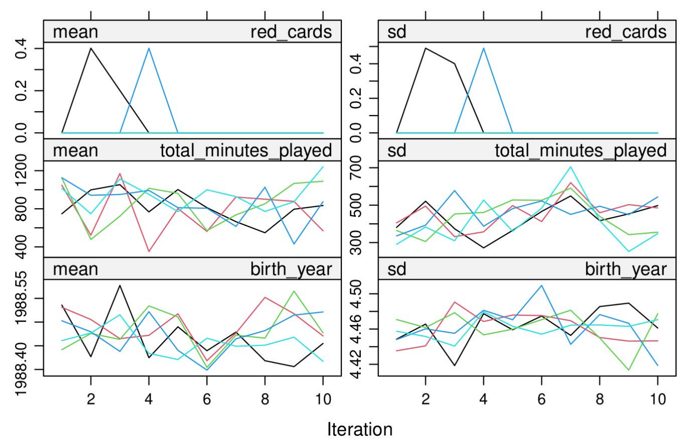


Distributions





Chain Convergence



Predictive

Random forest - Predict whether the home team wins with 73%

K-fold CV shows:

15 = best number of parameters to use

7500 = best number of trees

Day of match -> Most important predictor

nmtry	ntree	accuracy
15	7500	0.7517198
9	7000	0.7459324
15	8000	0.7453066
15	5500	0.7429644
15	6500	0.7429644

Interpretable

Elastic Net predicting a home win.

- 2nd best performing model (after RF)

	Coefficient
(Intercept)	3.3639205
pr_home_win	3.3098051
dayJ4	0.5867708
teamNarbonne	0.5321195
teamCA Brive	0.3733338
dayJ25	0.2617654
dayJ8	0.1760355
teamAuch	0.1530271
drop_goals_in_match_home	0.1133388
dayJ2	0.0535995
dayJ24	0.0252845
yellow_cards_in_match_home	0.0102652
elo_home	0.0021414
tries_in_match_home	0.0003301
teamAviron Bayonnais	0.0001680

Sources

Sas, W. (21 July, 2025). *Elo Calculator*. Omni Calculator
https://www.omnicalculator.com/sports/elo

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*
https://stefvanbuuren.name/fimd/sec-pmm.html

Buuren, S. (27 May, 2025). *Multivariate Imputation by Chained Equations.*
https://cran.r-project.org/web/packages/mice/refman/mice.html#mice

Enders, C. (2024). *Modern Missing Data Analysis.* CenterStat. [Lecture Notes]
https://centerstat.org/wp-content/uploads/2024/04/MISS-Notes.pdf

ISLR Version 2