Northwestern

CBEST.COM

Intelligent Soccer Player Classifier

Angela Chen

Problem Statement

Good Player?

Bad Player?

- Low cost-effective player influences team performances
- Soccer is a traditional sport field with little help from data science
- Generation of a dynamic model to classify soccer players

Data

- 18000 players
- 80 attributes
- 3 fields
- 2018 up to data

- 100 NA entries
- 30 irrelevant attributes
- Weight, height and currency measurements

	ID	Name	Age	Photo	Nationality	Overall	Potential	Club	Value	Preferred.Foot	...	Penalties	Composu
0	158023	L. Messi	31	https://cdn.sofifa.org/players/4/19/158023.png	Argentina	94	94	FC Barcelona	110500	Left	...	75	
1	20801	Cristiano Ronaldo	33	https://cdn.sofifa.org/players/4/19/20801.png	Portugal	94	94	Juventus	77000	Right	...	85	
2	190871	Neymar Jr	26	https://cdn.sofifa.org/players/4/19/190871.png	Brazil	92	93	Paris Saint-Germain	118500	Right	...	81	
3	193080	De Gea	27	https://cdn.sofifa.org/players/4/19/193080.png	Spain	91	93	Manchester United	72000	Right	...	40	
4	192985	K. De Bruyne	27	https://cdn.sofifa.org/players/4/19/192985.png	Belgium	91	92	Manchester City	102000	Right	...	79	

Original Dataset: https://www.kaggle.com/karangadiya/fifa19

Algorithm

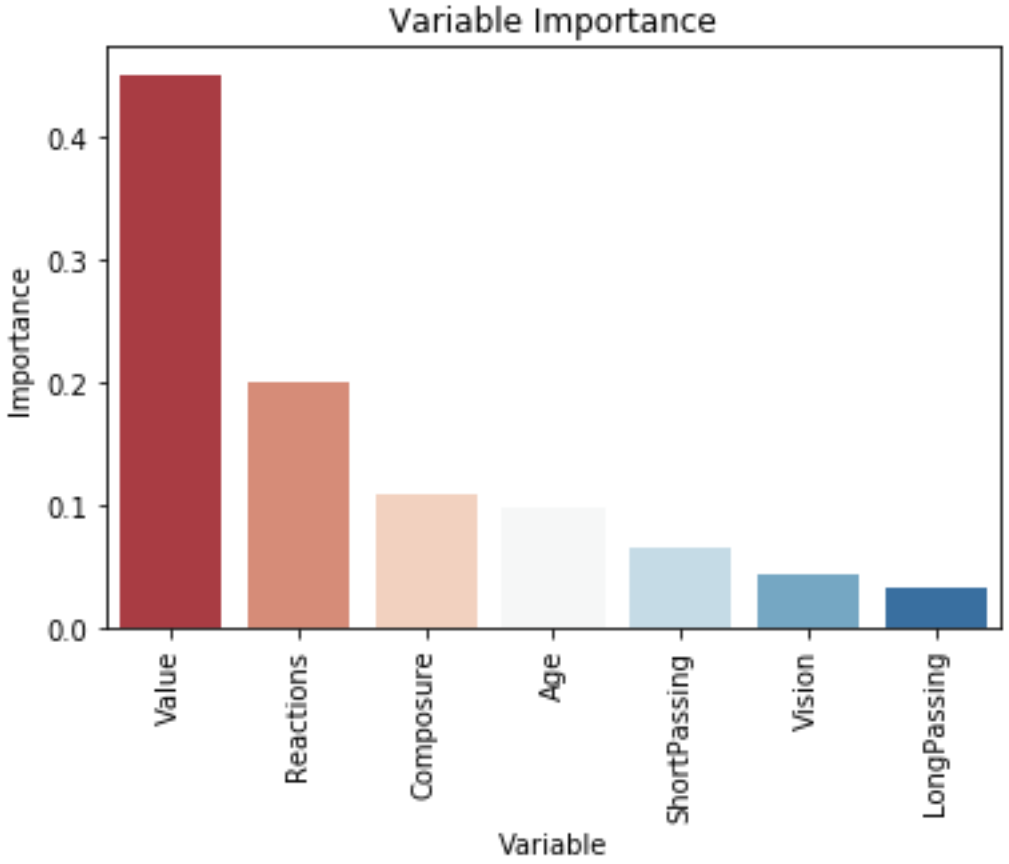
Step 1: Feature Selection

- 0 & 1 binary variable as response
- 7 features with most correlations to response
- 30% percent of test set

Step 2: Model Building

Model	Cross Validation	Misclassification Rate
Logistic Regression	No	7.8%
Random Forest	Yes	3.4%
Decision Tree	Yes	2.8%
K Nearest Neighbor	Yes	11.2%

Insight



- Value is the most essential factor
- Reaction is an important quality
- Age is not so influential as supposed

Thank You!

sichen2019@u.northwestern.edu
https://www.linkedin.com/in/angela-chen-nu/