Who Will be the Next Football Superstar?

Using FIFA game's data to predict young player's potential

MACS 30200 Project
Presenter: Siyuan Peng

Practical Application

Daryl Morey

1. The First General Manger of a NBA team with non-traditional basketball backgrounds
2. Received a bachelor's degree in computer science with an emphasis on statistics from Northwestern University
3. trend of employing more advanced statistical-based analysis rather than the traditional use of qualitative scouting and basic statistics
4. Trade of James Harden and pick great second round

Basic Information of a player

Position?

Age?

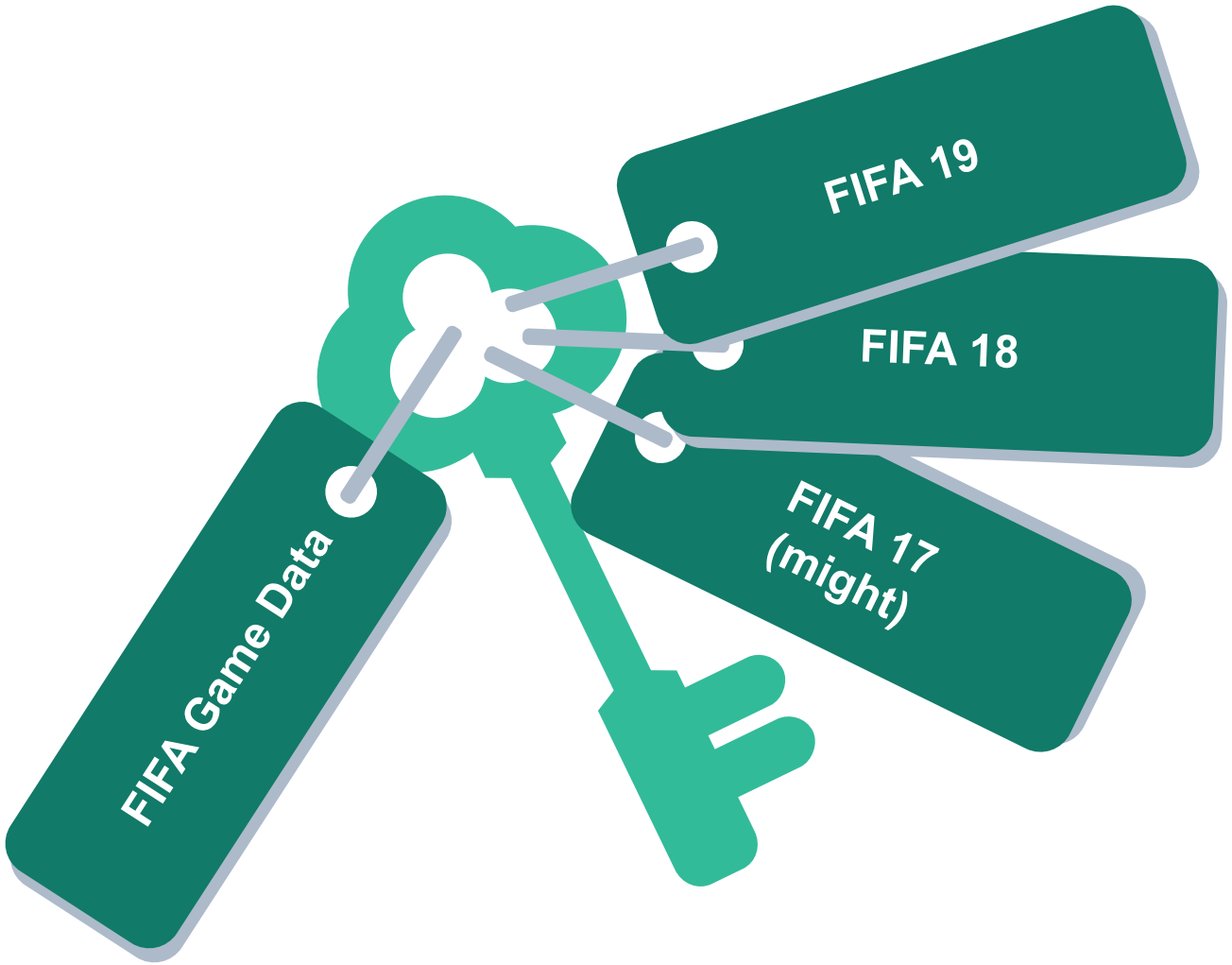
Physic?

Skills?

Seasonal Score or Ability score?

Market Value or Wage?

Data Source



Source

All databases are downloaded from Kaggle. (Data scraped from https://sofifa.com/)

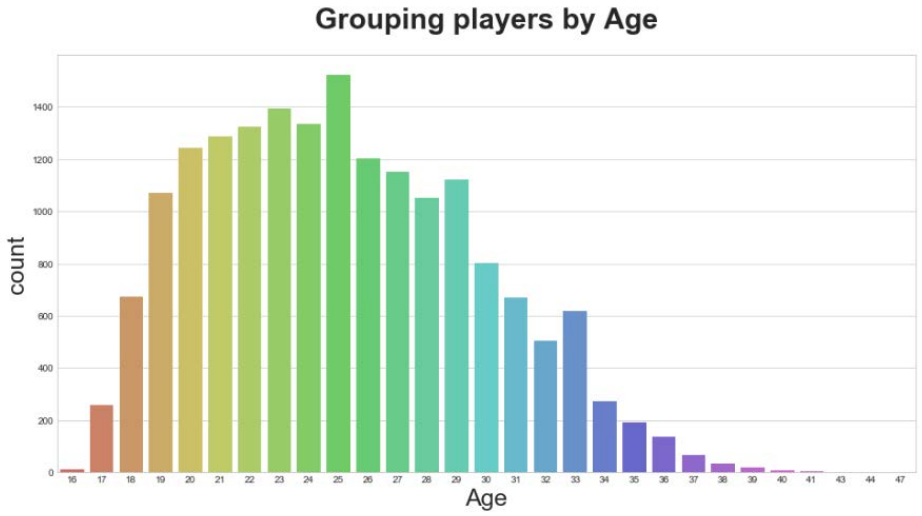
Attributes

Age, Nationality, Overall, Potential, Club, Market Value, Wage, Preferred Foot, International Reputation, Weak Foot, Skill Moves, Work Rate, Position, Height, Weight, Crossing, Finishing, Heading, Accuracy, Short Passing.

Size

More than 15,000 players and 40 numeric attributes.

Glance of the Data

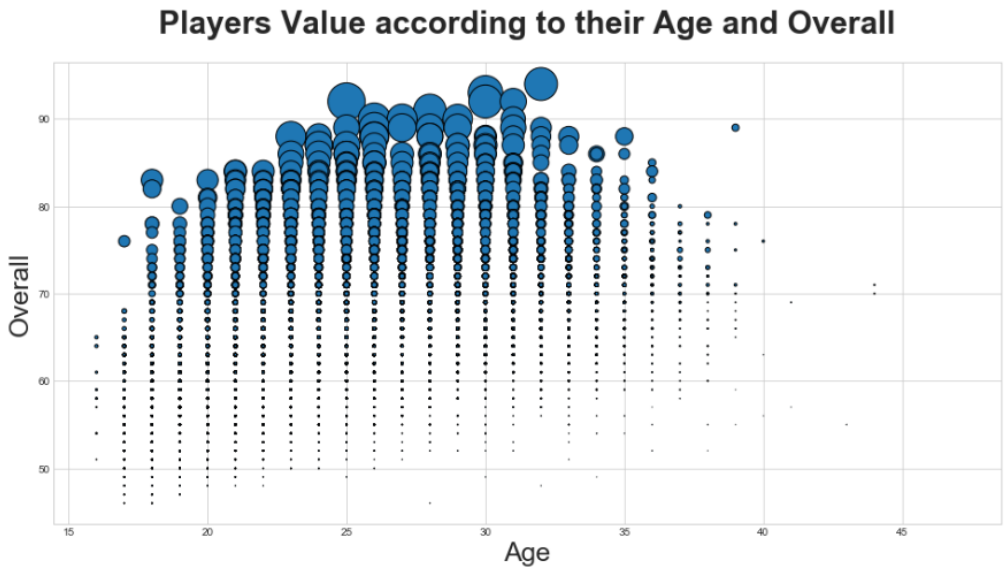


Hope to find those potential superstar

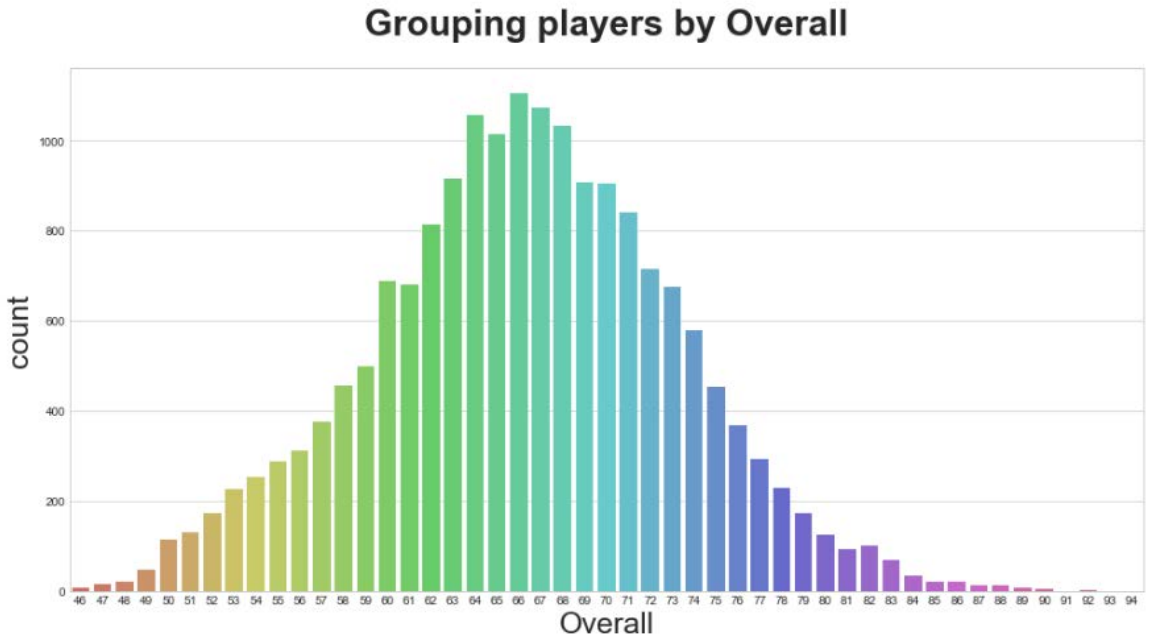
Hot 100 Players

Toggle Columns: Basic Attacking Skill Movement Power Mentality Defending Goalkeeping Special

						BASIC		SPECIAL	RESET
	NAME	AGE	O...	POT	TEAM& CONTRACT	VAL...	WA...	TOT...	HITS / COMM...
	M. Kean ST	18	73	88	Juventus 2016 ~ 2020	€8.5M	€36K	1682	9.5K / 2K
	Matheus Cunha ST RM CAM	19	72	86	RB Leipzig 2018 ~ 2023	€6.5M	€18K	1758	5.3K / 0.1K
	N. Zaniolo CAM CM RW	18	74	89	Roma 2018 ~ 2023	€11.5M	€22K	1879	3.6K / 1.1K
	K. Piątek ST	23	78	84	Milan 2019 ~ 2023	€14.5M	€76K	1808	3.4K / 4.3K
	Y. Tielemans L CM CDM	21	78	87	Leicester City Jun 30, 2019 (AS Mon...	€16M	€47K	2087	2.8K / 4.6K



The structure of the data



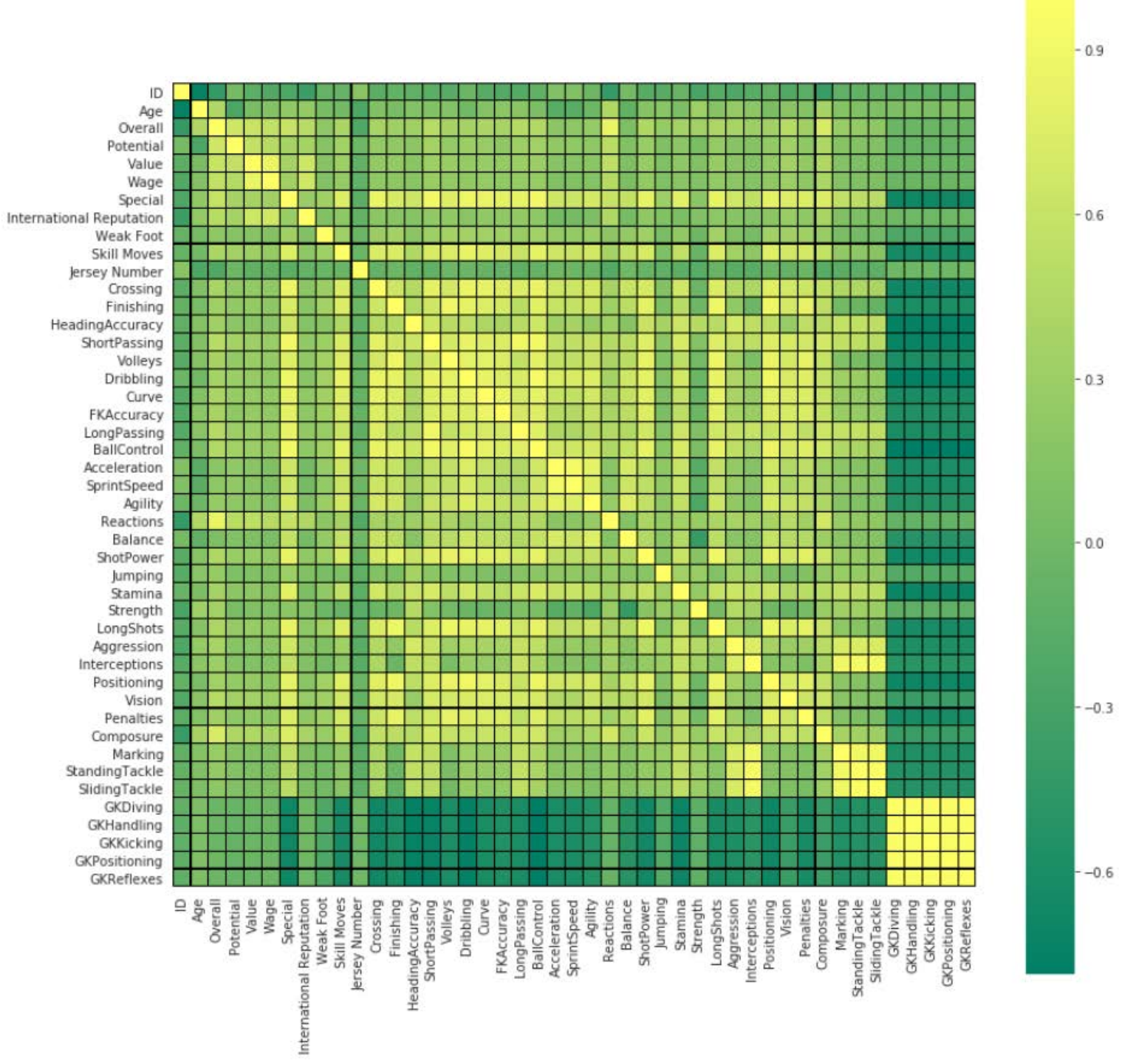
Model Selection

GMM and OLS

$$y_i = \beta_0 + \beta_{n,i} x_{n,i} + \varepsilon_i$$

Where yi is the potential of the player and there are n independent variables

Considering that there are too many variables, I need to select them cautiously.



Model Selection

Logit Regression

$$Pr(y_i = 1 | X_i, \theta) = \frac{e^{X_i \beta}}{1 + e^{X_i \beta}} = \frac{e^{\beta_0 + \beta_1 x_{1,\,i} + \ldots \beta_P x_{P,\,i}}}{1 + e^{\beta_0 + \beta_1 x_{1,\,i} + \ldots \beta_P x_{P,\,i}}}$$

How to define a successful player?

One of the great thing about the FIFA database is that EA engineers sort players by their fame.

TOP 1000 players?

Horse Race

Random Forest

Support Vector Machine

Neural Network

Data

Training Set: FIFA 18

Test Set: FIFA 19

Find players who are not currently in the TOP 1,000 but will be in

Or the TOP 100 player with highest predicted potential minus currently overall

Initial Guess

Kylian Mbappé

Physics

FIFA17->18->19

Sprint Speed:89 -> 93 -> 96

Agility: 84 -> 86 -> 92

Potential

FIFA17->18->19

87 -> 94 -> 95

Skills

FIFA17->18->19

Skill Move(5):3 -> 4 -> 5

Ball Control: 81 -> 87 -> 91

Overall

FIFA17->18->19

71 -> 83 -> 88

Contribution

Reveal the mystery of **Scout Report**

The Evaluation process of young player is somewhat trade secret.

Former works seldom focus on potential but rather on overall rating or wage.

Future star or Chinese Football team?