Data Incubator

Introduction

Dataset

Where

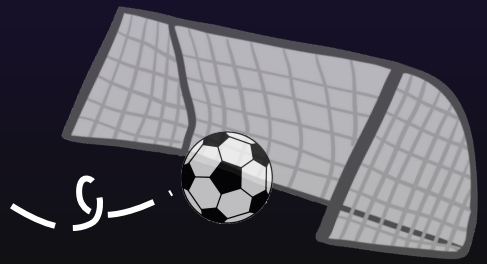
Kaggle - FIFA 19 complete player dataset

Data scraped from Sofifa website

Why

Interested in learning more about the domain

Great Potential of dataset



Main Goals

1. Market Value
2. Striker Position

Multiple Regression for Market Value

- Age vs. Market Value
- Striker = stocky and muscular?

Position Classification

- Which position should I play?
- Which skills to focus on?
- How important are these skills?

Predictive Modeling for Market Value -- Multiple Regression

Multiple Regression

Data Preparation

- Encode categorical variables to numeric values
- Remove attributes like *Nationality* to avoid overfitting
- Filter out “Super Star” players to ensure normalization of data
- Identify outliers through Norm-QQ Plot

Model Specification

- Input: 11 Characteristics of Players
  - e.g. *Age, Potential Rating, Overall Rating, Wage, Body type,* etc.
- Output: **Market Value** for Strikers (in €)
- Goal: help club managers detect underrated football strikers and make informed decisions

Multiple Regression (con't)

Best Model Result

Input Variable	Effect on Market Value	P-Value
Overall Rating	69010	Approximate 0
Age	-30966	Approximate 0
Wage	16	0.00074
Stocky Body Type	73278	0.01518
Potential Rating	5785	0.05746

Best Model Summary

- Methond: Forward selection using lowest AIC
- Performance: adjusted R-squared 0.897
- Application: Identify "hidden gems", improve ROI on players

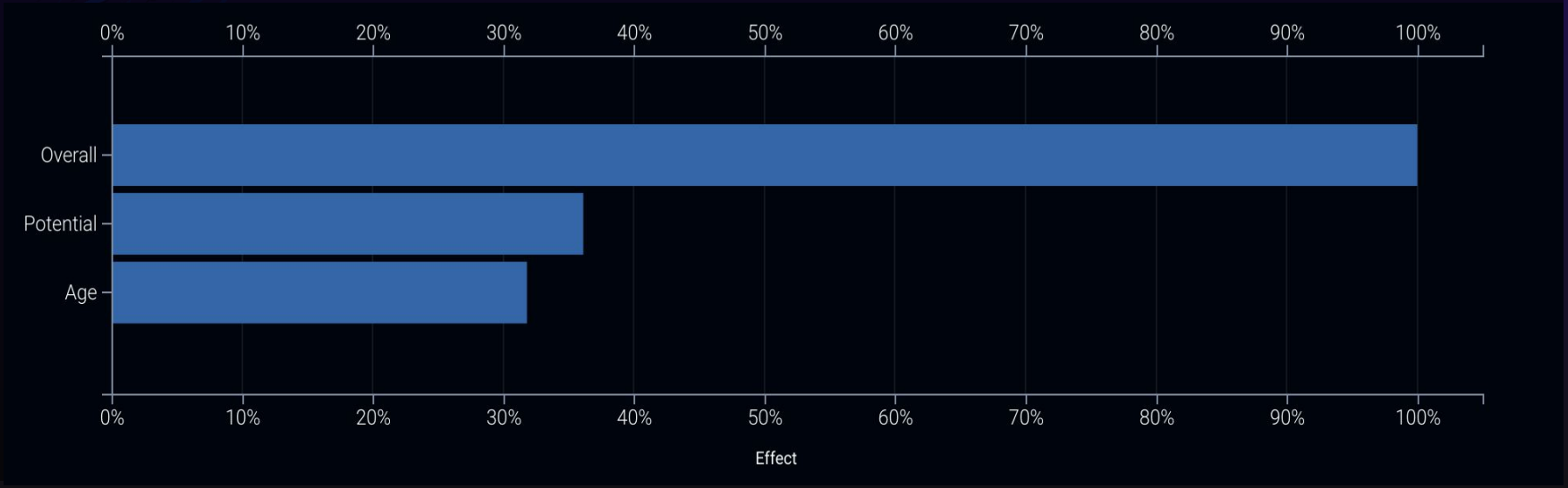
Insights

- Increase in overall rating boosts market value
- Aging is a big problem for strikers
- Stocky body type is an advantage for strikers

DataRobot

- eXtreme Gradient Boosted Tree Regressor

  RMSE: 15602 vs. 87650



Position Classification

- Striker

Logistic Regression

- Input: 30 skill ratings
- Output: Striker or not

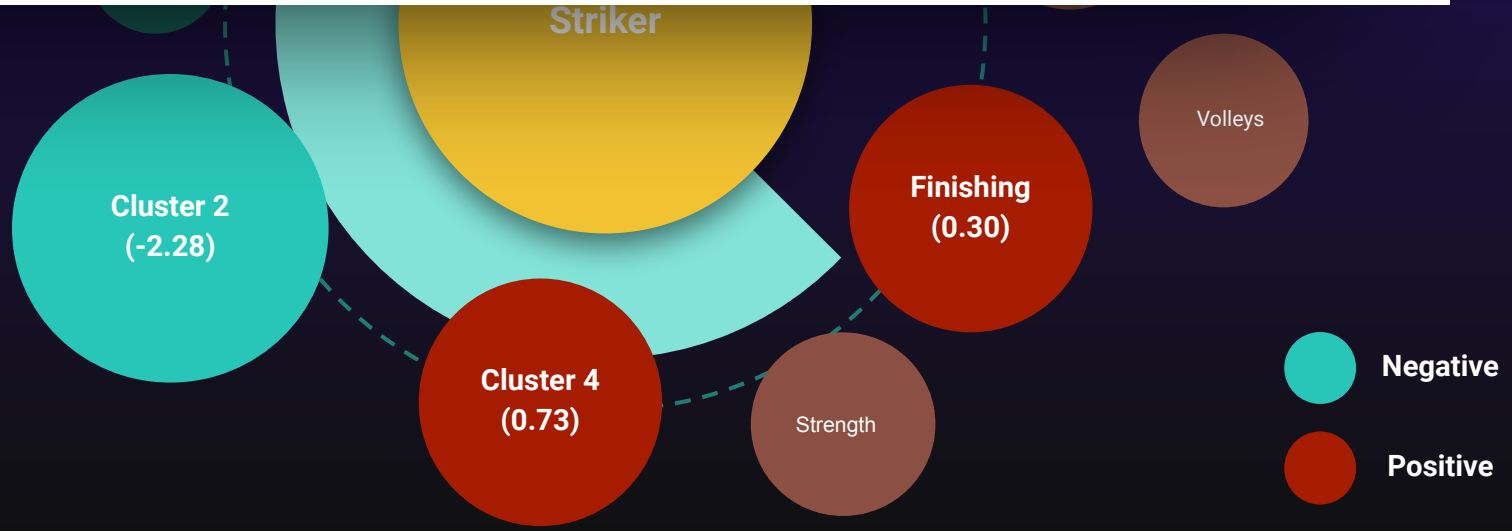
Clustering

- Dimension Reduction : 30 to 6

Attacking	Power
Mentality	Defending
Skill	Movement

https://sofifa.com/

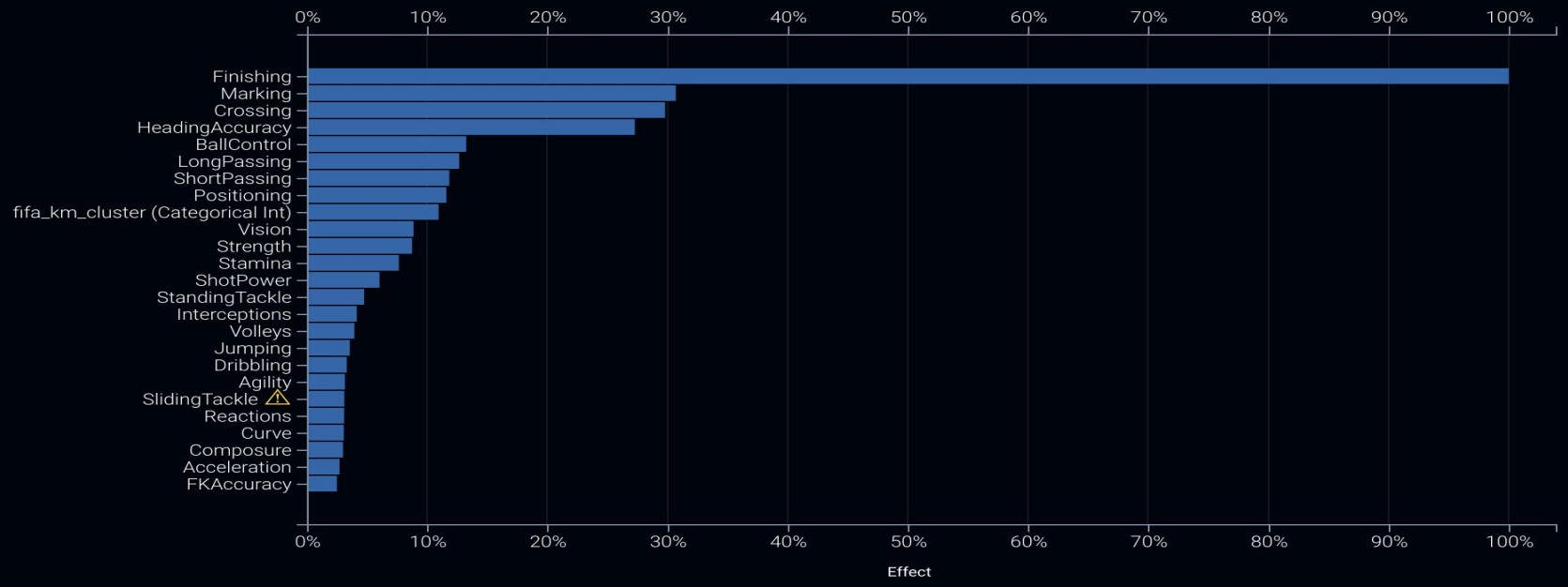
Group	Release.Clause	Attacking	Skill	Movement	Power	Mentality	Defending
1	-0.3412883	-0.1928480	-0.4019056	0.0151148	-0.7428919	-0.9965899	-1.2150539
2	-0.3145556	-1.3739665	-1.3210001	-1.0427549	-1.0270469	-0.9887795	0.5247145
3	0.4205543	0.8986570	1.0791981	0.4730293	1.0489878	1.3415202	0.7678633
4	5.4496254	1.8111452	1.7725916	1.3116219	1.4568298	1.8473548	0.2837951
5	0.0136318	0.9241627	0.5933497	0.6714985	0.5638330	0.2682620	-1.0998389
6	-0.2118450	-0.2210196	-0.0080833	-0.1161675	0.0532365	0.1827190	0.6508161



DataRobot

- Average Blender **Accuracy: 95.18% vs. 94.68%**

  (Light GBM & Nystroem Kernel SVM)



Easy to Interpret

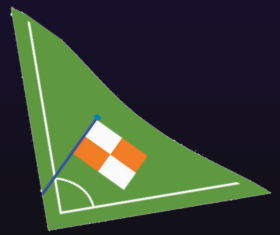
Accuracy

Tradeoff

Insights

Multiple Regression for Market Value

- Age > 25: Market Value ⤵️
- Striker ≠ Stocky and Muscular



Position Classification

- Coaches: Assign the right man to the right position
- Players: Which skills I should spend my time working on

Difficulties

outliers

unbalanced data

dimension reduction

Who will be the next star?

Let’s see!

- **The Nyström Approximation:**

$$\mathbf{K} \approx \tilde{\mathbf{K}}_c^{\mathrm{nys}} = \mathbf{C}\mathbf{W}^{\dagger}\mathbf{C}^T$$

(A low-rank factorization).