Data Driven Estimation of Market Values of Soccer Players

Anubhav Pareek (Ridge)

Presented January 22, 2021

What Am I Trying To Do ?

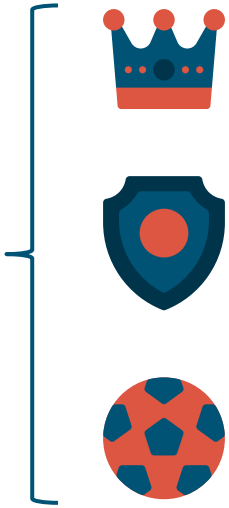
"People are overlooked for a variety of biased reasons and perceived flaws. Age, appearance, personality."

JONAH HILL - Peter Brand

So What Am I Really Trying to Do Then ?

Current Market Valuation

?



Market Information

+

Player Attributes

+

Player Statistics/Ability

Features & Target

Market Information

- 2016 Market Valuations
- Wikipedia Page Views(2016-2017)

Player Attributes

- Height
- Weight
- Age
- Foot(Right/Left)
- Nation

Player Ability

- Position
- Matches Played
- Match Starts
- Goals
- Assists
- Penalty Kicks
- Cards(Y/R)

Target - 2017 Market Valuations

(Source of Truth : TransferMarkt Val.)

Exploratory Data Analysis | Overview

141 Players across 20 Teams

Mean Age : 25.95
Mean Height(cm) : 181.88
Mean Weight(Kgs) : 75.68

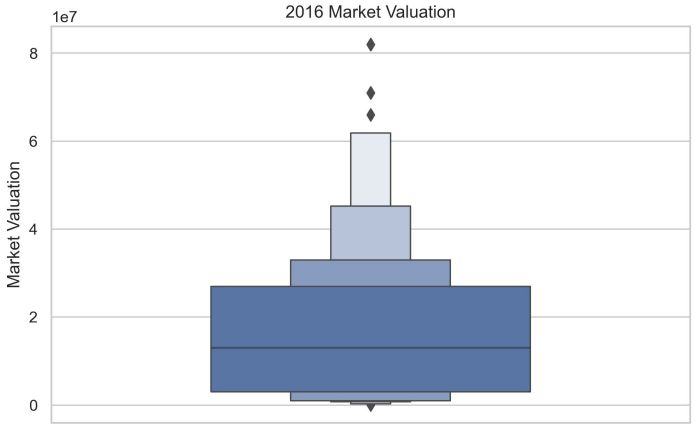
Wikipedia Page Views

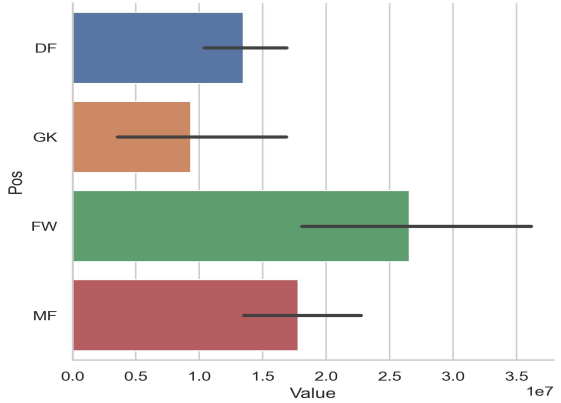
Dates - 08/01/2016 - 08/01/2017

2016 Market Valuations

Highest Mean (Team) - 32.5 Mn USD

Exploratory Data Analysis | Market Valuations

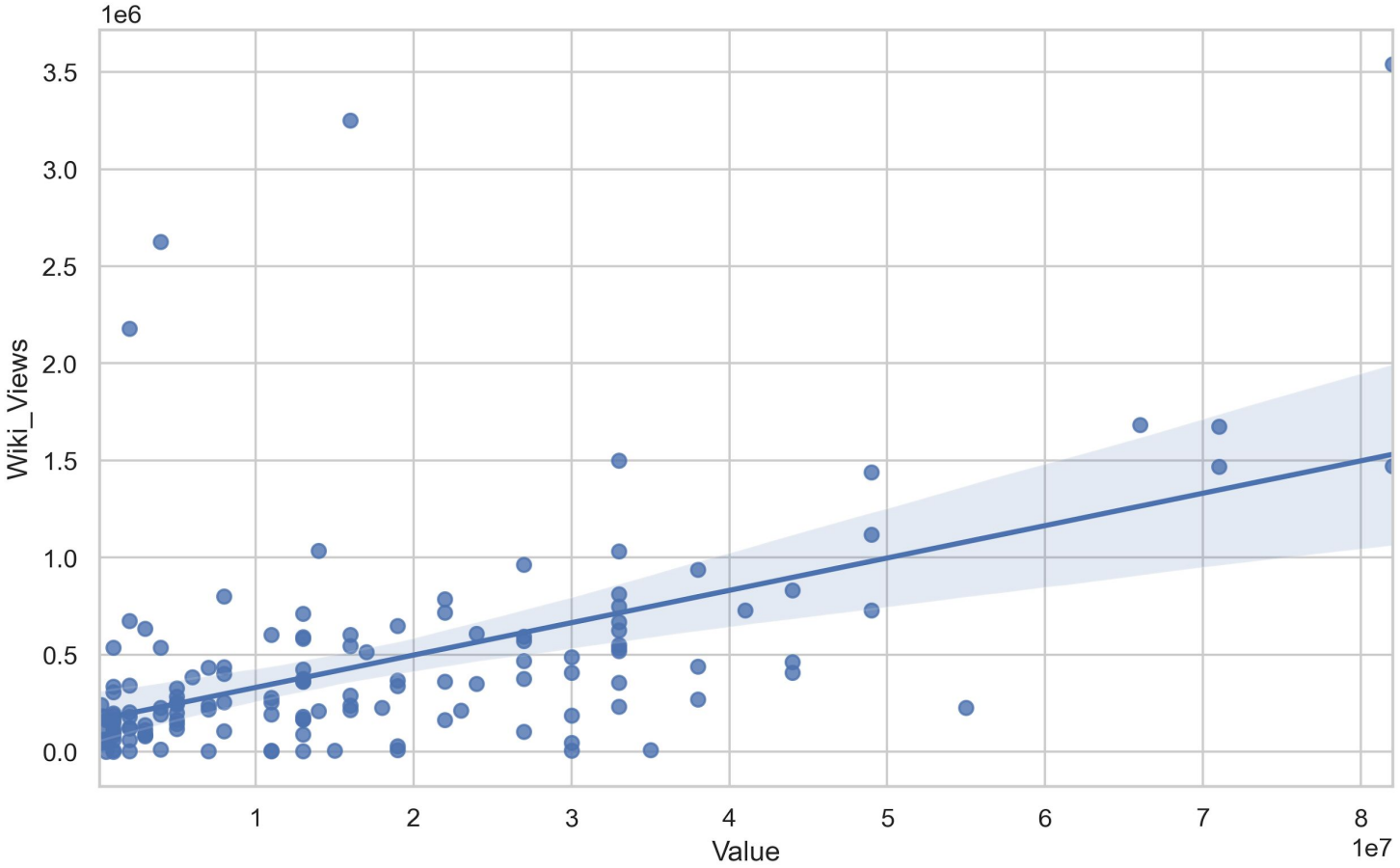




Key Observations:

- **Mean Valuation ~** 18 Mn USD
- **Forward Position Players** are valued the highest at ~ 27 Mn USD
- **DF -** Defender, **GK -** Goal-Keeper, **FW -** Forward, **MF -** Mid Fielder

Exploratory Data Analysis | Popularity Impact



Model | Results

A

#	Coefficient	Coefficient Estimate
1	Red Cards	-3.75E+06
2	Age	-2.15E+06
3	Yellow Cards	-1.19E+06
4	Matches Played	-7.22E+05
5	Height	-2.65E+05
6	Value - 2016	8.73E-01
7	Wiki_Views	6.73E+00
8	Minutes on Field	3.88E+03
9	Weight	1.86E+05
10	Starts	7.53E+05
11	Assist	9.45E+05
12	Goals	1.01E+06

B

OLS Regression Results

Dep. Variable:	Target	R-squared (uncentered):	0.832
Model:	OLS	Adj. R-squared (uncentered):	0.816
Method:	Least Squares	F-statistic:	53.05
Date:	Fri, 22 Jan 2021	Prob (F-statistic):	6.13e-44
Time:	02:14:12	Log-Likelihood:	-2535.8
No. Observations:	141	AIC:	5096.
Df Residuals:	129	BIC:	5131.
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
MP	-4.965e+05	3.01e+05	-1.647	0.102	-1.09e+06	1e+05
Age	-1.79e+06	3.99e+05	-4.492	0.000	-2.58e+06	-1e+06
Starts	6.892e+05	5.47e+05	1.260	0.210	-3.93e+05	1.77e+06
Min	967.6512	5978.511	0.162	0.872	-1.09e+04	1.28e+04
Gls	1.189e+06	3.71e+05	3.208	0.002	4.56e+05	1.92e+06
Ast	6.879e+05	7.17e+05	0.960	0.339	-7.3e+05	2.11e+06
CrdY	-9.104e+05	5.69e+05	-1.600	0.112	-2.04e+06	2.16e+05
CrdR	-2.928e+06	2.98e+06	-0.983	0.328	-8.82e+06	2.97e+06
Height	3.783e+05	1.32e+05	2.859	0.005	1.16e+05	6.4e+05
Weight	-2.968e+05	2.97e+05	-0.999	0.320	-8.85e+05	2.91e+05
Value	0.9152	0.104	8.833	0.000	0.710	1.120
Wiki_Views	5.8921	3.269	1.802	0.074	-0.577	12.361

Key Observations:

- **Model A** = Train : 60%, Validation : 20% and Test : 20% split
- Height, Weight, Matches Played display unexpected values

Model A -Various Models Tested

01 Linear Regression

Test data	R^2: 0.815
Validation data	R^2: 0.695

02 Ridge Regression

Test data	R^2: 0.824
Validation data	R^2: 0.702

03 LASSO Regression

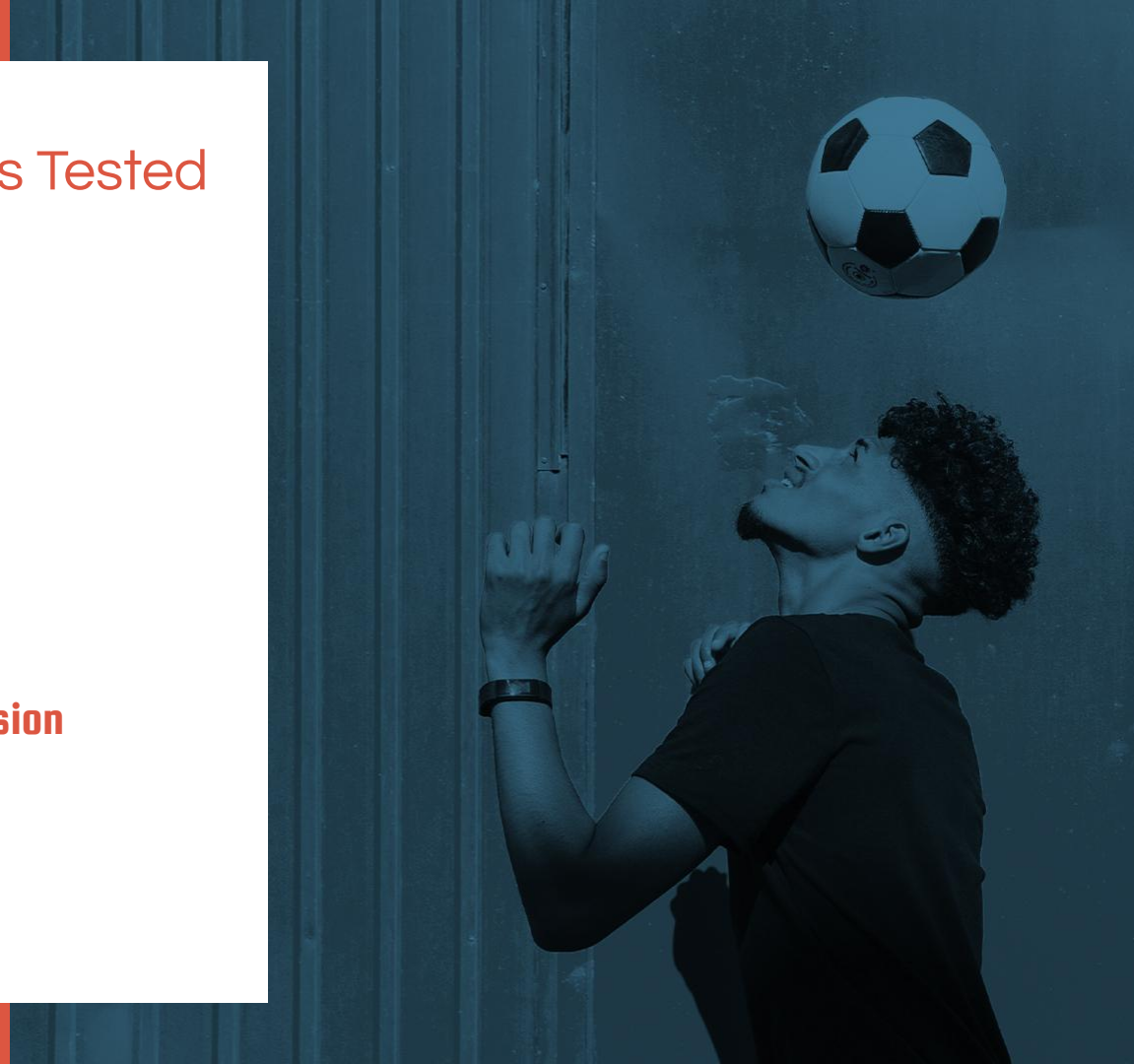
Test data	R^2: 0.81466
Validation data	R^2: 0.694

04 Degree -2 Polynomial Regression

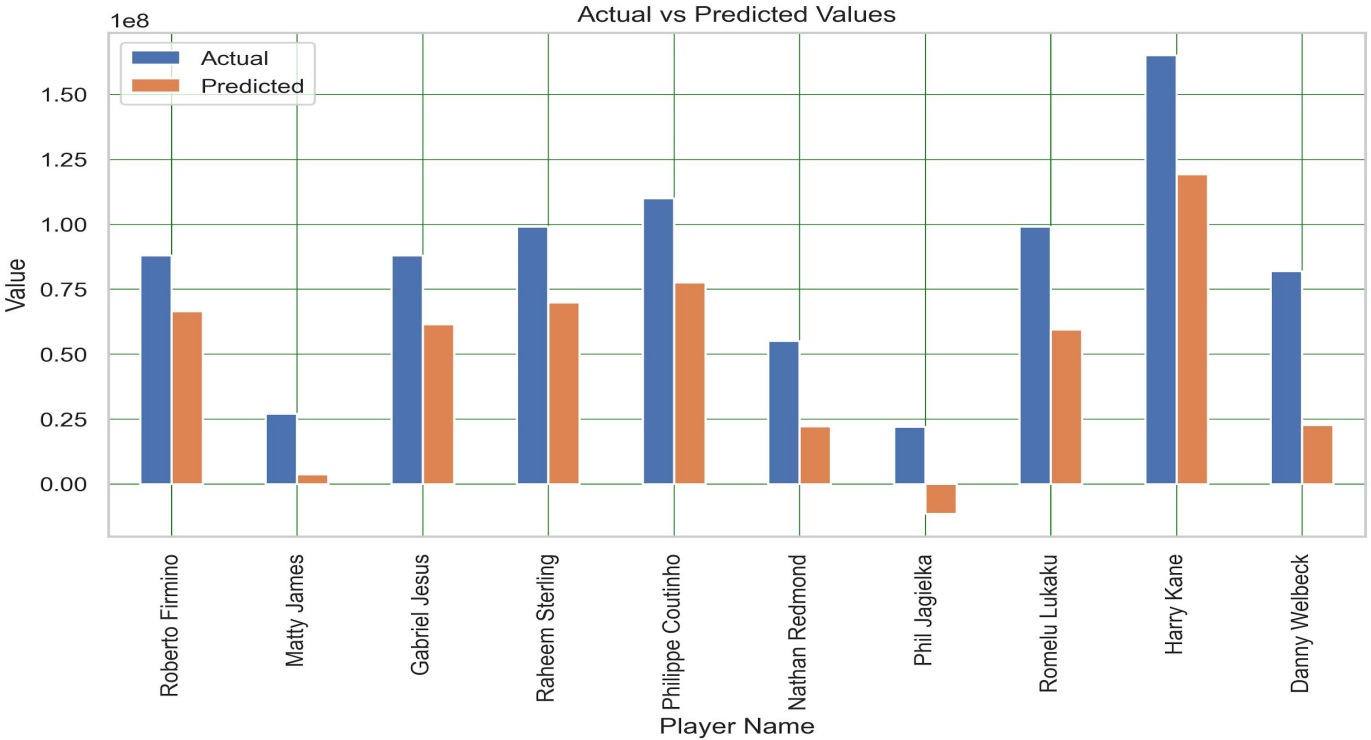
Test data	R^2: -12.47
Validation data	R^2: -11.713

05 Cross-Validation(10 Splits)

R^2 : 0.504

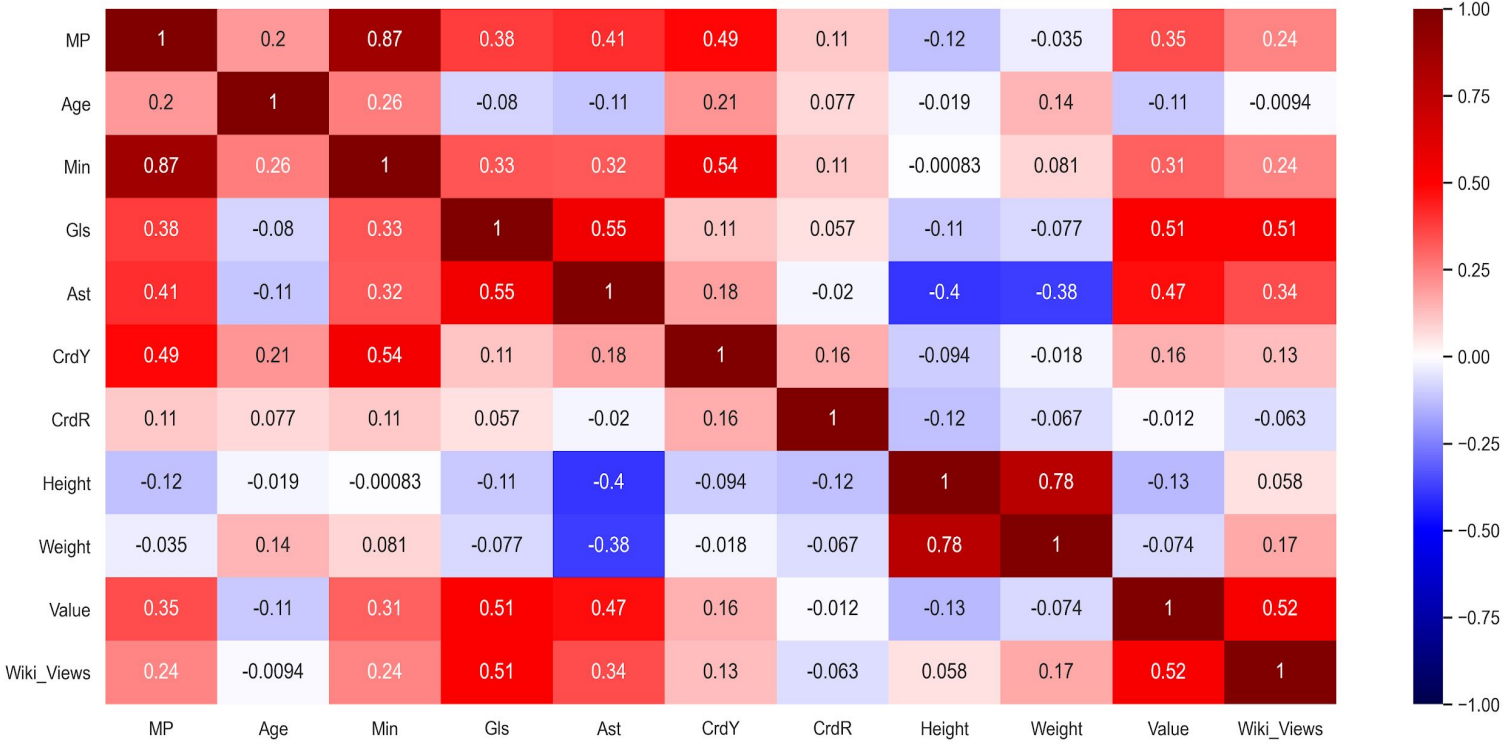


Model| Prediction



THANK YOU

Model | Correlation



Model | Correlation & Pair Plot