



Predicting Footballers Market Values

Using 2008 - 2017 data from transfermarkt.co.uk

Christopher Williams
Capstone
GA DSI
Austin
Feb 2019

Contents

- The Data Science Problem
- The Data
 - Getting it
 - Data Challenges
 - Seeing it
 - Data Assumptions
 - Facts
- Modeling
 - Regression
 - Classification
- Model Results
- Conclusions

Data Science Problem

In this project the problem I'm seeking to understand is how well annual statistics predict the estimated market values of professional footballers in the top 5 European leagues.

Using players annual data to see how applicable machine learning prediction models are on market values.

Data: comes from transfermarket.co.uk which has crowd sourced market values for players in the top flight of European football over the last 10 years. Think Penelope and wisdom of the crowds!



The Data

Getting it

Data Challenges

Most sites focus on the just high level stats like: shots, goals, assists, cards, minutes-played, and games played.

Lack of one site with all the stats desired for true merits of each different position.

Complex Process to unify all the stats needed to build a truly complete dataset at scale that could be used to predict market values of players in defensive positions along with forwards & midfielders.

Compact	Detailed	Gallery								
#	Player	Date of Birth (Age)	Nat.	Current club	Height	Foot	Joined	Signed from	Contract until	Market value
1	Marc-André ter Stegen Goalkeeper	Apr 30, 1992 (25)			1,87 m	right	Jul 1, 2014		30.06.2022	£22.50m
13	Jasper Cillessen Goalkeeper	Apr 22, 1989 (28)			1,85 m	right	Aug 25, 2016		30.06.2021	£8.10m
-	Adrián Ortíz Goalkeeper	Aug 20, 1993 (23)			1,87 m	left	-	-	-	£720k
3	Gerard Piqué Centre-Back	Feb 2, 1987 (30)			1,94 m	right	Jul 1, 2008		30.06.2022	£36.00m
23	Samuel Umtiti Centre-Back	Nov 14, 1993 (23)			1,82 m	left	Jul 12, 2016		30.06.2023	£27.00m
24	Yerry Mina Centre-Back	Sep 23, 1994 (22)			1,95 m	right	Jan 11, 2018		30.06.2023	£4.50m
25	Thomas Vermaelen Centre-Back	Nov 14, 1985 (31)			1,83 m	left	Aug 9, 2014		30.06.2019	£2.70m
-	David Costas Centre-Back	Mar 26, 1995 (22)			1,84 m	right	-	-	-	£900k
18	Jordi Alba Left-Back	Mar 21, 1989 (28)			1,70 m	left	Jul 1, 2012		30.06.2020	£28.80m
19	Lucas Digne Left-Back	Jul 20, 1993 (23)			1,78 m	left	Jul 13, 2016		30.06.2021	£14.40m
-	Marc Cucurella Left-Back	Jul 22, 1998 (18)			1,75 m	left	-	-	-	£630k
20	Sergi Roberto Right-Back	Feb 7, 1992 (25)			1,78 m	right	Jul 1, 2013		30.06.2022	£22.50m
2	Nélson Semedo Right-Back	Nov 16, 1993 (23)			1,77 m	right	Jul 14, 2017		30.06.2022	£18.00m
5	Sergio Busquets Defensive Midfield	Jul 16, 1988 (28)			1,89 m	right	Sep 1, 2008		30.06.2023	£54.00m

Data Collection

Market Values

2008 - 2017

DOB

Nationality

Height

Foot

Joined

Signed From

Transfer Fee

Contract Until

Market Value

FC Barcelona have played 59 games so far and achieved a points average of 2,31 points per game.

		Compact	Detailed	#	Player	Age	Nat.	In squad	H	S	G	Y	R	P	PPM	Time		
				1	Marc-André ter Stegen Goalkeeper	25	GER	57	48	-	-	-	-	-	2,25	4.320'		
				13	Jasper Cillessen Goalkeeper	28	ESP	59	11	-	-	-	-	-	2,55	990'		
				*	Adrián Ortolá Goalkeeper	23	ESP	2	Was not used during this season									
				18	Jordi Alba Left-Back	28	ESP	54	48	3	11	10	-	-	4	4	2,33	4.000'
				23	Samuel Umtiti Centre-Back	23	FRA	43	40	1	-	7	-	-	1	1	2,28	3.539'
				20	Sergi Roberto Right-Back	25	ESP	50	48	1	8	6	-	2	11	9	2,29	3.439'
				3	Gerard Piqué Centre-Back	30	ESP	56	49	4	-	12	1	-	1	7	2,37	4.145'
				2	Nélson Semedo Right-Back	23	POR	47	36	-	2	4	-	-	8	7	2,31	2.494'
				24	Yerry Mina Centre-Back	22	COL	13	6	-	1	2	-	-	2	-	1,83	377'
				19	Lucas Digne Left-Back	23	FRA	47	20	1	2	3	-	-	5	3	2,10	1.337'
				*	Marc Cucurella Left-Back	18	ESP	1	1	-	-	-	-	-	1	-	3,00	7'
				*	David Costas Centre-Back	22	ESP	3	1	-	-	-	-	-	1	-	3,00	32'
				25	Thomas Vermaelen Centre-Back	31	BEL	36	20	-	1	4	-	-	3	2	2,20	1.512'
				14	Philippe Coutinho Attacking Midfield	25	BRA	22	22	10	6	1	-	-	5	11	2,32	1.483'
				5	Sergio Busquets Defensive Midfield	28	ESP	52	50	1	5	11	-	-	2	13	2,26	4.164'

Data Collection

Player's Yearly Stats

2008 - 2017

In-squad Appearances

Goals

Assists

Yellow Cards

Second Yellows

Red Cards

Substituted On

Substituted Off

PPM

Time Played

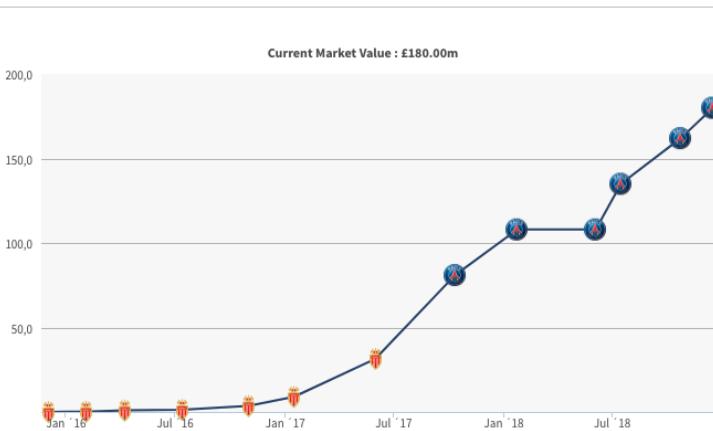
#7 Kylian Mbappé



Date of Birth (Age): **Dec 20, 1998 (20)** Height: **1,78 m** Current international: **France**
Place of Birth: **Bondy** Position: **Right Winger** Caps/Goals: **28/10**
Citizenship: **France** Agent: **Relatives**

PROFILE STATS ▾ MARKET VALUE TRANSFERS RUMOURS NATIONAL TEAM NEWS HONOURS

MARKET VALUE DEVELOPMENT



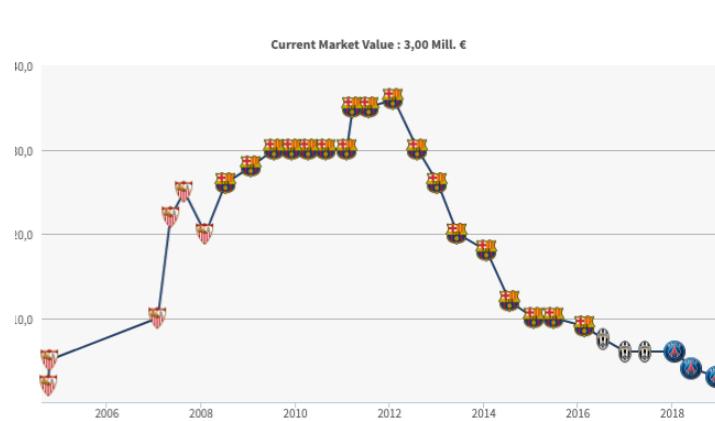
#13 Dani Alves



Date of Birth (Age): **May 6, 1983 (35)** Height: **1,72 m** International: **Brazil**
Place of Birth: **Juazeiro** Position: **Right-Back** Caps/Goals: **107/7**
Citizenship: **Brazil** Agent: **Flashforward, S.L.**

PROFILE STATS ▾ MARKET VALUE TRANSFERS RUMOURS NATIONAL TEAM NEWS HONOURS

MARKET VALUE DEVELOPMENT



Kylian Mbappé vs Dani Alves



Sergio Aguero, hat trick
Man City 6, Chelsea 0. Feb
10
£ 67.5m market value 2018

#10 Sergio Agüero



Date of Birth (Age): Jun 2, 1988 (30)

Height: 1,73 m

International: Argentina

Place of Birth: Buenos Aires

Position: Centre-Forward

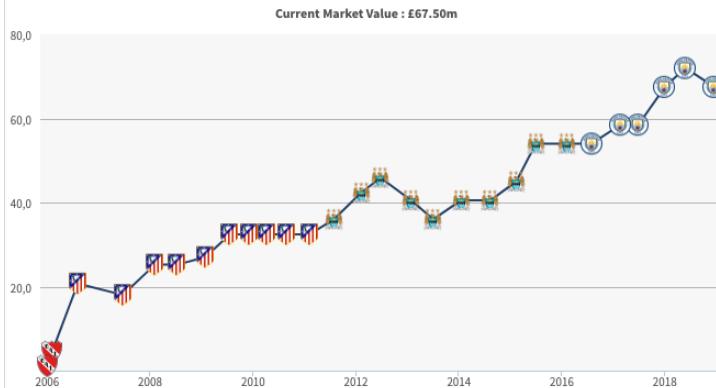
Caps/Goals: 89/39

Citizenship: Argentina

Agent: Eleven GT

PROFILE STATS ▾ MARKET VALUE TRANSFERS RUMOURS NATIONAL TEAM NEWS HONOURS

MARKET VALUE DEVELOPMENT



#10 Lionel Messi



Date of Birth (Age): Jun 24, 1987 (31)

Height: 1,70 m

International: Argentina

Place of Birth: Rosario

Position: Right Winger

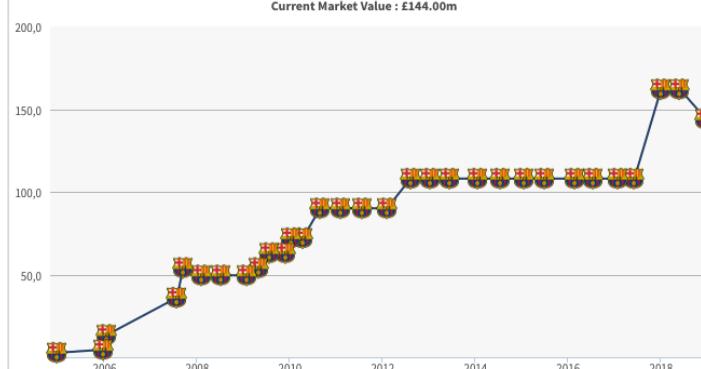
Caps/Goals: 128/65

Citizenship: Argentina

Agent: Relatives

PROFILE STATS ▾ MARKET VALUE TRANSFERS RUMOURS NATIONAL TEAM NEWS HONOURS

MARKET VALUE DEVELOPMENT



Sergio Agüero vs Lionel Messi

The Data

Seeing it

Data Assumptions. How to deal with it.

How to deal with defenders & goalies compared to midfielders & forwards who score more goals, have more assist?

Min number of appearances?
Min number of time played?

Facts about the data

Top 5 European Leagues

- Spanish La Liga
- French Ligue 1
- English Premier League
- Italian Serie A
- German Bundesliga (only 18 teams)

30504 observations over 2008-2017

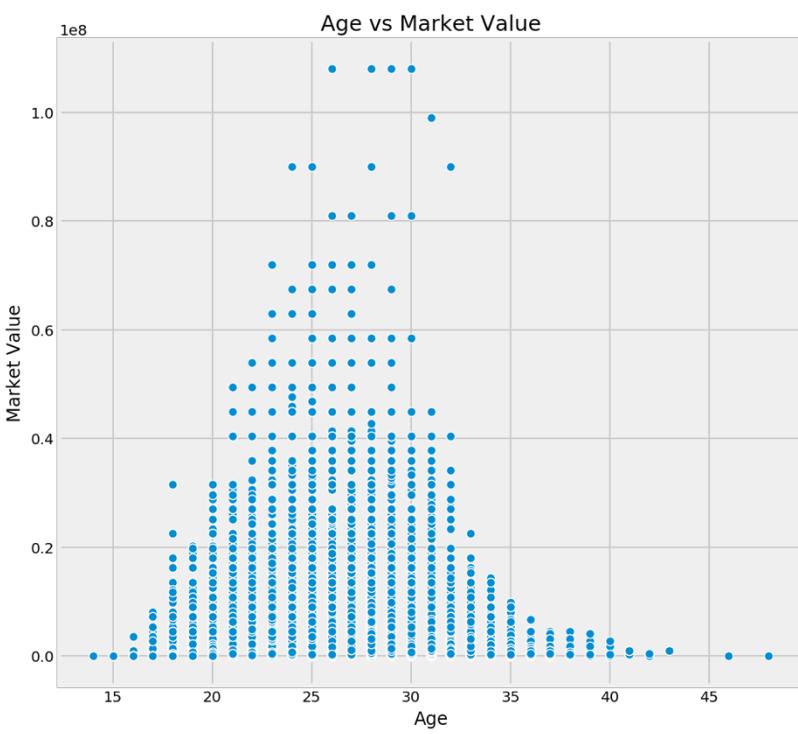
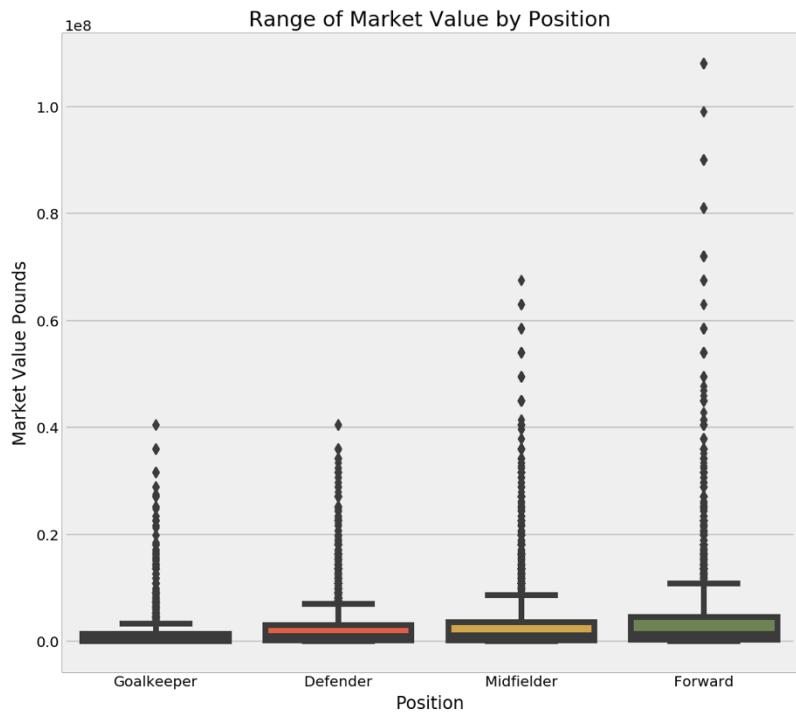
- Dropped to 24,632
- Average Market Value of 3,257,152
- Median Market Value of 900,000
- 10584 unique players
- 1.85 average goals per season per player
- 17.3 average appearances
- 24 years old on average

Market Value Distribution of All Players and Amounts



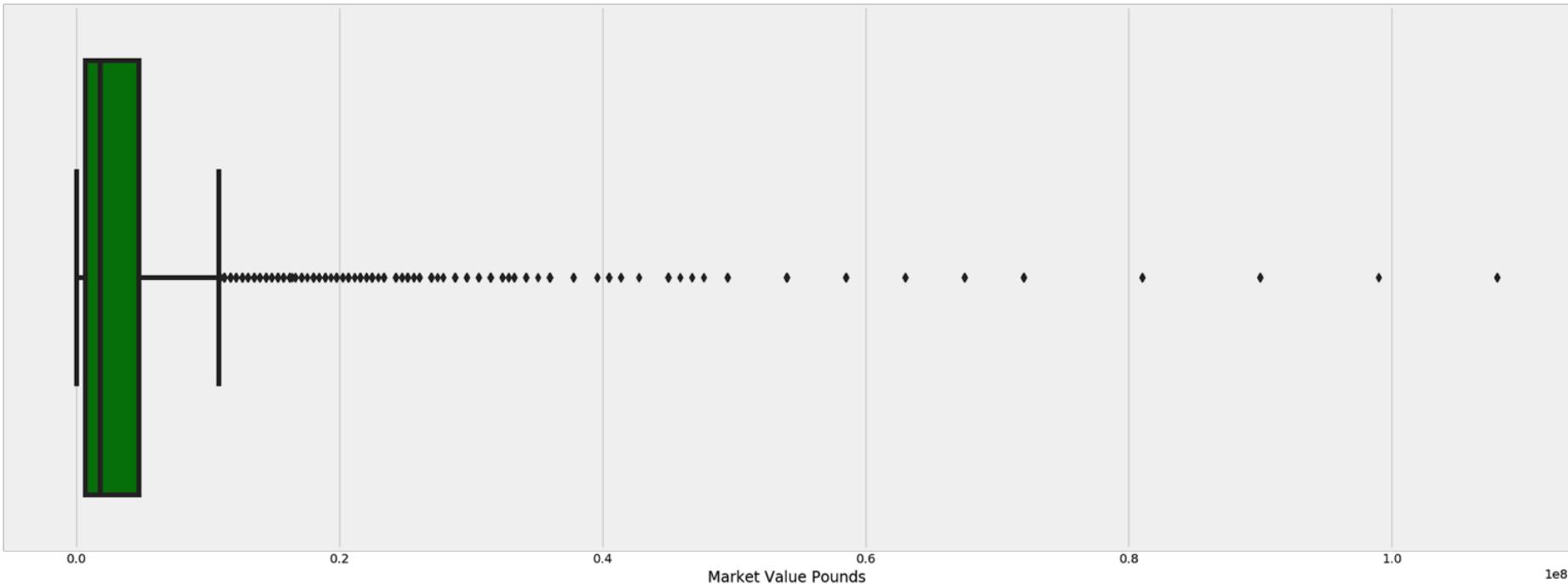
Market Value Distribution Less than 15m



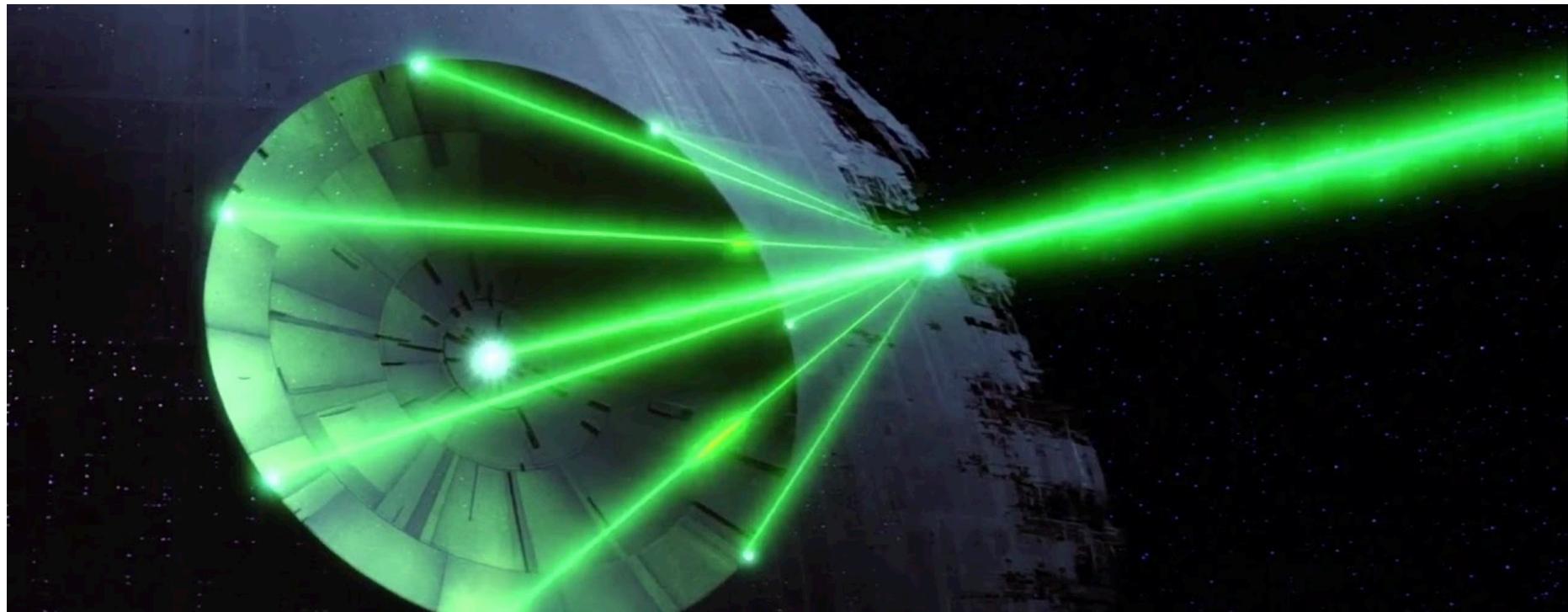


Position and Age vs Market Value

Distribution of Market Value



Distribution of Market Values



Distribution of Market Values

Correlation with market value



Modeling

Doing it

THE
BEST

FOOTBALL AWARDS 2018

הYEAR

39

FIFA FOOT

הYEAR
39

FIFA FOOTBALL AWARDS

Dani Alves
Paris Saint-Germain
£ 6m market value 2018
£ 9m market value 2016
£ 20m market value 2009



Regression Models

- LinearRegression
- Ridge
- RidgeCV
- Lasso
- LassoCV
- ElasticNet
- ElasticNetCV
- DecisionTreeRegressor
- BaggingRegressor
- KNeighborsRegressor
- RandomForestRegressor
- AdaBoostRegressor
- SVR



Targets

Main Target is Market Value

- Primary objective for this project
- 21,859 observations

Transfer Fee

- Not primary objective for this project
- Limited Observations 4030

Classification of players by general position

- Forward
- Midfielder
- Defender
- Goalkeeper

Features

Categorical Features

- 98 clubs
- Position
- Foot
 - Right
 - Left
 - Both

Hard Features:

- Age
- Market -Value (not included when target)
- Transfer-Fee (not included when target)
- In-Squad (included on bench)
- Appearances (games played)
- Goals
- Assists
- Yellow Cards
- Second Yellow Cards
- Red Cards
- Substituted On
- Substituted Off
- PPM
- Minutes Played

Regression Models

Can we use the stats to predict the market value of a player?

Let's do it with just the raw numerical stats and then the dummied categorical ones.

How do we slice the data?

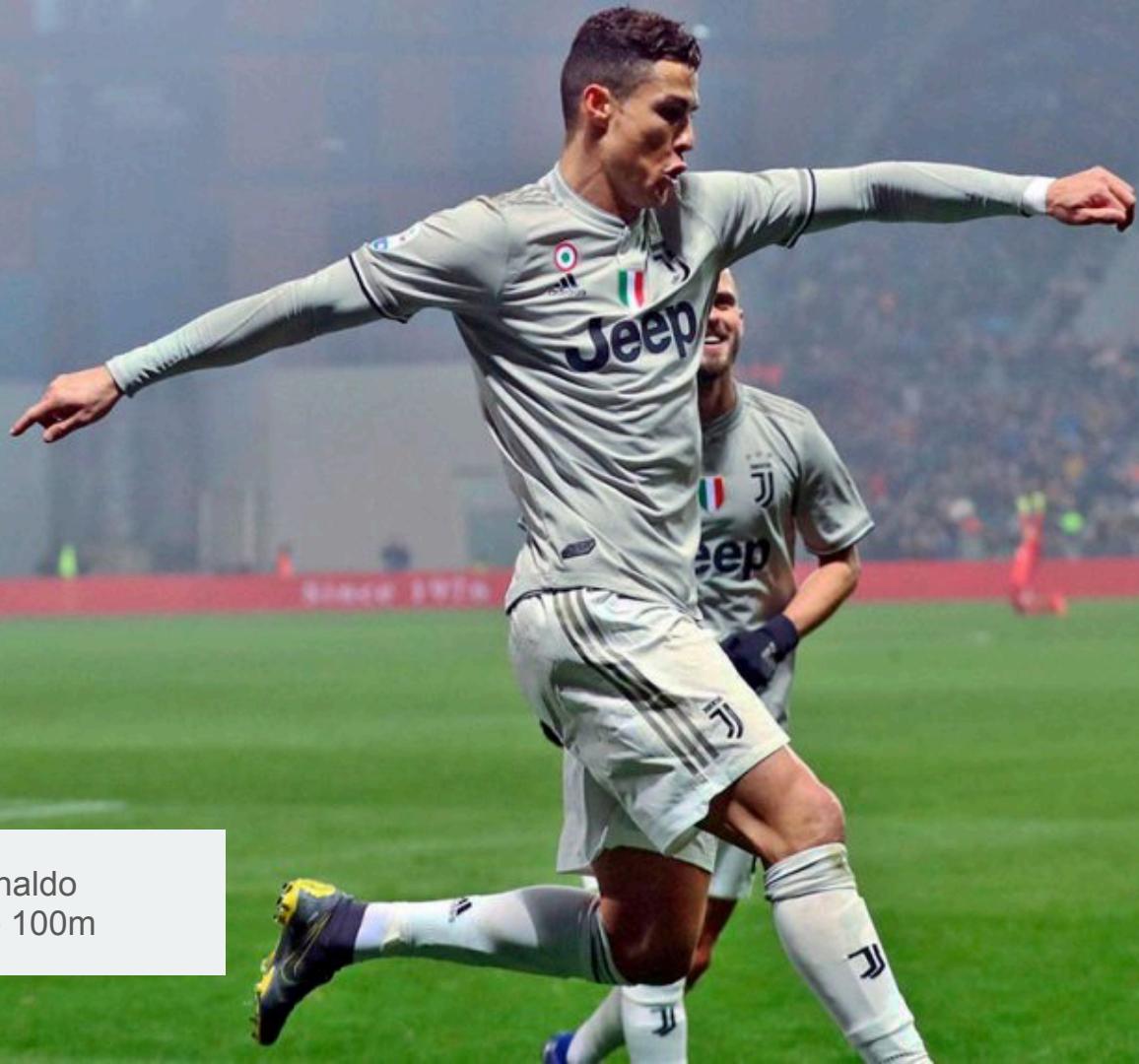
What happens when we feed it to the machine?

Classification Models

Can we use the stats to determine which position a player is?

Use all numerical stats and include market value and transfer value.

Model Results



Cristiano Ronaldo
Market value 100m

Purely Statistics

Without dummying data - taking into account solely the 14 numerical annual statistics:

- Transfer Fee - Included and not in some various as correlation was high at 51%
- Age
- Height - meters
- In-Squad
- Appearances
- Goals
- Assists
- Yellow Cards, Second Yellows and Red Cards
- Substituted On & Off
- PPM
- Minutes Played

Regression Model Performance

No Categorical

Without dummying data

Only numerical statistics, no categorical data.

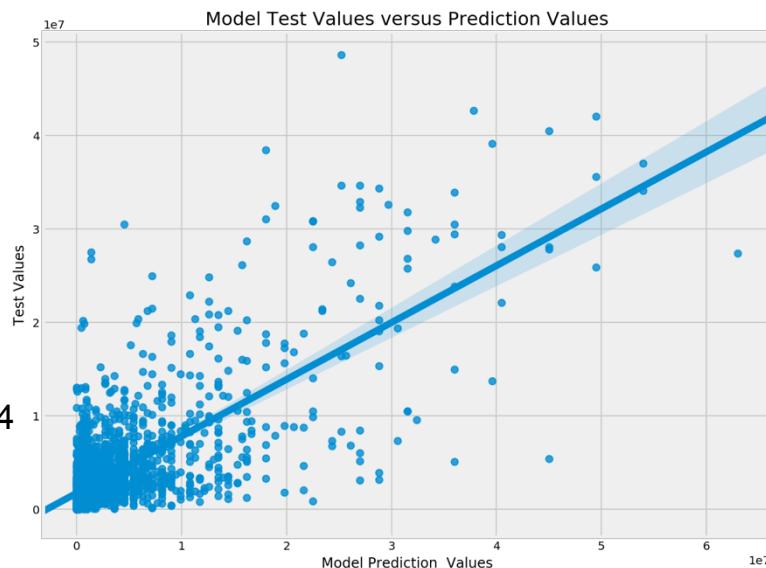
Model for forwards (including transfer fee) :

LinearRegression model R2 Score: 0.585

LinearRegression model MSE: 42,765,455,460,624

LinearRegression model RMSE Train: 6,819,546

LinearRegression model RMSE Test: 6,539,530



Regression Model Performance

No Categorical

Performance of players with more than 540 minutes played ~ 6 games.

16183 observations with 28 features - no dummies

LinearRegression model R2 Score: 0.585

LinearRegression model MSE: 42765455460624.98

LinearRegression model RMSE Train: 6819546.765

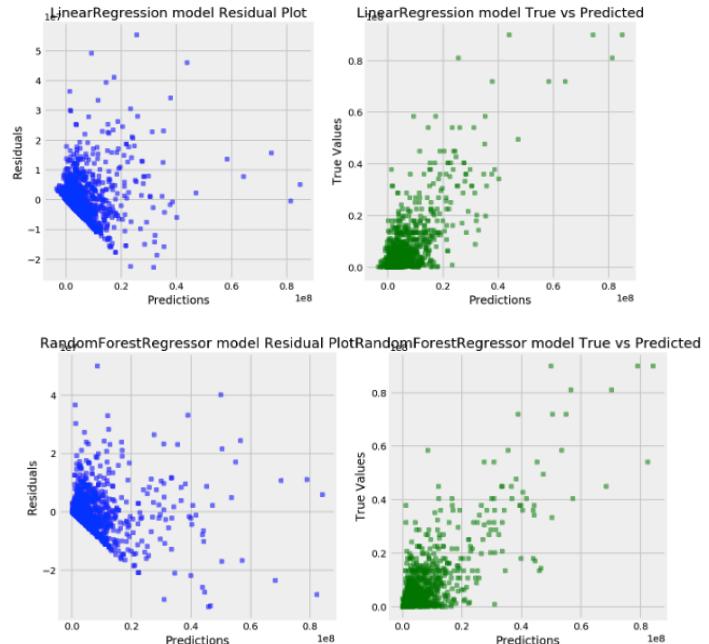
LinearRegression model RMSE Test: 6539530.217

RandomForestRegressor model R2 Score: 0.607

RandomForestRegressor model MSE: 40548077649726.484

RandomForestRegressor model RMSE Train: 2792440.133

RandomForestRegressor model RMSE Test: 6367737.247



Dummying Features

With dummying data - taking into account categorical data:

- Foot
 - right, left, both
- Team
 - 98 different ones
- Position
 - 4 major
- Total Features - 120

Model Performance with Categorical

Best Performance was on players with more than 540 minutes played ~ 6 games.

16183 observations with 130 features - dummied included

LinearRegression model R2 Score: 0.621

LinearRegression model MSE: 23950821762912.28

LinearRegression model RMSE Train: 4860018.106

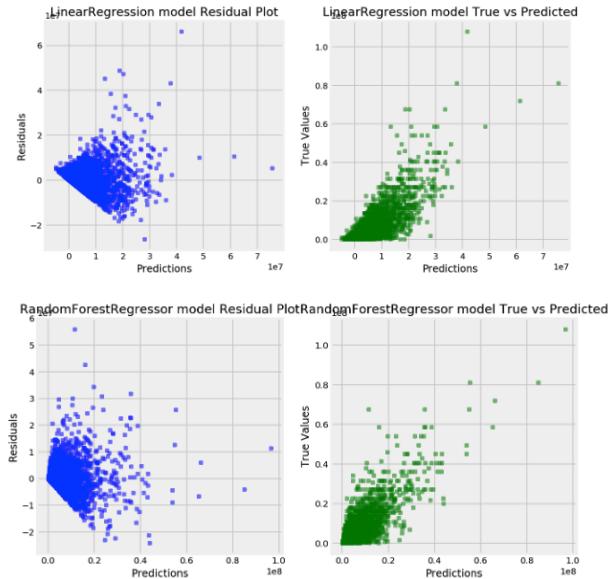
LinearRegression model RMSE Test: 4893957.679

RandomForestRegressor model R2 Score: 0.675

RandomForestRegressor model MSE: 20523282524001.484

RandomForestRegressor model RMSE Train: 2036772.958

RandomForestRegressor model RMSE Test: 4530262.964



Classification Model Performance

Support Vector Classification

SVC model Training Accuracy Score: 0.755

SVC model Testing Accuracy Score: 0.74

DecisionTreeClassifier model Training Accuracy Score: 1.0

DecisionTreeClassifier model Testing Accuracy Score: 0.628

	y_test	predictions	diff	players_position_y_test
0	1.0	2.0	-1.0	Forward
1	1.0	1.0	0.0	Forward
2	3.0	3.0	0.0	Defender
3	4.0	4.0	0.0	Goalkeeper
4	4.0	4.0	0.0	Goalkeeper
5	3.0	3.0	0.0	Defender
6	3.0	3.0	0.0	Defender
7	3.0	3.0	0.0	Defender
8	2.0	1.0	1.0	Midfielder
9	2.0	2.0	0.0	Midfielder



Conclusions

Gather more stats

Get stats that demonstrate strengths of defenders, midfielders and goalkeepers

Dealing with Outliers

Leave them in if you have enough data that supports their value

Much of Futbol is Not Tracked

There is opportunity to find more golden players like Kylian Mbappé, can you use data to do it?

