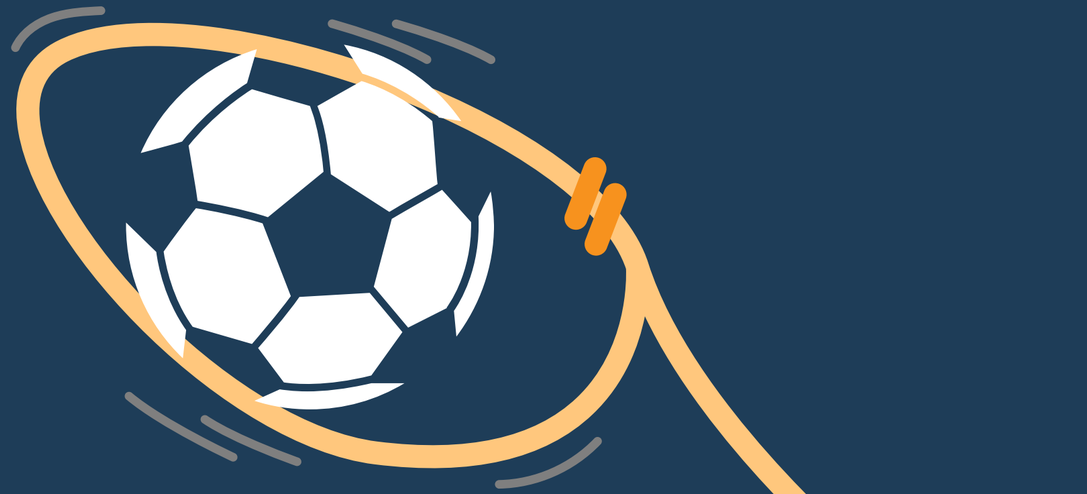
DSI-11

Presented by Niall Anthony McNulty

DSI-11



LASSO THAT FOOTBALL!

Presented by Niall Anthony McNulty

Covered Today

A brief outline

1. Recap of Problem Statement, Goals, Audience & Data Sources
2. Data Collection incl. Methodologies
3. Data Cleaning, Feature Engineering
4. EDA - Visualisations
5. Models
6. Conclusion & Personal Reflection

Recap

Problem Statement

Can footballer valuations be predicted using FIFA player attribute scores & player sentiment?

Recap

Audience

- Sporting directors, head of recruitment, managers etc.
- Clubs in regards to FFP rules
- Football Agents
- Companies engaging in sponsorship deals
- Players
- Betting Companies
- Gamblers
- Football Stock Exchanges

Recap

Goals

- Collect & clean data.
- Add in more features which might correlate with player valuation using web scraping. E.g. How popular a player is via sentiment analysis, minutes played per season, goals scored, goals defended etc.
- Fit a regression model - Evaluate.
- See which factors influence player valuations most/least.
- Improve upon Baseline score.

Recap

Data Sources

- FootyStats
https://footystats.org/download-stats-csv

- FIFA 19 Dataset
https://www.kaggle.com/stefanoleone992/fifa-20-complete-player-dataset#players_20.csv

- Sofifa (Online database for players)
https://sofifa.com/player/183277/eden-hazard/20/159586

Data Collection

The Longgg Hours of Scraping!

- Only included players from the top divisions of each country in Europe
- Data for the 2018/19 season
- Added features to my FIFA Dataset using web scraping
- Beautiful Soup to scrape the league each player played in using each players unique URL from FIFA dataset.
- Selenium to scrape each players likes, dislikes and comments on the SOFIFA website via their unique URL from FIFA dataset.

Standford Core NLP

```
for sentence in annot_doc["sentences"]:
    print ( " ".join([word["word"] for word in sentence["tokens"]]) + " => " \
        + str(sentence["sentimentValue"]) + " = "+ sentence["sentiment"])
```

```
by far best attribute composure and penalties . => 3 = Positive
if you ask me his composure in the box is World class and I kinda Hate to say it cause hes a spurs player . => 1 = Ne
gative
Pen 9394 Comp 93 ew Dude is really not that slow agility is fine but sprint should be a cool 79 . => 1 = Negative
imo hes not that slow ffs Ritmo de 72 o 70 almenos didnt he have 80 sprint speed not so long ago . => 1 = Negative
No hes always had underrated pace 91 potential dont make me laugh . => 3 = Positive
Wtf Kane ist never that slow His Sprint Speed should BE 80 . => 1 = Negative
80 is way too high nahhh 7475 is accurate 72 de ritmo estaria bien 8890 . => 1 = Negative
I think he should have leaved Tottenham when he was at his prime . => 1 = Negative
But if he wants to be a club legend its his choice and I dont see it bad => 1 = Negative
```

Data Cleaning

The Pain of Merging with Fuzzy Wuzzy!

I had a secondary dataset which I wanted to include, which provided goals, clean sheets & minutes played for each player for the given season. I assumed these would be valuable features to include into my modelling to improve accuracy.

HOWEVER I ran into some problems!

- My two dataset did not merge on ANY columns
- The columns they could merge on (NAMES) did not match, as each dataset had either short or long names for the respective players!
- My solution (thanks to Claudia) was to utilize Fuzzy Wuzzy.

Data Cleaning

- Using Regex
- Time-stamp-ordinal
- Dummifying
- Getting rid of redundant features
- Removing Goal Keepers due to missing data

Dataset

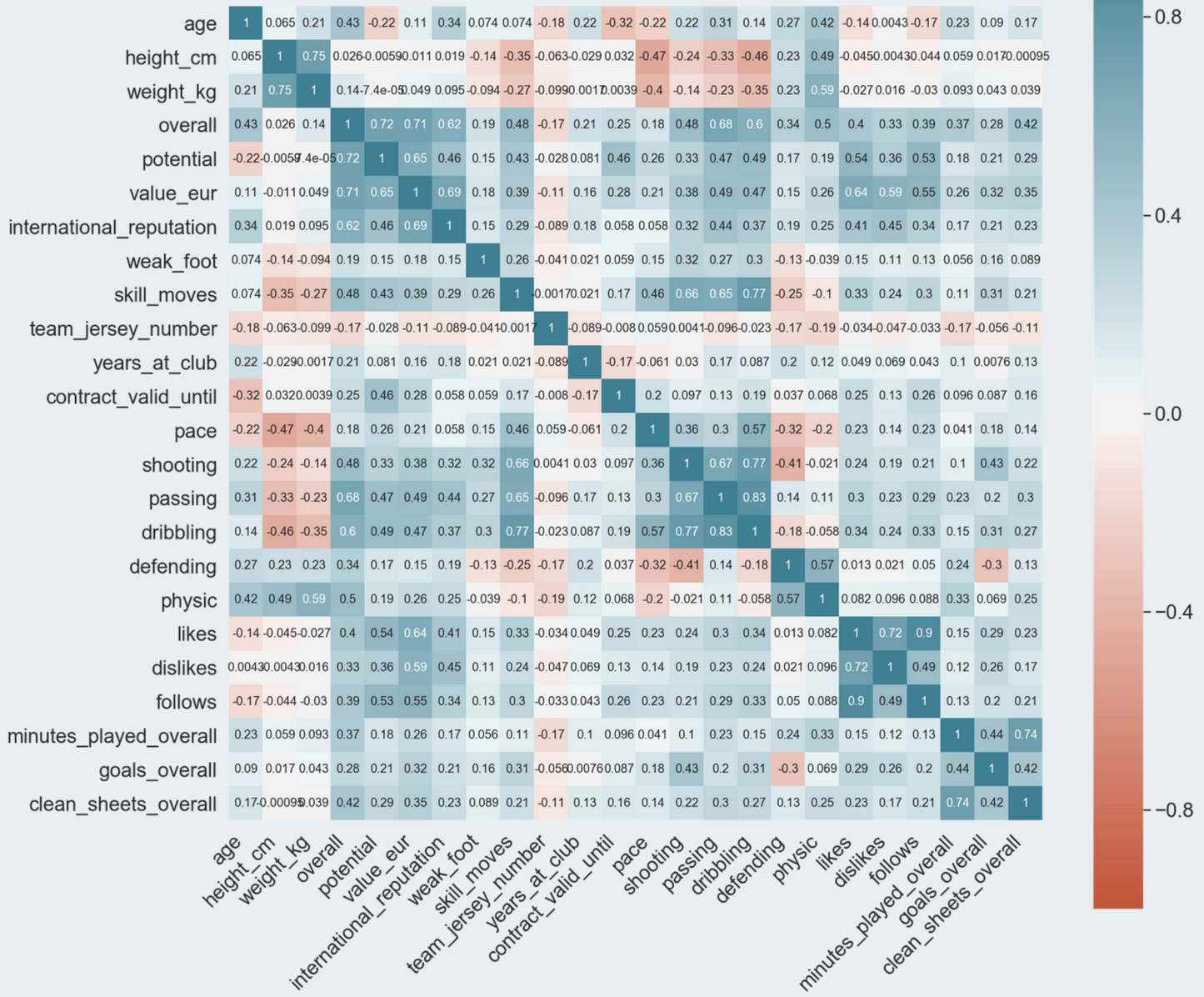
	name	age	height_cm	weight_kg	nationality	club	overall	value_eur	team_jersey_number	years_at_club	...
0	Kevin De Bruyne	27	181	70	Belgium	Manchester City	91	102000000	17	4.493151	...
1	Luka Modrić	32	172	66	Croatia	Real Madrid	91	67000000	10	8.136986	...
2	Eden Hazard	27	173	74	Belgium	Chelsea	91	93000000	10	8.139726	...
3	Diego Godín	32	187	78	Uruguay	Atlético Madrid	90	44000000	2	10.136986	...
4	Toni Kroos	28	183	76	Germany	Real Madrid	90	76500000	8	5.613699	...
...	...	...	...	...	...	...	...	...	...	...	...
4464	Shaun Kelly	28	180	74	Republic of Ireland	Limerick FC	54	50000	2	2.101370	...
4465	David McAllister	29	181	80	Republic of Ireland	Shamrock Rovers	59	150000	16	3.624658	...
4466	Maximilian Kilman	21	186	80	England	Wolverhampton Wanderers	56	180000	49	1.465753	...
4467	Jake Bennett-Rivera	22	180	70	England	Sheffield United	56	100000	33	3.150685	...
4468	Karl O'Sullivan	18	175	70	Republic of Ireland	Limerick FC	53	100000	22	2.024658	...

4469 rows × 100 columns

Features

'name', 'age', 'height_cm', 'weight_kg', 'nationality', 'club',
'overall', 'potential', 'value_eur', 'wage_eur', 'preferred_foot',
'international_reputation', 'weak_foot', 'skill_moves', 'body_type',
'release_clause_eur', 'team_position', 'team_jersey_number',
'years_at_club', 'contract_valid_until', 'pace', 'shooting', 'passing',
'dribbling', 'defending', 'physic', 'attacking_crossing',
'attacking_finishing', 'attacking_heading_accuracy',
'attacking_short_passing', 'attacking_volleys', 'skill_dribbling',
'skill_curve', 'skill_fk_accuracy', 'skill_long_passing',
'skill_ball_control', 'movement_acceleration', 'movement_sprint_speed',
'movement_agility', 'movement_reactions', 'movement_balance',
'power_shot_power', 'power_jumping', 'power_stamina', 'power_strength',
'power_long_shots', 'mentality_aggression', 'mentality_interceptions',
'mentality_positioning', 'mentality_vision', 'mentality_penalties',
'mentality_composure', 'defending_marking', 'defending_standing_tackle',
'defending_sliding_tackle', 'player_league', 'likes', 'dislikes',
'follows', 'comments', 'minutes_played_overall', 'goals_overall',
'clean_sheets_overall

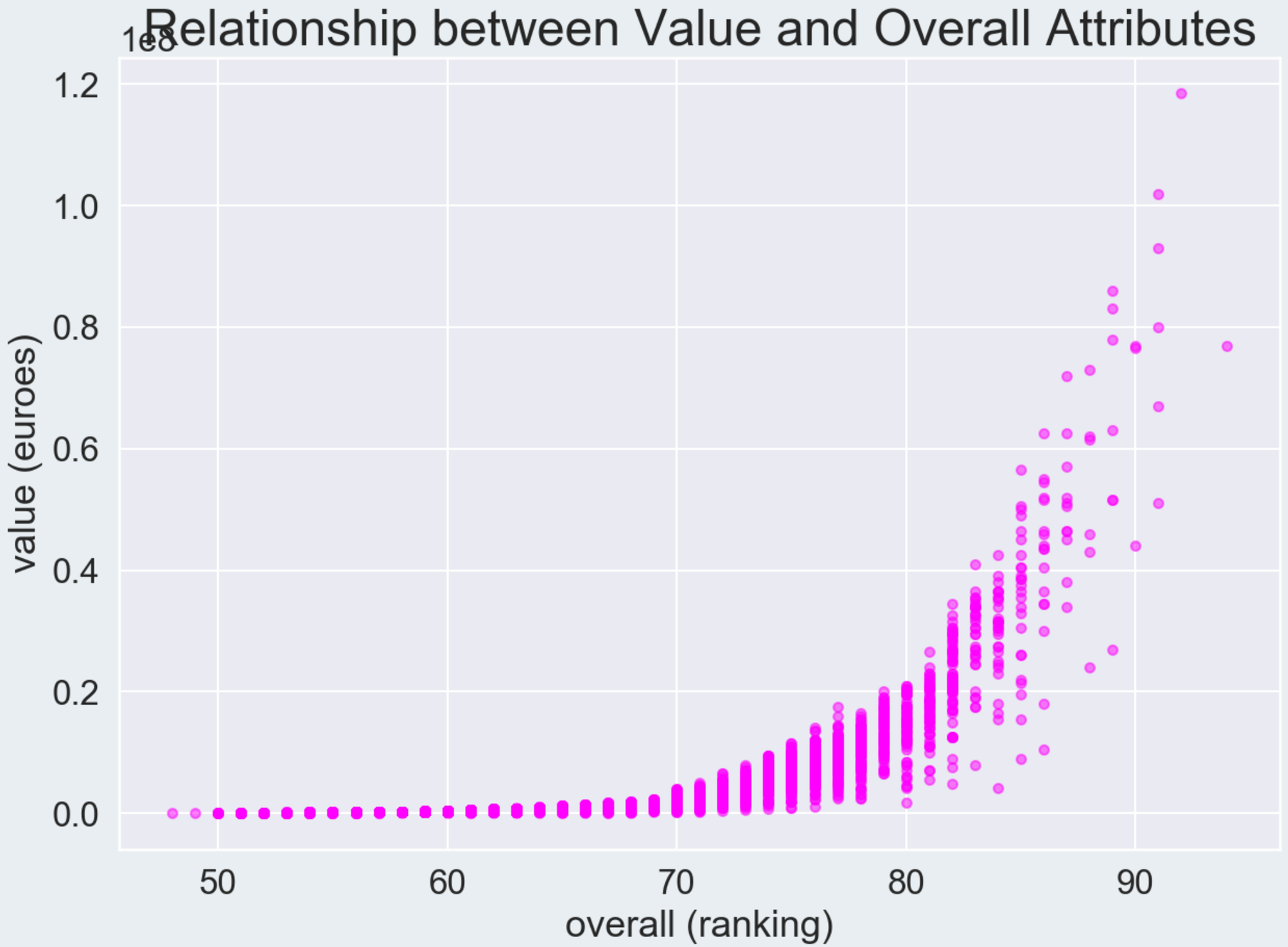
Exploratory Data Analysis



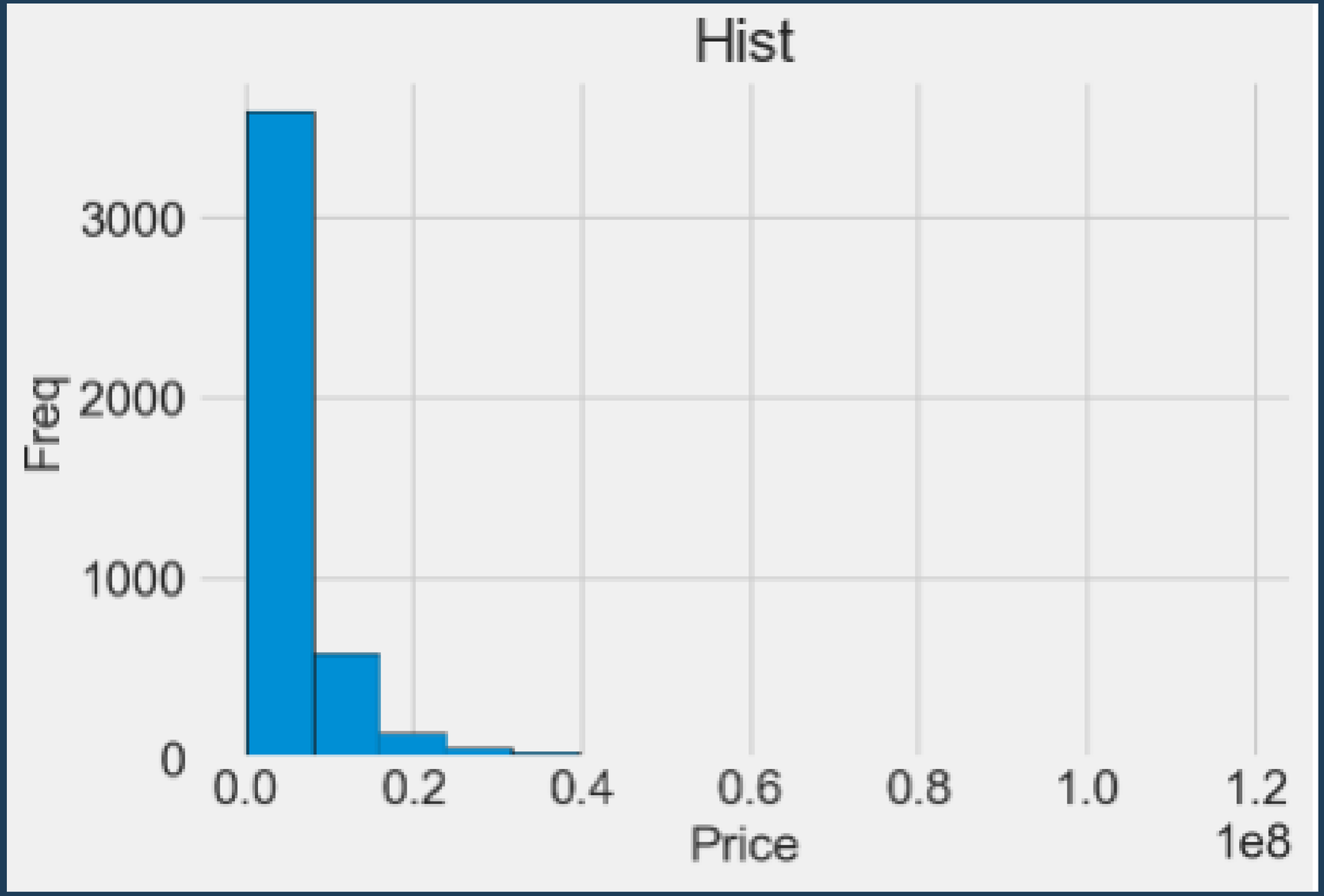
EDA

```
df.describe()
```

	age	height_cm	weight_kg	overall	potential	value_eur	wage_eur	international_reputation	
count	4469.000000	4469.000000	4469.000000	4469.000000	4469.000000	4.469000e+03	4469.000000	4469.000000	44
mean	25.078541	181.492728	75.369881	69.565227	74.391363	4.903731e+06	18744.685612	1.252629	
std	4.363965	6.168933	6.600704	7.179908	6.343963	8.572204e+06	34176.708126	0.569632	
min	16.000000	161.000000	54.000000	48.000000	54.000000	4.000000e+04	1000.000000	1.000000	
25%	22.000000	177.000000	70.000000	65.000000	70.000000	5.750000e+05	2000.000000	1.000000	
50%	25.000000	182.000000	75.000000	70.000000	74.000000	1.500000e+06	6000.000000	1.000000	
75%	28.000000	186.000000	80.000000	75.000000	79.000000	6.000000e+06	21000.000000	1.000000	
max	39.000000	202.000000	101.000000	94.000000	95.000000	1.185000e+08	455000.000000	5.000000	


Relationship between Value and Overall Attributes
1e8
1.2
1.0
0.8
0.6
0.4
0.2
0.0
value (euroes)
50
60
70
80
90
overall (ranking)

EDA



MODELLING

Linear Regression (Baseline 82%)
Lasso
Ridge

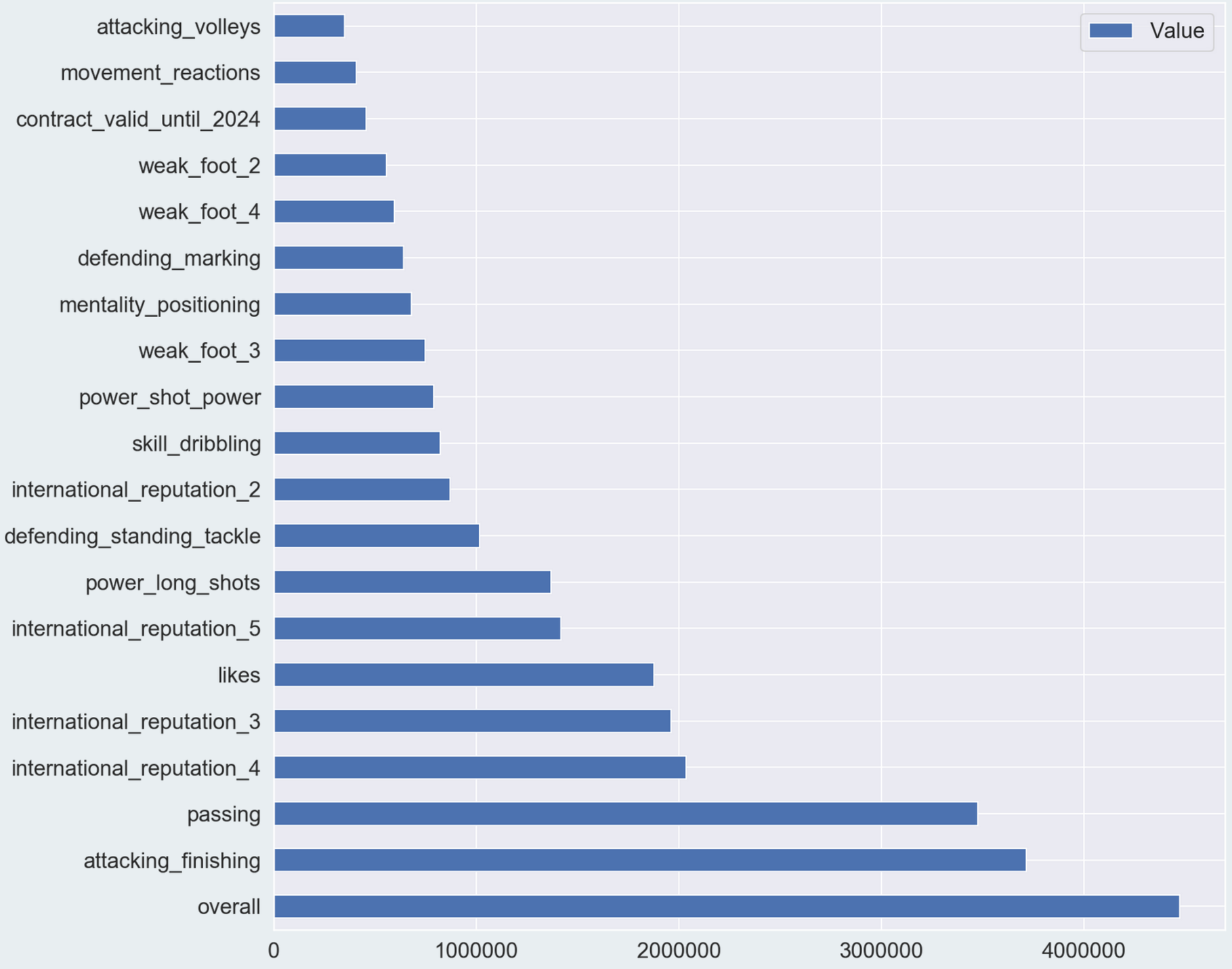
More to come...

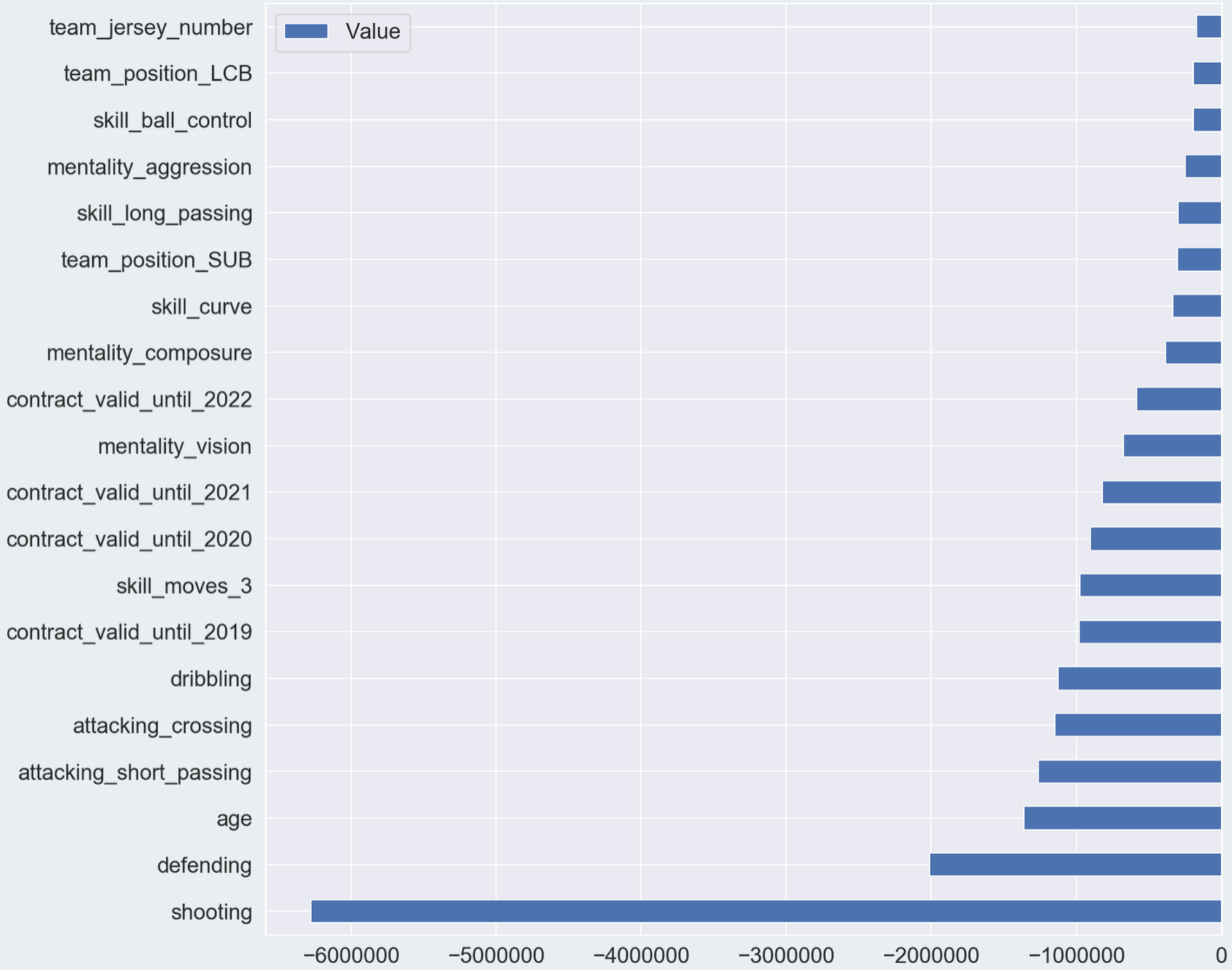
Time Series
Decision Trees

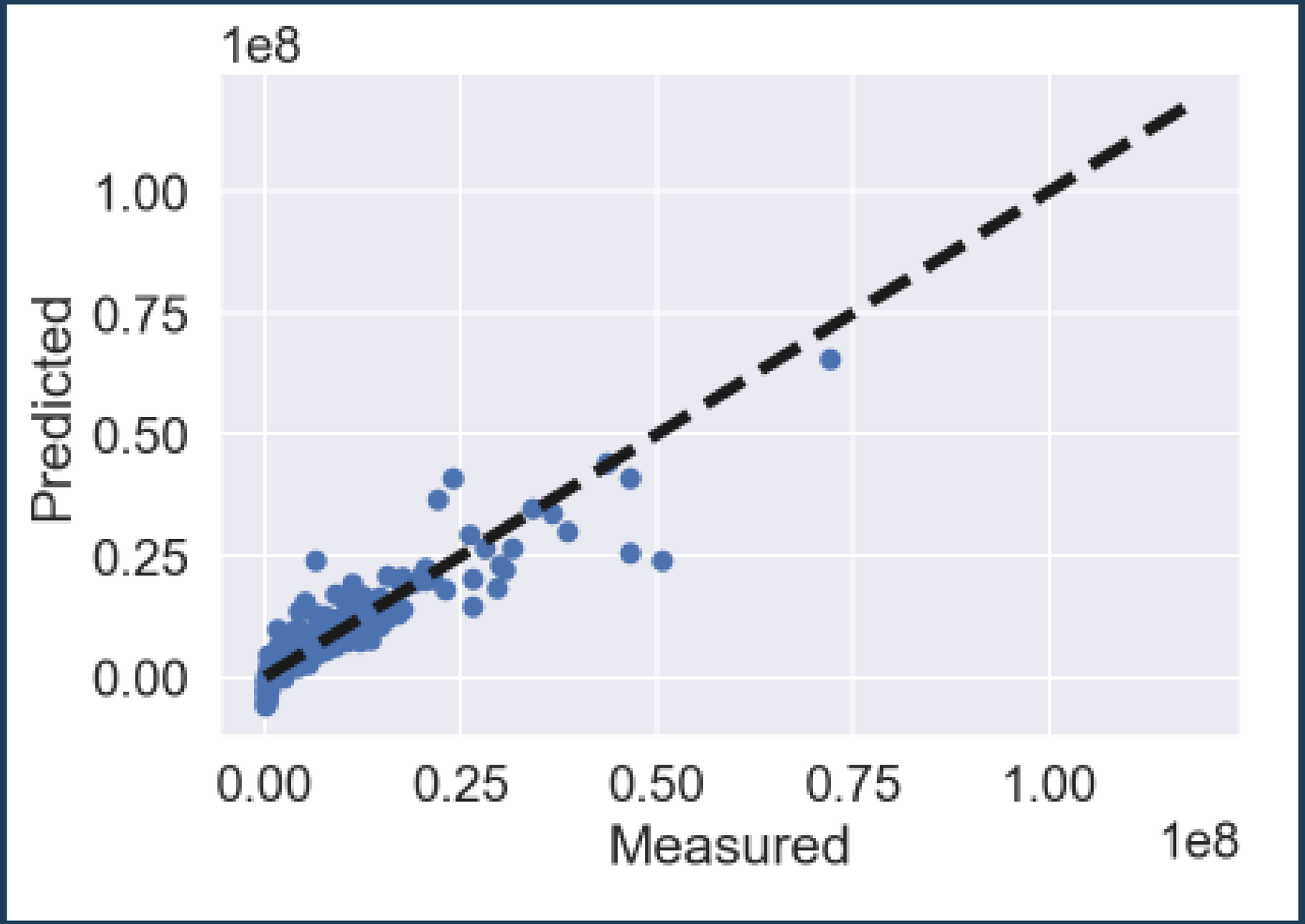
1.Linear Regression
Cross-validated training scores: [0.71650618 0.80680949 0.73366917 0.7005549
0.6502284]
Mean cross-validated training score: 0.7215536274177794
Training Score: 0.7847374440537804
Test Score: 0.8284794374627071

2. Ridge
Best alpha: 166.81005372000558
Training score: 0.7811607418035735
Test Score: 0.8337889352093124

3. Lasso
Best alpha: 10000.0
Training score: 0.7900953486726378
Test Score: 0.7851033056393415


Value
attacking_volleys
movement_reactions
contract_valid_until_2024
weak_foot_2
weak_foot_4
defending_marking
mentality_positioning
weak_foot_3
power_shot_power
skill_dribbling
international_reputation_2
defending_standing_tackle
power_long_shots
international_reputation_5
likes
international_reputation_3
international_reputation_4
passing
attacking_finishing
overall
0
1000000
2000000
3000000
4000000


Value
team_jersey_number
team_position_LCB
skill_ball_control
mentality_aggression
skill_long_passing
team_position_SUB
skill_curve
mentality_composure
contract_valid_until_2022
mentality_vision
contract_valid_until_2021
contract_valid_until_2020
skill_moves_3
contract_valid_until_2019
dribbling
attacking_crossing
attacking_short_passing
age
defending
shooting
−6000000
−5000000
−4000000
−3000000
−2000000
−1000000
0


1e8
1.00
0.75
0.50
0.25
0.00
Predicted
0.00
0.25
0.50
0.75
1.00
Measured
1e8

Conclusion

- Consider doing Time Series
- Consider finding some more relevant features. E.g. Player's agent, Endorsements
- Figure out Sentiment Analysis (use Vader - with a compound score) and potentially use players Twitter rather than the Fifa forum
- Improve upon models
- Improve on my time management (avoid going off on tangents and wasting hours upon hours)
- Improve on the technical side of things - have a clearer understanding of how the models work etc.
- Test the model on new data (2019/2020 Season Stats)