Project EA Sports - FIFA

CS109a Final Project

11 December 2020

Group 4 – Clement Brenot, Luis Corrales, Ryuichiro Hashimoto, Yuki Yaguchi

1

Agenda

1. Overview
2. Part A – Rank the Players
3. Part B – Classify Player Position
4. Part C – Which Club has the Best Staff?
5. Part D – How Things will be in 2021?
6. Possible Future Improvements

2

1. Overview

3

Data and Selected EDA Results

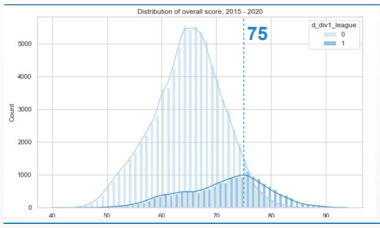
Data pre-processing

Raw Data
5 files
104 variables

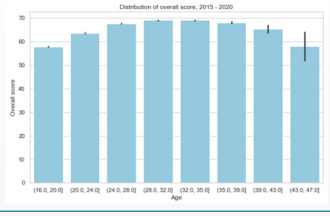
→ Merge → One-hot encode → Simplify & clean → Impute → Standardize →

Pre-processed data
1 dataset
100,995 x 146

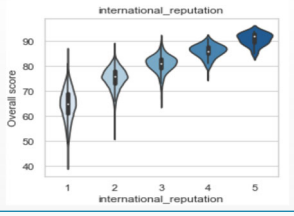
A good overall score? 75 and up



Old players decline… but not always



International reputation is a guarantee



4

2. Part A – Rank the Players

5

Methodology

Pre-processing

Target variable:
Overall (player's overall rating)

Overall (co)

*co: continuous

Training dataset:
All players in FIFA 2019 dataset

FIFA 2019 dataset

Test dataset: players in the clubs below (FIFA 2020 dataset)

- FC Barcelona
- FC Bayern München
- Real Madrid
- Paris Saint-Germain
- Juventus
- Manchester City
- Liverpool FC

Data processing flow:

One-hot-encoding for categorical variables
Ex.) player traits, tags, etc.

Imputation and Standardization
Ex.) goal keeping abilities for field players, etc.

EDA and Regression

Model creation

Model type	Tuning parameters
XGBoost	Tree parameters; Regularization parameters
Random Forest	Tree depth; # of trees
LASSO	L1 norm parameter

Hyperparameter tuning using random search with cross validation

Final model

- XGBoost model is chosen

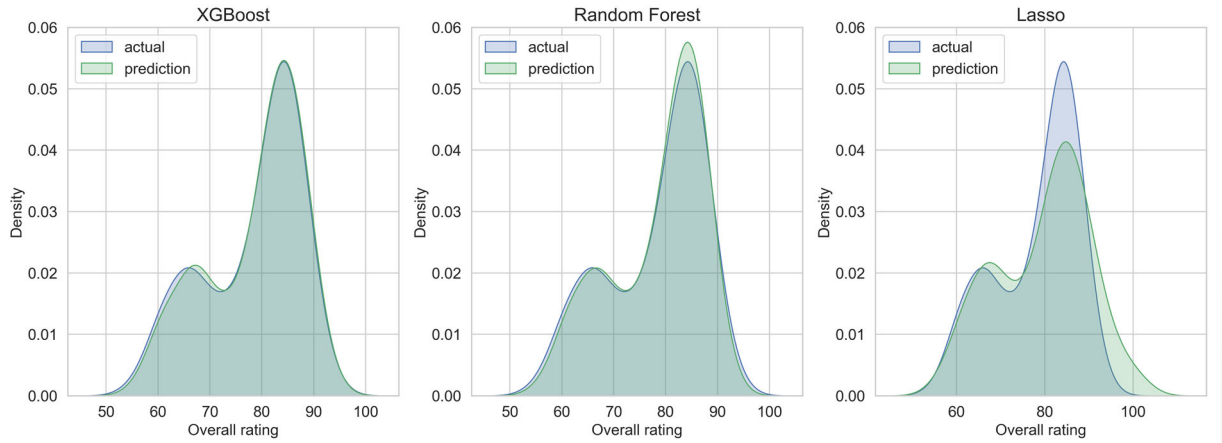
	Evaluation metrics: Test MSE	Evaluation metrics: % of obs within +/- 2
XGBoost	0.68	97.7%
Random Forest	0.72	96.4%
LASSO	7.99	60.2%

"% of obs within +/- 2" shows the proportion of observations that fall into +/- 2 of predicted overall rating.

6

Overall Score in Test Set – Actual vs. Prediction, by models

- XGBoost almost completely predicted overall ratings for almost all players.
- LASSO especially misses the highest range, potentially due to its linear equation form.



3. Part B – Classify Player Position

Methodology

Pre-processing

Target variable:
Player position

Player position (mc)

* mc: multiclass

Training dataset:
All players in FIFA 2019 dataset

FIFA 2019 dataset

Test dataset:
All players in FIFA 2020 dataset

FIFA 2020 dataset

Data processing flow:

Creation of new grouped category of players
Forward, Midfilder, Defender and Goalkeeper.

Drop of Variables that bias the sample
Ex.) goal keeping abilities only available for goalkeepers and missing values for the rest

Classification

Model creation

Model type	Tuning parameters
Logistic	Ridge-like regularization
Decision Tree	Tree depth
AdaBoost	Tree depth (same as Decision Tree)
XGBoost	Hyperparameter tuning using random search with cross validation
Random Forest	Tree depth

Final model

- AdaBoost model is chosen

Evaluation metrics: Grouped Positions

	Test Accuracy	Precision	Recall
Logistic	0.951	0.956	0.958
Decision Tree	0.938	0.946	0.945
AdaBoost	**0.962**	**0.967**	**0.967**
XGBoost	0.958	0.959	0.967
Random Forest	0.958	0.962	0.963

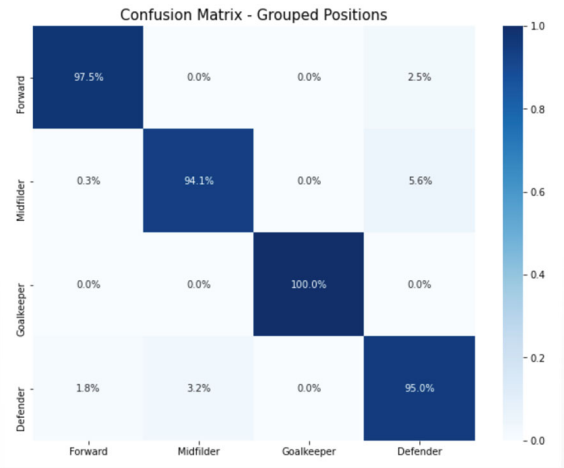
Evaluation metrics: All Positions

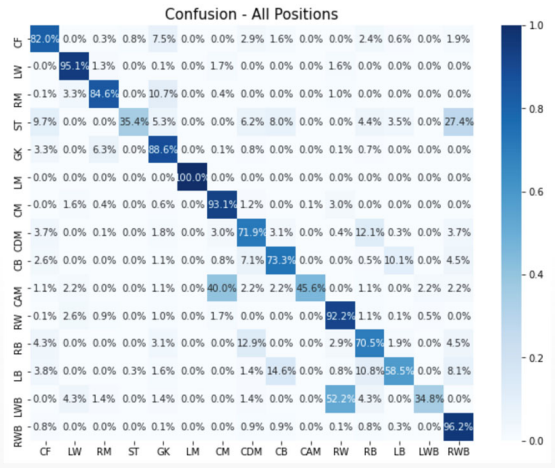
	Test Accuracy	Precision	Recall
Logistic	0.865	0.783	0.720
Decision Tree	0.826	0.687	0.616
AdaBoost	**0.885**	**0.838**	**0.753**
XGBoost	0.883	0.799	0.792
Random Forest	0.868	0.816	0.715

9

9

Confusion Matrix for Predicted Positions

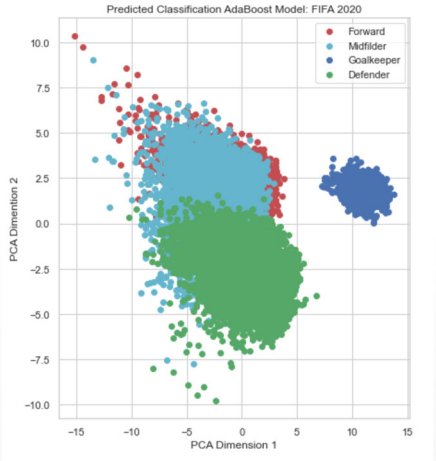


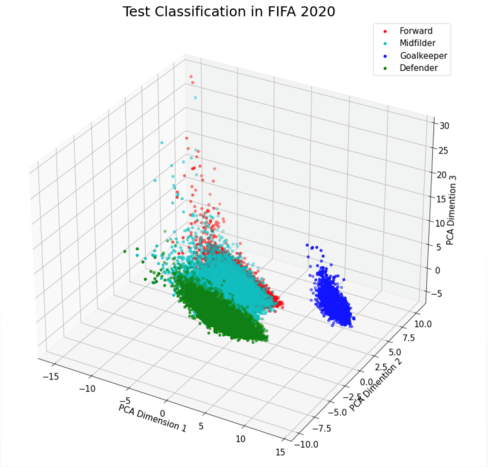


10

10

PCA components of Grouped Positions:
2 and 3 Components





11

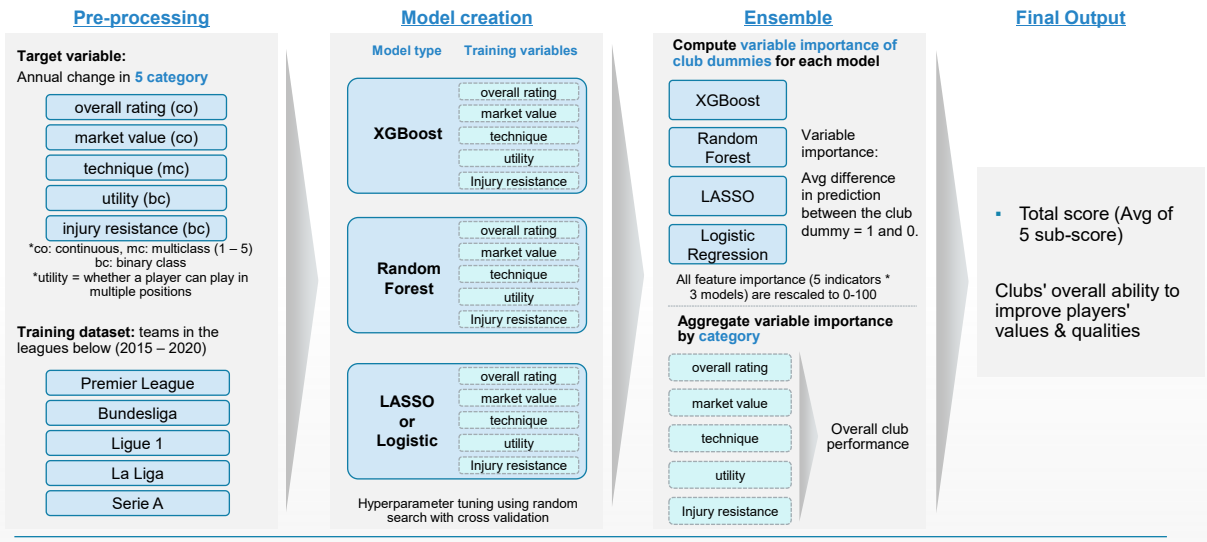
11

4. Part C – Which Club has the Best Staff?

12

12

Methodology



Scoring Methodology

- Each sub score (1 – 5) is scaled to 0 – 100 based on variable importance from 3 models.
- Total score is the average of sub scores

	League	Club	Total Score	(1) overall rating	(2) Market value	(3) Technique	(4) Utility	(5) Injury resistance
1	French Ligue 1	Amiens SC	70.9	100.0	81.4	72.2	39.2	67.4
2	Italian Serie A	Atalanta	66.9	77.8	100.0	62.9	37.6	58.1
3	English Premier League	Tottenham Hotspur	62.7	57.8	60.2	72.9	62.1	67.6
4	French Ligue 1	RC Strasbourg Alsace	61.9	52.2	75.0	93.2	30.6	62.2
5	Spain Primera Division	RCD Espanyol	59.3	62.3	50.3	58.3	76.1	50.6
:	:	:	:	:	:	:	:	:
94	French Ligue 1	Stade de Reims	38.2	12.9	16.8	59.9	33.4	58.0
95	French Ligue 1	Stade Brestois 29	37.8	23.4	34.6	39.9	25.5	50.1
96	German 1. Bundesliga	Fortuna Düsseldorf	35.2	18.3	35.1	33.3	15.3	57.6
97	Spain Primera Division	CA Osasuna	32.6	4.0	6.3	31.9	55.7	51.1
98	Spain Primera Division	Granada CF	28.4	0.0	0.0	40.4	31.4	52.7
		Max	70.9	100.0	100.0	100.0	100.0	100.0
		Mean	48.2	38.3	46.7	58.7	35.2	55.1
		Min	28.4	0.0	0.0	0.0	0.0	0.0
		Standard deviation	6.8	13.5	12.9	17.2	15.8	12.7

Top 3 clubs vs. Bottom 3 clubs

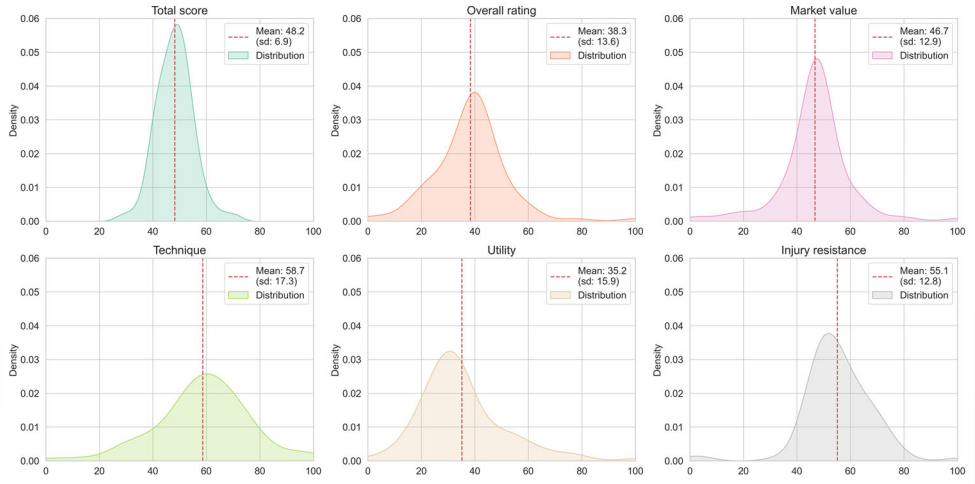
- Amiens SC scores the best in player overall score, while Atlanta improves market value.
- Bottom 3 clubs' performance are poor, especially for these two metrics.

League Average

- English Premier League and French Ligue 1 are, on average, better than other 3 leagues.

Score Distributions

- Technique and Utility have a wider distribution than other scores.

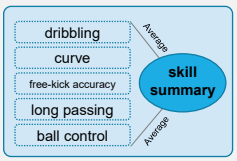


5. Part D – How Things will be in 2021?

Methodology

Pre-processing

Target variable:
skill_summary(average of all skills) of the following year



Training dataset: teams in the leagues specified in instruction (below)

- Premier League
- Bundesliga
- Ligue 1
- La Liga
- Serie A

Model creation

Model type	Training data
XGBoost	2015 / 2016, 2016 / 2017, 2017 / 2018, 2018 / 2019, 2019 / 2020
Random Forest	2015 / 2016, 2016 / 2017, 2017 / 2018, 2018 / 2019, 2019 / 2020
Elastic Net	2015 / 2016, 2016 / 2017, 2017 / 2018, 2018 / 2019, 2019 / 2020

Hyperparameter tuning using random search with cross validation

Ensemble

Weight by year of training data

2015 / 2016	0.15
2016 / 2017	0.15
2017 / 2018	0.20
2018 / 2019	0.20
2019 / 2020	0.30

* Model trained on recent year had lower test MSE in general.

Weight by model

XGBoost	0.30
Random Forest	0.40
Elastic Net	0.30

* Result (better to worse): Random Forest > Elastic Net > XGBoost

Final model

- Ensemble of 15 models
- Predict on 2020 dataset to predict skill_summary score in 2021

The model result (test MSE score)

Test dataset: skill_summary score in 2020 predicted by data in 2019
Above number in each cell is test MSE score, the number below is its weight in the entire 15 models (%)

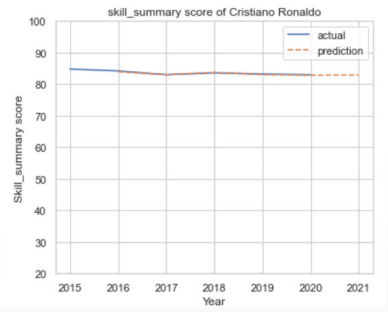
Test MSE / Weight in final ensemble model	2015 / 2016 (year weight: 0.15)	2016 / 2017 (year weight: 0.15)	2017 / 2018 (year weight: 0.20)	2018 / 2019 (year weight: 0.20)	2019 / 2020* (year weight: 0.30)
XGBoost (model weight: 0.30)	8.15 4.5%	6.33 4.5%	6.25 6.0%	6.07 6.0%	2.15 9.0%
Random Forest (model weight: 0.40)	6.58 6.00%	5.27 6.00%	5.24 8.00%	5.12 8.00%	0.76 12.0%
Elastic Net (model weight: 0.30)	7.90 4.5%	5.05 4.5%	5.02 6.0%	4.94 6.0%	4.81 9.0%

* Because 2019 / 2020 model is trained on the same data as test dataset, this is for reference only

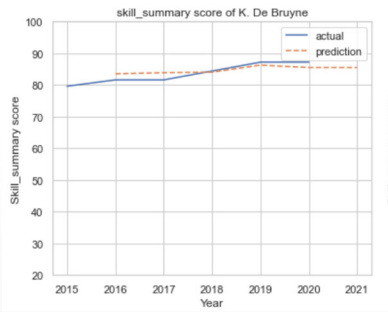
The model result (examples)

We applied the ensemble model to make prediction for skill_summary of specified players in the past years and compared them with actual result.

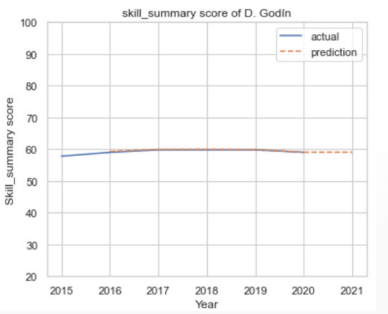
Cristiano Ronaldo



Kevin De Bruyne



Diego Godín



6. Possible Future Improvements

Possible Future Improvements

Do we agree with our client on definitions and methodological choices?

Could neural networks further improve predictions?