**CS 109a – Fall 2020**

**Final Project**

**EA Sports – FIFA**

**Final Report**

This final report summarizes the work undertaken by Group 4 on the EA Sports – FIFA project.

It is structured as follows:

* Introduction - Presentation of data and main results of exploratory data analysis (EDA)
* Part A - Rank the Players
* Part B - Classify player position
* Part C - Which Club has the Best Staff
* Part D - How will things be in 2021
* Conclusion – Caveats and possible future improvements

Impact statement

This project is solely based on-universe data from EA’s Sports FIFA series of games, as opposed to real-life football player characteristics and attributes. In addition, it is mining data that is mostly not directly visible to the player but blended into a comprehensive player experience.

However, this is not to say that there are no ethical and societal considerations to be had. Given the popularity of the game and the increasing realism in the graphical rendering of at least some players’ traits, it is likely that a player’s portrayal in the videogame does influence the perception of the actual players.

ADD point about nationality?

**Introduction - Data and main results of EDA**

**The data**

The data used originates from website sofifa.com. It was provided by the teaching team and did not require web scraping or data collection from the project team.

The data was contained in six csv files, each covering a year from 2015 to 2020. Each file contained between 14,881 and 18278 observations, each observation corresponding to a player. For each observations, 104 variables were available, which were a mix of categorical and quantitative variables covering elements of identification and numerous characteristics of the player. The data format was well harmonized between all six files. A separate file contained the match between teams and leagues.

To prepare the data for analysis, the following main steps were taken:

* **Merger** - Creation of a year variable and merger of the 6 datasets
* **One-hot encoding** - Creation of dummy variables for each player trait and each player tag
* **Simplification** - Selection of one primary position for each player (the first one listed) and creation of a dummy for primary positions
* **Other cleaning steps**, including deleting +/- signs and deleting unnecessary columns
* **Data imputation** – Missing goalkeeping skills for non-goalkeepers were imputed as 0.
* **Standardization** – All quantitative variables were normalized with a mean of 0 and standard deviation of 1

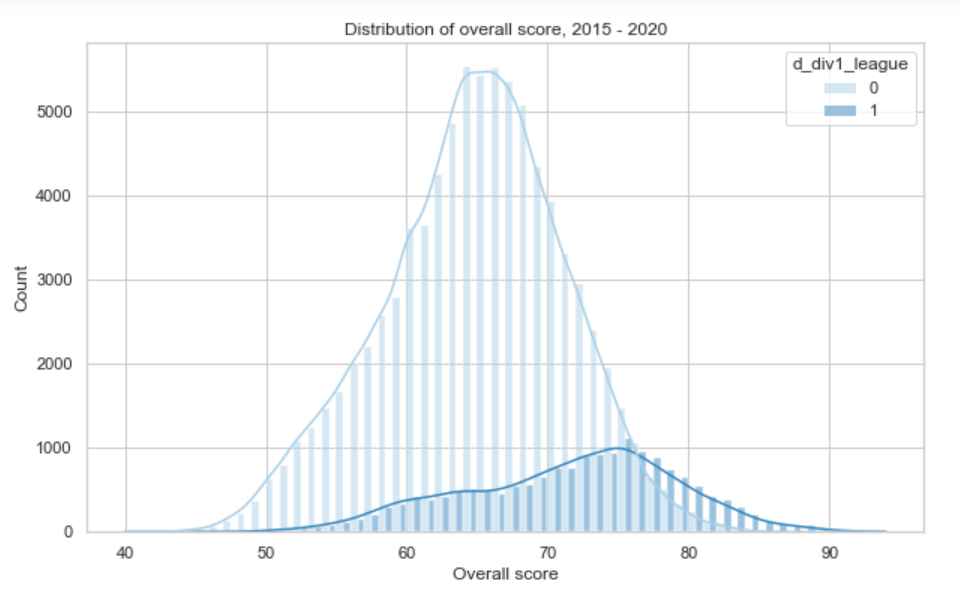
The resulting dataset contains **100,995 observations (player x year) of 146 variables.**

**Main results of EDA**

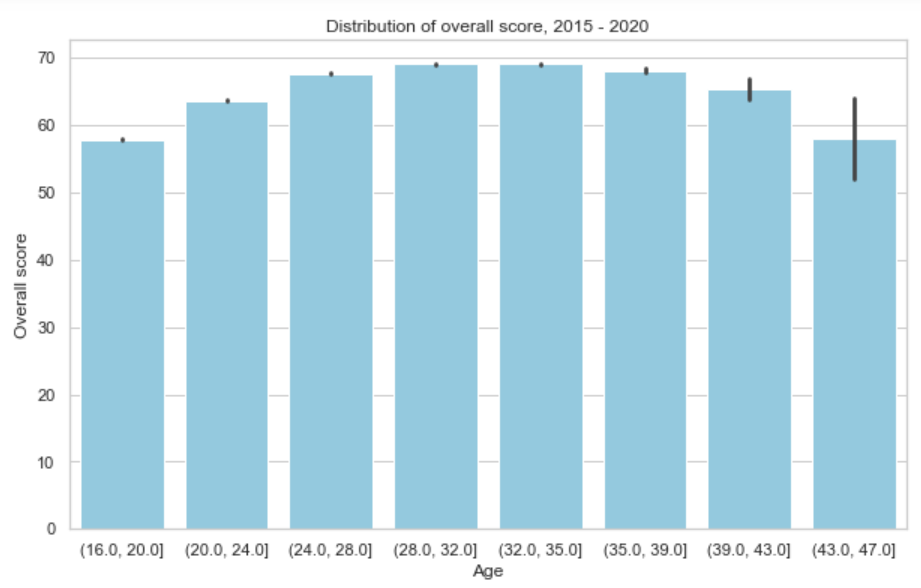
The EDA was mainly carried out with Part A in mind (please see below).

The most salient facts from the EDA include:

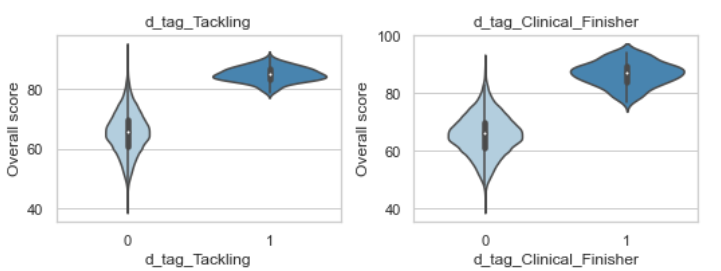
* **The distribution of OVR scores seems almost normal at first sight**. In fact, it is normal for players who do not play in Division 1 Leagues, but seems left-skewed for players of Division 1 Leagues. The modal OVR score for a Division 1 player is about 75, and about 65 for a player which does not play in a Division 1 club.



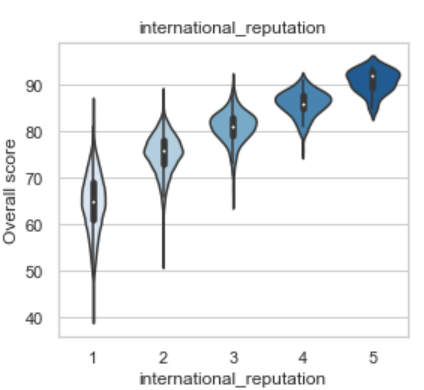
* There seems to be **a quadratic relationship between age and OVR.**



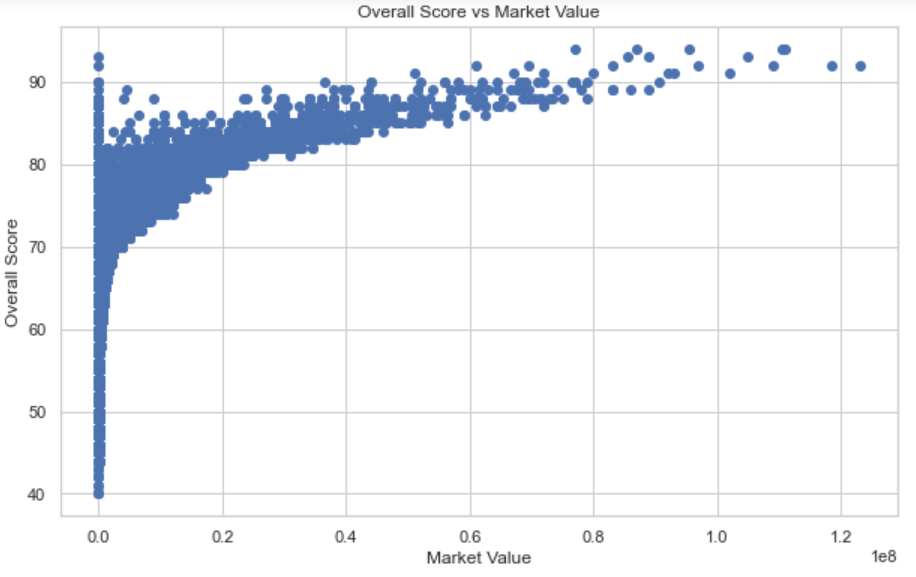
* Certain player tags seem associated with outstanding OVR values, including “tackling”, “clinical finisher” and several others. All tags can be considered positive, as none is negatively associated with a lower OVR. Player traits seem much less associated with OVR. A notable exception is “1 on 1 rush”.



* A high international reputation is associated with high OVR. A low international reputation is less predictive of OVR level.



* Market value seems to have a complex relationship with OVR and is a particularly weak predictor for high market values



**Part A - Rank the Players**

**Methodology**

*Objective*

The objective of Part A is to predict the OVR statistic for 2020 based on 2019 data. The OVR variable is quantitative, this is a regression task.

*Modelling, training and evaluation*

Three models were used: a boosted tree model (XGBoost), a random forest model and a lasso model. The models were trained on the 2019 full dataset, and model hyperparameters were tuned using a grid search on the training set. Results were evaluated as requested on a 7-club subset of the 2020 data.

**Results**

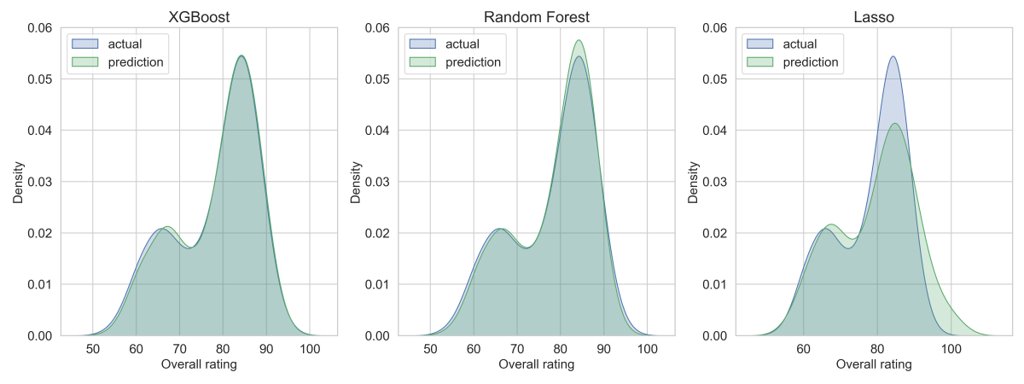
The XGBoost model yielded the most accurate results, with an MSE of 0.68 on the test set and 97.7% of predictions within +/- 2 of the actual values on the test set (Table A.1). The random forest model performed similarly, while the LASSO model performed markedly less well. The superior performance of tree-based models is likely to reflect the benefit of the flexibility of this class of models.

**Table A.1.**

|  |  |  |
| --- | --- | --- |
| **Model class** | **MSE**  **(test set)** | **% of predictions within +/-2 of actual (test set)** |
| **XGBoost** | **0.68** | **97.7%** |
| Random Forest | 0.72 | 96.4% |
| LASSO | 7.99 | 60.2% |

The superior performance of the XGBoost model is most obvious when graphing the distribution of predicted values against the distribution of actual values (Fig. A.1.). It is visible that the XGBoost algorithm better matches OVR score all along the score curve. In contrast, the random forest model seems to slightly underestimate extreme values. The LASSO model presents the opposite bias, missing a large part of the modal values.

**Figure A.1.**



**Part B - Classify player position**

**Methodology**

*Objective*

The objective is to predict player positions based on other statistics, training on 2019 data and testing on a subset of 2020 data. Player positions being categorical variables, this is a classification task.

*Modelling, training and evaluation*

Four competing models were used: a multinomial model based on logistic regression, a simple decision tree model, a boosted tree model (AdaBoost) and a random forest model. Models were trained on the full 2019 dataset. Model model hyperparameters were tuned using a grid search on the training set.

Models were used to predict primary player position, that we define as the first position given in the raw dataset. There were 15 classes. In addition, noting that the choice of primary position may be arbitrary and result in noisy data, we also ran the exact same four approaches to predict broad player position, defined as goalkeeper, defender, midfielder or forward (4 classes). As a result, a total of 8 models were trained and evaluated.

Results were evaluated as requested on a 7-club subset of the 2020 data, using accuracy as a metric.

**Results**

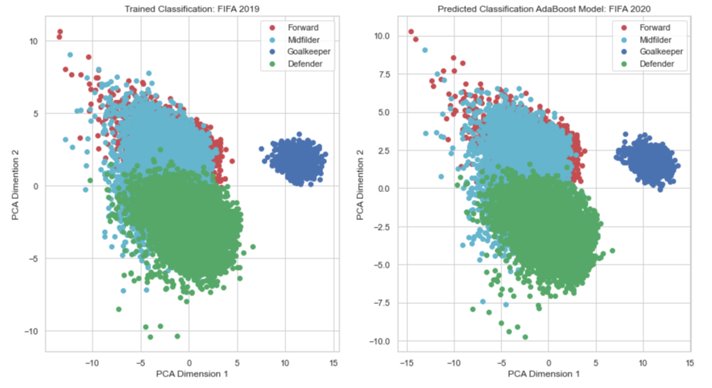
The best predictions came from the XXX model, with an accuracy of xx.x% on the test set. The xxx model performed second (xx%), xxx model third (xx%), and sss model came fourth (xx%).

**Table A.1.**

|  |  |  |
| --- | --- | --- |
| **Model class** | **Accuracy on test set**  **Primary position (15 classes)** | **Accuracy on test set**  **Broad position**  **(4 classes)** |
| Multinomial | XX% | 95.1% |
| Simple decision tree | XX% | 93.8% |
| **AdaBoost** | **XX%** | **96.2%** |
| Random Forest | XX% | 95.8% |

A visual analysis of the classification results using principal components analysis (PCA) illustrates the non-obvious character of the classification task. Even in the 4-class version of the problem, we observe considerable overlap between classes when visualized in a 2-dimension space, especially between forwards and midfielders (Figure B.1). In contrast, goalkeepers appear very separated.

**Figure B.1.**



Possibly some more discussion using the confusion matrix.

**Part C - Which Club has the Best Staff**

**Methodology**

*Objective*

The objective is to infer staff quality on the basis of player improvement while at the club. This is a regression task.

*Definitions*

We define player improvement as the year-on-year increase in the simple average of 5 standardized variables: overall rating, market value, technique, utility, injury resistance. We define staff quality as the parameter associated with a club dummy in the model.[[1]](#footnote-2)

*Data preparation*

To add

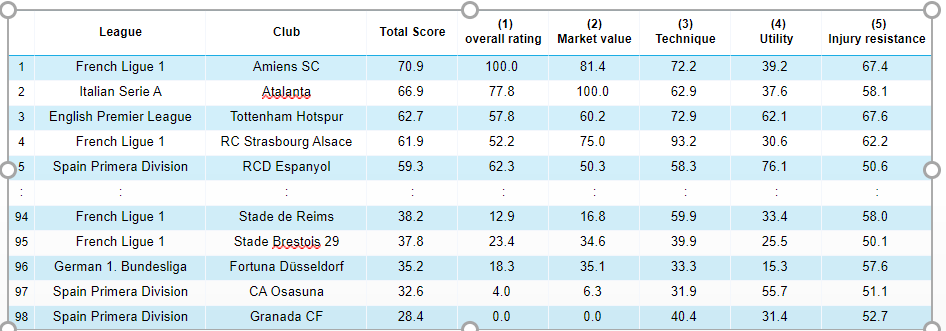
*Modelling, training and evaluation*

For each of the 5 sub-dimensions player improvement that we used, we trained three types of model: an XGBoost boosted tree model, a random forest model, and a LASSO (for quantitative variables) or a logistic model (for categorical variables). Each model’s hyperparameters were tuned using a random search by cross-validation. Staff quality is approximated by the variable importance associated with the club dummy (predicted variable with dummy = 1 and predicted variable =0). These 3 values are averaged and then rescaled on a 0-100 scale. The 0-100 values obtained on each variable are then averaged to result in the main club total score.

**Results**

Results by club may provide a clearer idea of the construction of our synthetic indicator, but may be difficult to read (Table B.1). Observing top 5 and bottom 5 performers, we identify top, middle and low-ranking clubs. We also identify clubs from all 5 leagues studied.

**Table B.1.**



An visualization of results by league is more revealing, showing that the English Premier League, French Ligue 1 and German Bundesliga seem to benefit from better staff than the Italian Serie A and Spanish Liga (Fig B.1). It also seems that across all leagues club staff has most influence on technique and injury resistance, while having little influence on player rating or market value.

**Table B.1.**



**Part D - How will things be in 2021**

**Methodology**

*Objective*

The objective here is to predict the 2021 skills values. This is a regression task.

*Response variable selection*

We defined as skills values a single synthetic indicator (skill\_summary), defined as the simple average of 5 variables: dribbling, curve, free-kick accuracy, long passing, ball control.

*Data preparation*

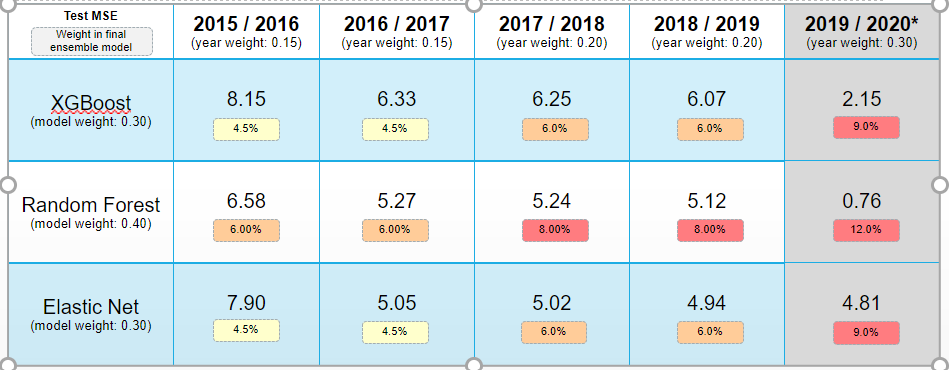
To add

*Modelling, training and evaluation*

Three different classes of models were used: XGBoost, random forest and ElasticNet. In order to take advantage of the multi-year dataset, we trained each of these three models on 5 pairs of years. Each model was then evaluated on its MSE on 2019-2020, used as validation data.

The final prediction of skill-summary for 2021 is made by aggregating these 15 models (Table D.1). Observing that XGBoost performs better, the 5 XGBoost models were given a collective weight of 0.40. The random forest and Elastic Net model were both given a total weight of 0.30. Each year was also weighted differently, giving recent years more weight in the aggregation process. This is based on the hypothesis that evolution pattern may evolve over time.

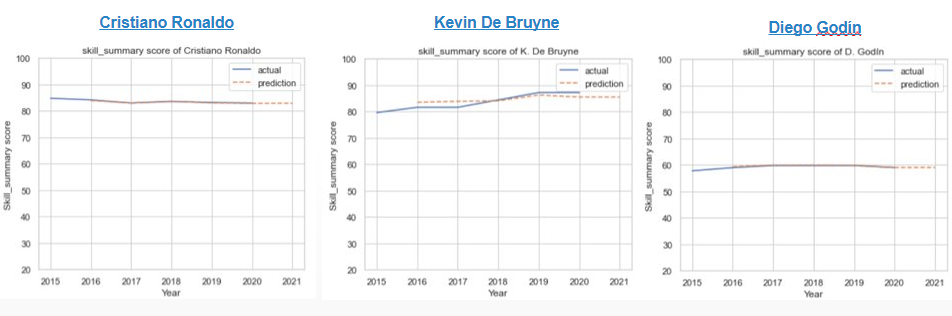
**Table D.1.**



**Selected predictions**

As the data for 2021 is not available, it is not possible to evaluate the prediction performance on a test set. However, we can examine predictions for selected players. Our model forecasts a slight decrease in Cristiano Ronaldo’s skill summary score. It predicts a stable score for Diego Godin. Results for Kevin De Bruyne are interesting, as they show a larger prediction error in previous years.

**Figure D.1.**

**

**Conclusion – Caveats and possible future improvements**

**Main caveats**

There are two main limitations in our results. First, some of our approaches were computationally intensive. Fine-tuning hyperparameters in several dimensions, even using relatively computationally efficient approaches such as XGBooset, is not immediate.

Second and most importantly, there were a number of ambiguities in the definition of the tasks that we solved as a team. In a professional setting, it would be important that the choices are made with the end user of the results and not by the modelling team on its own.

**Possible future improvements**

A number of improvements and extensions to the work presented could be made. They include exploring additional model classes, more rigorous consideration of the time series properties in the data and possibly an extension to real-life football statistics.

Exploring additional model classes could contribute to improving further the quality of the predictions. Given the reasonable size of the dataset, support vector machines could constitute a tractable option. Neural networks approaches could also be used and may prove especially useful for the difficult classification problem tackled in Part B, where a certain overlap in classes seems to limit current approaches.

It may also prove possible to more rigorously take into account the time series properties in the data. While this is to some extent a theoretical question, it may also enable better feature engineering and lead to better prediction performance.

Going beyond in-universe game data, one could also consider extending the approach to real-life player data. This is likely to require a number of adaptations, as a large number of variables would need to be operationalized in a way that could be observed in games.

**Annex – Additional visuals and tables**

1. A player’s improvement between Year N and Year N+1 is attributed to the Year N club. [↑](#footnote-ref-2)