CS 109 – Fall 2020

Final Project

EA Sports – FIFA

Milestone 3

**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 observation, 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 harmonised 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:

* Creation of a year variable and merger of the 6 datasets
* Creation of dummy variables for each player trait and each player tag
* 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

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

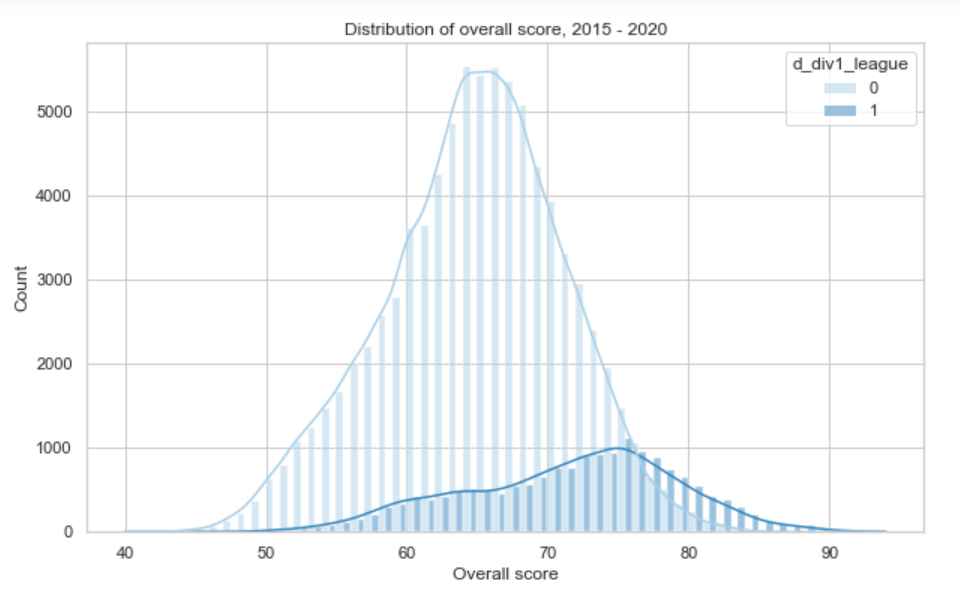
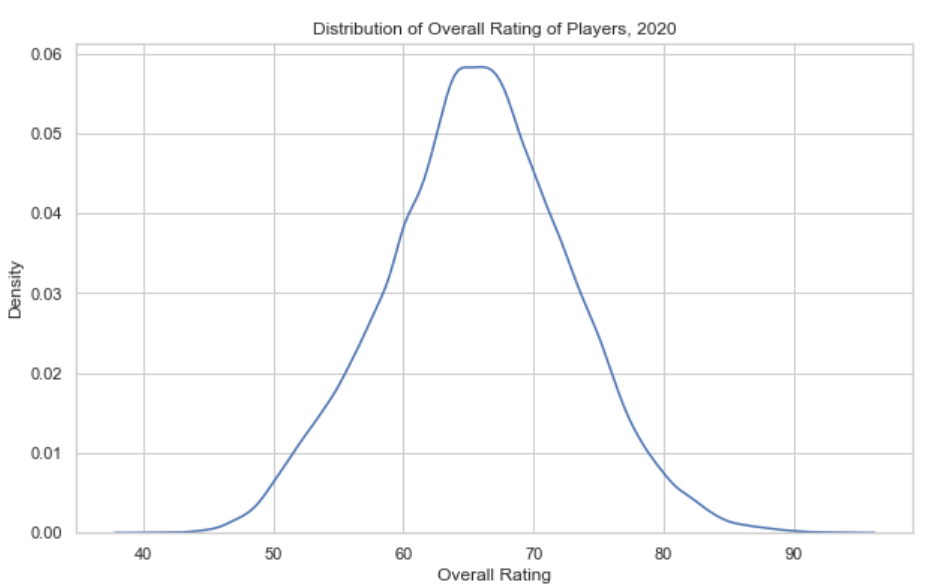
While some limited data imputation was carried out, more remains to be done to improve the quality of predictions.

**Exploratory data analysis – Key findings and visualizations**

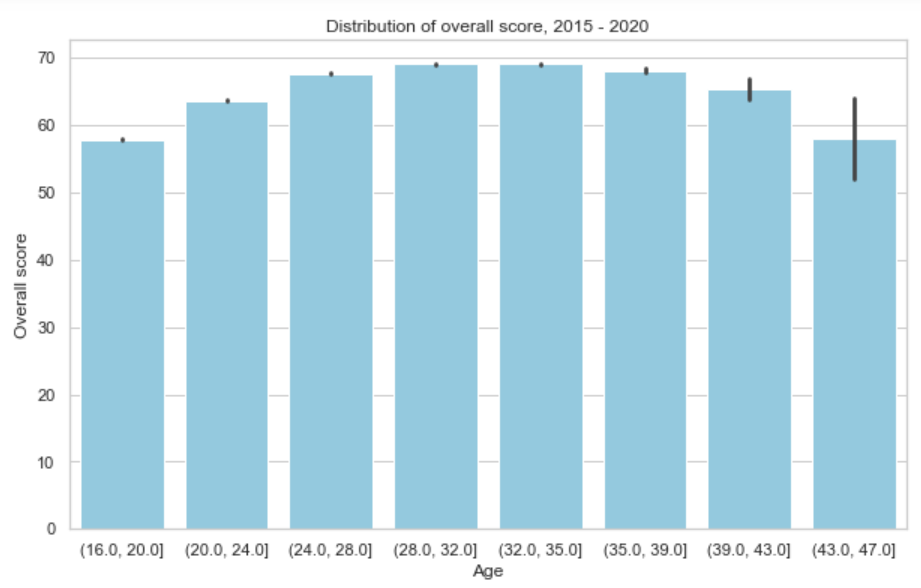
The EDA was mainly carried out with Part A in mind (please see below).

The most salient facts from the EDA include:

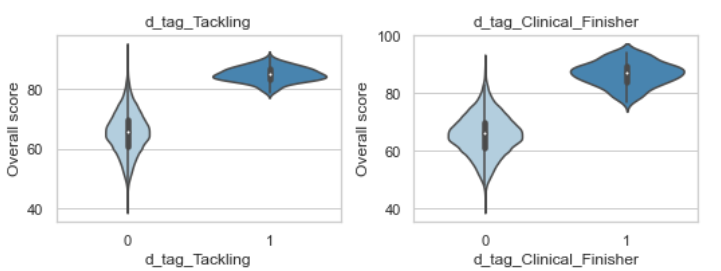
* **The distribution of OVR scores seems almost normal at first sight**. In fact, it is normal for players who do not play in Division 1 Leagues, but seems right-skewed for players of Division 1 Leagues. The modal OVR score for a Division 1 player is about 75, and about 65 for a player which does not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



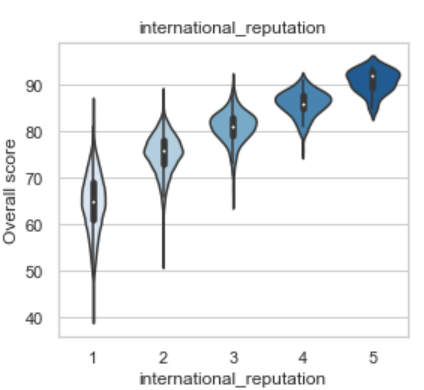
* There seems to be **a quadratic relationship between age and OVR.**



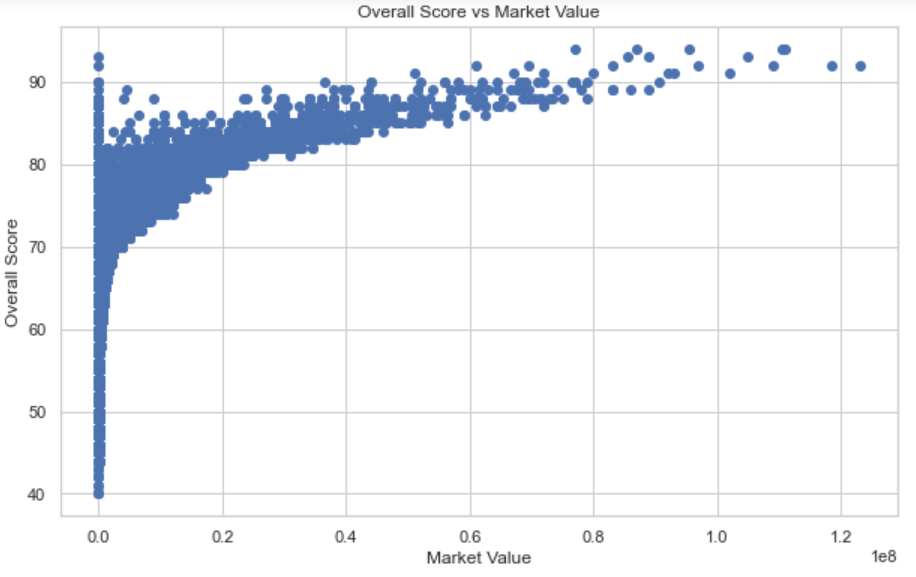
* Certain player tags seem associated with outstanding OVR values, including “tackling”, “clinical finisher” and several others. All tags can be considered positive, as none is negatively associated with a lower OVR. Player traits seem much less associated with OVR. A notable exception is “1 on 1 rush”.



* 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



**Revised project question(s) based on EDA**

The project structure was already determined in the project description. As a result, we will mainly focus on answering the four key questions as is they were formulated by our client.

* Part A - Rank the Players

The objective here is to predict the OVR statistic for 2020 based on 2019 data. One key observation is that we will evaluate the performance of the model on players of 7 top European clubs only, which is a very special set.

* Part B - Classify player position

The objective is to predict player positions based on other statistics, training on 2019 data and testing on a subset of 2020 data. One difficulty here will be the large number of different positions. We will also remain mindful of the peculiarities of the test set.

* Part C - Which Club has the Best Staff

The objective here is to infer staff quality on the basis of player improvement while at the club. It will be important to control for a number of variables when doing so, notably player age (possibly with a quadratic term).

* Part D - How will things be in 2021

The objective here is to predict the 2021 skills values. One key dimension to keep in mind here is that the evaluation will not be made against actual values, but against the average of values predicted by other groups.

Time allowing, we will try to focus also on one of the questions that we formulated ourselves as part of Milestone 2.

**Baseline model**

A number of models were already used, including a LASSO regression, a random forest and a XGBoost model. Their respective performance on the test set was:

|  |  |
| --- | --- |
| **Model** | **MSE on test set** |
| LASSO | 17.89 |
| Random Forest | 5.36 |
| XGBoost | 5.85 |

**Next steps**

More work remains to be done on variable imputation, feature engineering, model tuning for part A, as well as a model development for parts B, C and D.