Determining When a Soccer Player Will Reach Their Peak

Employing several machine learning models to predict when a player's skills will start to decline

Authors: Divya Kumari, Fernando Arancibia, & Peter Mullen

The dataset used in this post can be accessed on Kaggle - European Soccer Database

Image Source: https://www.desertsun.com/picture-gallery/sports/2017/02/08/75-the-minimum-age-for-this-soccer-match/97661350/

Soccer (or Football if you prefer) is the world's most popular sport and, as a result, is big business. Winning is critical to attracting fans and financially supporting a club. To make their team better, the management of a club will often purchase players from other teams during specified windows during a year, called transfer windows. In recent years, the transfer fees paid by one club to another for a player have continued to grow, making multi-million dollar transfers commonplace. The high cost of player transfers makes the decision to execute a transfer highly risky. A series of bad transfers can create issues for a club. For example, bad transfers by Barcelona have been a major contributor to their recent financial instability[1].

You can think of a player as an asset for a club. If a player is purchased for $50M and performs at a high level, the club will likely benefit from improved on-field performance and value generation from that success. Additionally, a player's financial value will either remain consistent or grow. However, if a player performs poorly, they likely are not generating value and their

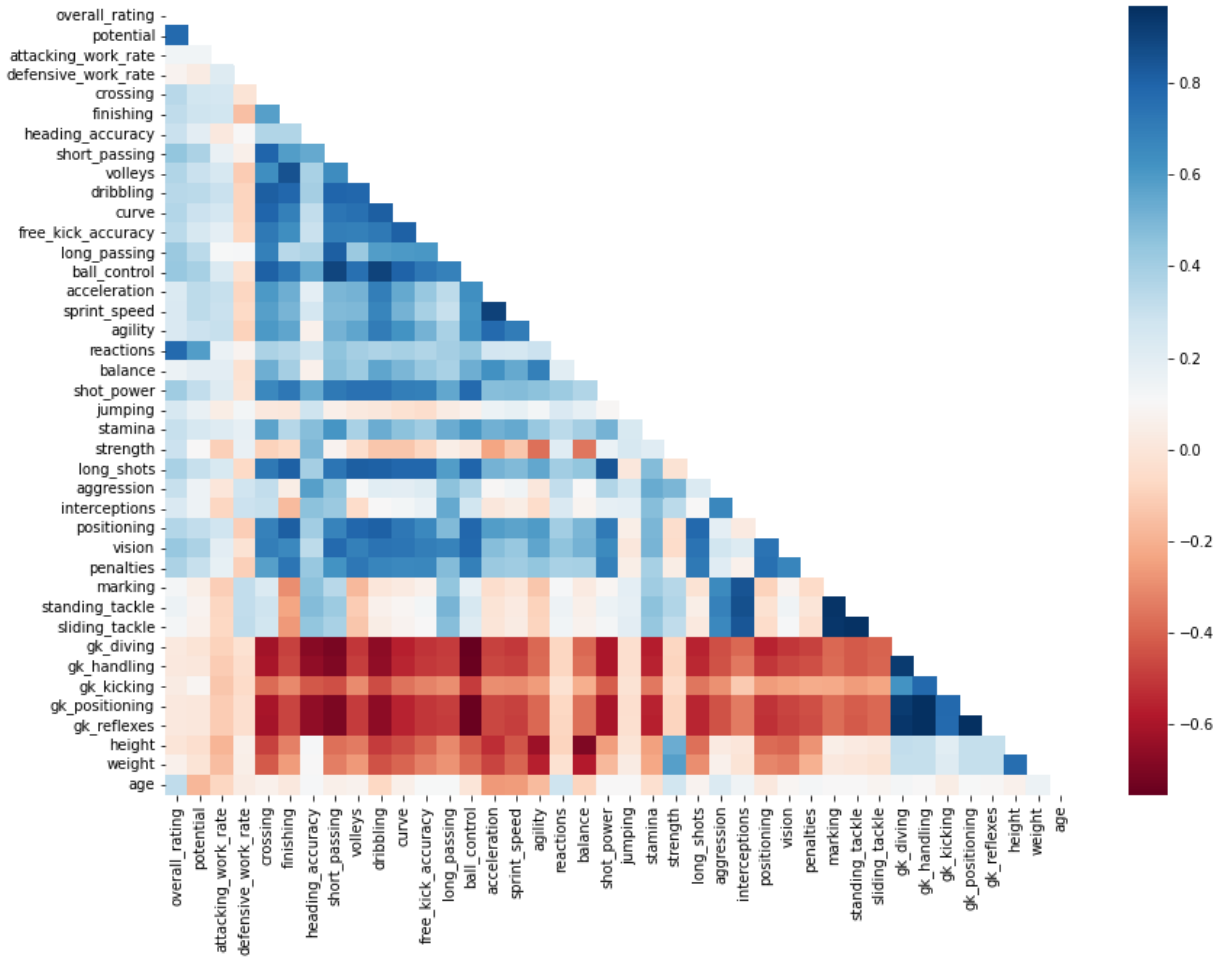
[1] https://theathletic.com/3468740/2022/08/03/barcelona-money-finances-crisis/

value as an asset to the club will decrease. Therefore, understanding when a player's abilities are going to start declining is a critical consideration of a transfer. Knowing if an asset is going to decline in the near or long term can drastically change negotiations and decisions.

With this in mind, we decided to explore how to predict whether a player is going to decline in the next year. To do this, we leveraged data from FIFA by EA Sports between the 2008 and 2016 seasons. The data from this game overviews players' attributes–both skill-based and physical–creating a valuable longitudinal dataset to understand the relationship between a player's rating over time and their attributes. Ultimately, we were able to create a model from this data that predicts whether a player is going to start declining based on their age and attributes.

Initial Data Exploration

The FIFA data provides a wide range of attributes, as well as an overall rating that generally represents a player's ability. In 2016, the three highest-rated players were Lionel Messi (94), Cristiano Ronaldo (93), and Neymar (90)[2]. With this data (after some basic cleanup), we started by wanting to understand the correlation between the various attributes and created a heat map:

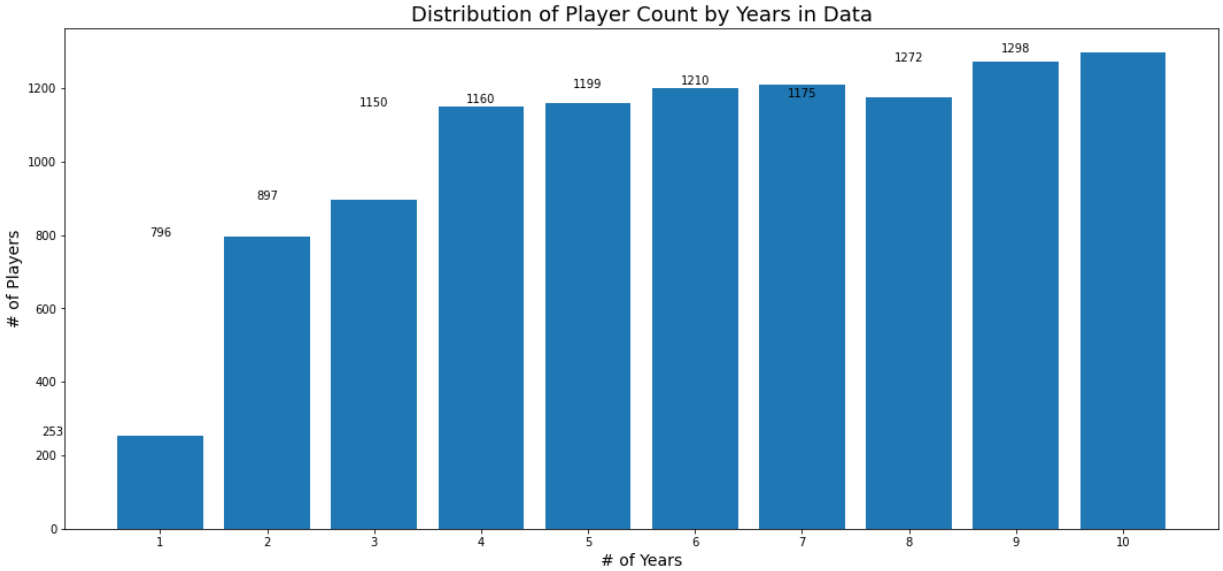


[2] https://www.fifaindex.com/players/fifa16/

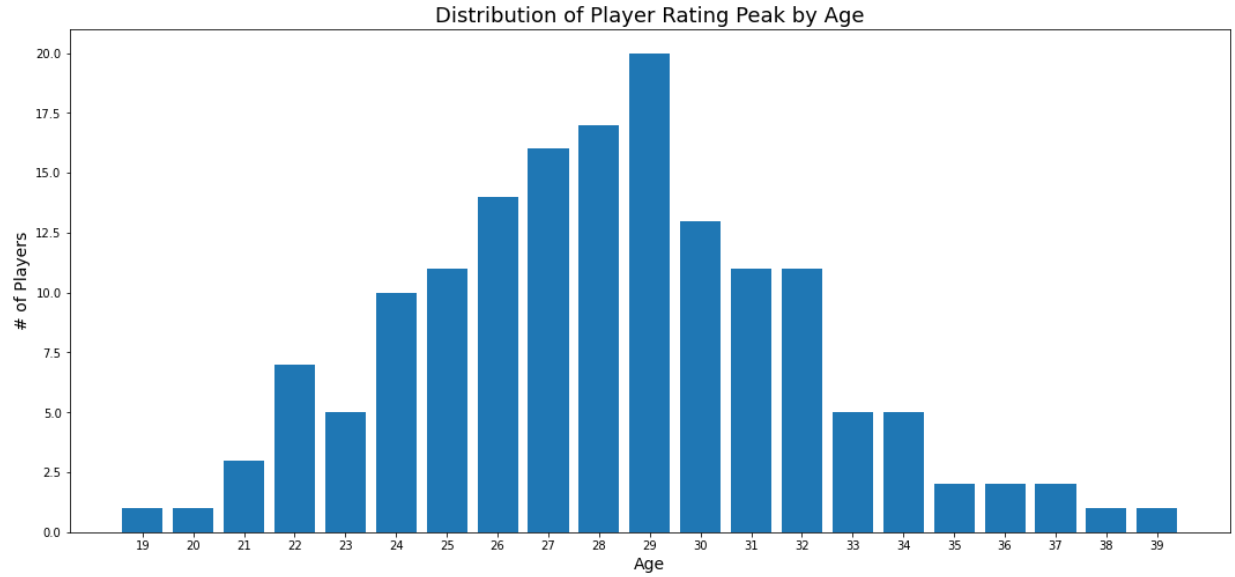
We plotted the matrix using the ‘RdBu’ color scheme so that positive correlations stand out in blue, negative correlations in red and unrelated features in white.

What stood out immediately was the group of *gk attributes*, which were highly uncorrelated with other attributes. Additionally, there were pockets of other attributes that were highly correlated, such as ‘sprint speed’ with ‘acceleration’ and ‘standing tackle’ with ‘sliding tackle’. These correlations made sense since these are related skills. We then went on to separate goalkeeping skills from player skills so our machine-learning models can predict accurately.

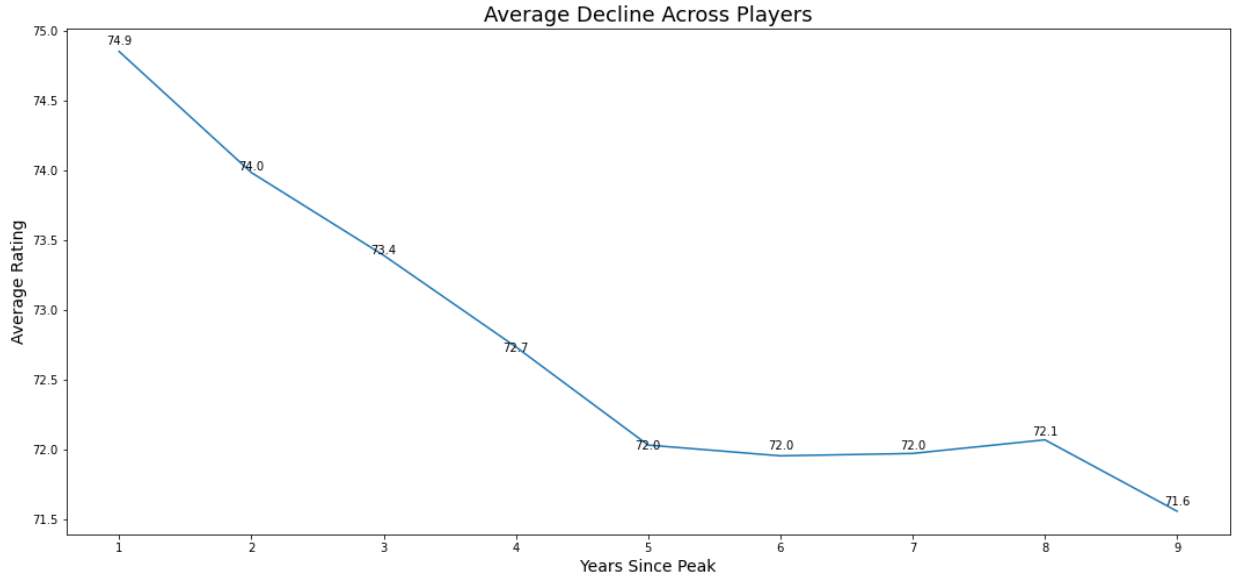
The next item we wanted to understand was how many years we have for each player in the dataset across the 10 years. We plotted this distribution and found that we had many players with multiple years, which we could use to observe the pattern of decline.



The group with ten years of data piqued our interest, so we decided to investigate to understand the patterns in this group. We proceeded to identify the “peak” for these players, defined as the highest overall rating they received. If a player received that same rating in multiple years, we determined the final instance to be their peak and all dates after as their “decline”. We plotted out the age of peak for players with 10 years of data and found a reasonably normal distribution centered around age 29:



Finally, we were curious about what “decline” looked like on average across all of these players. We calculated and observed the average rating of all players across their decline:

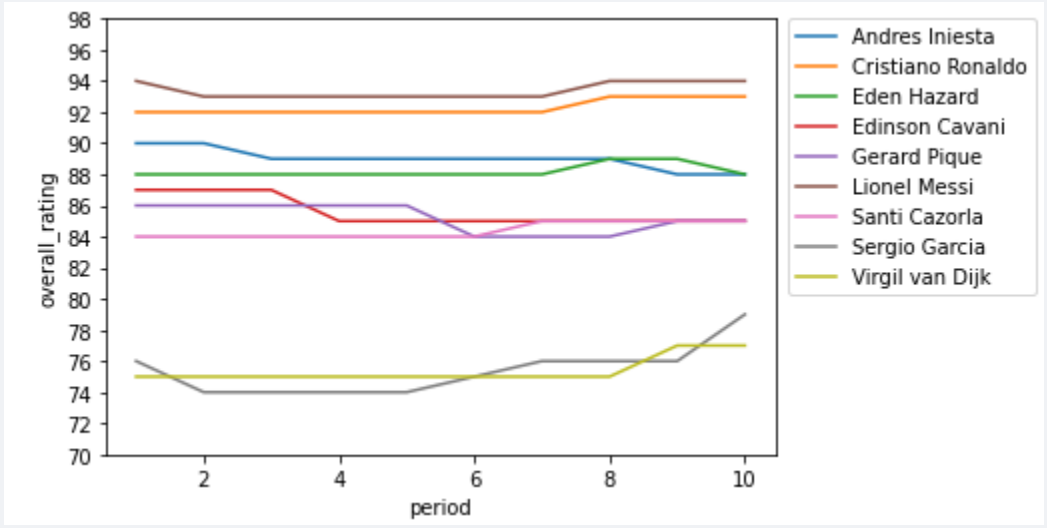


This dramatic drop-off and following flattening is interesting. We hypothesized that this leveling may be a result of two causes: (1) players will typically retire before they can drop off dramatically or (2) the data may be skewing upwards as players drop out in later years. Ultimately, this pattern demonstrates broadly that we are able to identify a decline across players.

Dataset Preparation

From our initial analysis we decided to make a few changes to our dataset. The first change was to limit to players with at least 3 years of data in order to ensure we could have some information about whether they were RISING or DECLINING. The next major step was to identify a player's peak and create a new column. In order to have a binary category, we decided to include a player's PEAK with their RISE, creating a new 'peak flag' indicated as either RISE or DECLINE. This classification will be used to train our model. We additionally separated out goalkeeper-related attributes since players' performance will likely be low on all of these.

We then selected a small group of well-known players to visualize their evolution more closely. We selected a heterogeneous set of players in terms of their positions in the field and their career stages. We included their ratings over the previous ten periods. This group included: Lionel Messi, Santi Cazorla, Sergio Garcia, Gerard Pique, Cristiano Ronaldo, Edinson Cavani, Eden Hazard, Virgil van Dijk, and Andres Iniesta. We plotted the overall ratings for these players to visualize their current trends.



Now that we have a usable dataset, we can proceed to make a prediction to answer the question: **Will a player's overall rating be in decline over the next period?**

We will build different Binary Classification models and then compare their performance. Finally, we tested our best model on the players(9) subset we showed earlier, in order to predict how their career stage will look like over the next period.

Modeling

In order to perform Binary Classification, we implemented 3 different models: Logistic Regression, SVM's LinearSVC, and Random Forest. We will briefly describe them and compare their performance.

Logistic Regression Classifier

Logistic regression is a supervised classification algorithm. In a classification problem, the dependent variable, y, can take only discrete values for a given set of independent variables, X. The model builds a regression model to predict the probability that a given data entry belongs to the category numbered "1". Logistic regression models the data using the sigmoid function.

We used the following hyperparameters: `max_iter` equal to 100, and `'liblinear'` as the solver method. We got an accuracy of 77% for both training and testing sets.

SVM's LinearSVC

The Linear Support Vector Classifier (SVC) method applies a linear kernel function to perform classification and it performs well with a large number of samples. The algorithm aims to find a hyperplane to maximize the distance between classified samples. Linear SVC has additional parameters (compared to the SVC model) such as penalty normalization and loss function.

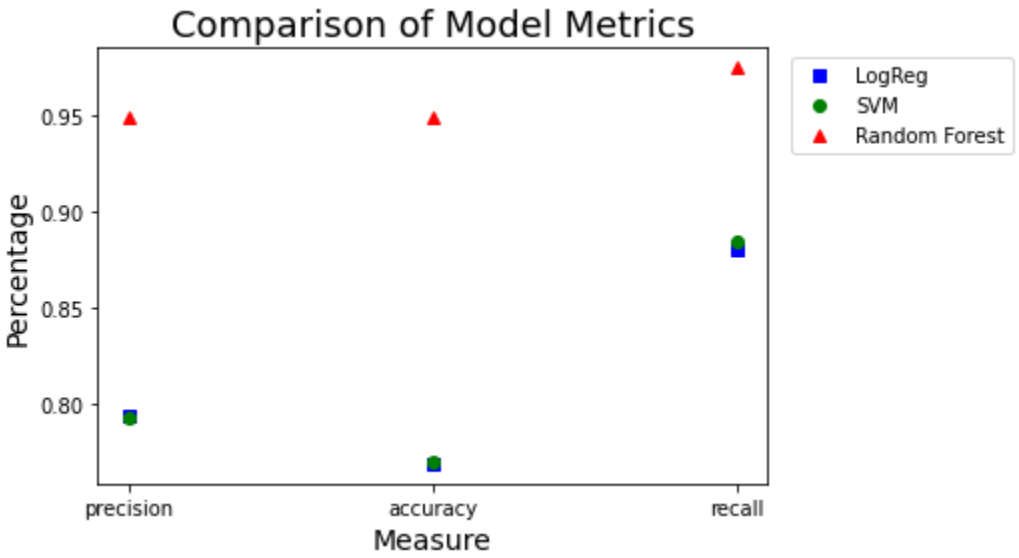
We used the default parameters: `penalty='l2'`, `loss='squared_hinge'`, and we adjusted the tolerance to `tol=1e-5`. Similar to our Logistic Regression model, we got an accuracy of 77% for both training and testing sets.

Random Forest Classifier

Random Forests are a superior option to linear or logistic models for many reasons. Firstly, Random Forests can incorporate non-linear effects and are preferred to alternate methods for modeling complex interactions when the interactions cannot be pre-specified. Additionally, this type of model inherently provides a much greater predictive ability for overfitting data, making them suitable for large datasets. Given the above advantages over existing methods and that Random Forest has been used previously to classify player performance, we chose this model for classification on our dataset.

After declaring the required libraries, we applied RandomForestClassifier on the dataset with the hyperparameters `max_depth` and `random_seed`. `max_depth` is the number of splits that each decision tree in the forest is allowed to make. It helps reduce overfitting if we limit the number of splits after a particular threshold. Moving on, we fit the data on the training set and then predicted on our training and testing set respectively. Our training and testing set received an accuracy of 99% and 94% respectively. Since the training set accuracy is not drastically larger than our testing set's accuracy, we assume our model is not overfitting.

We observe that our Random Forest Classifier obtained the highest accuracy compared to the 2 previous models. We plotted a graph to compare our models against each other in terms of 3 different measures. As expected, Random Forest is superior not only in terms of accuracy but also in terms of precision and recall.



We finally selected Random Forest as our Classification model.

Our Model in action

Now that we have selected our final model, it is ready to be used on our test dataset of well-known players! We will predict whether the player skills will be in the RISE or DECLINE phase over the next period. The following table indicates the player's age, his current phase (RISE/DECLINE), his current overall score, and his potential overall score, plus one extra column indicating our model's prediction for the next period:

Player name	Age	Current period phase	Overall score	Potential overall score	Next period phase *(prediction)*
Andres Iniesta	31	DECLINE	88	88	DECLINE
Cristiano Ronaldo	30	RISE	93	93	RISE
Eden Hazard	25	DECLINE	88	90	RISE
Edinson Cavani	29	DECLINE	85	85	DECLINE
Gerard Pique	28	DECLINE	85	85	DECLINE
Lionel Messi	28	RISE	94	94	RISE
Santi Cazorla	32	RISE	85	85	DECLINE
Sergio Garcia	32	RISE	79	79	DECLINE
Virgil van Dijk	25	RISE	77	83	RISE

Let's take a closer look at some notable cases. Eden Hazard (25 y.o.) was recently on a declining curve but our model predicts he will break this trend next season and start improving his skills again. This makes a lot of sense as his Potential Rating (given by FIFA) is greater than his current Rating and -at this point- he was only 25 years old.

Another notable case is Virgil Van Dijk (25 y.o.), whose Potential Rating is greater than his current Rating, and whose skills started to rise over the last 2 periods. We can observe that our model predicts that he will continue in the RISE stage over the next period.

Finally, we have some different cases such as Santi Cazorla (32 y.o.) and Sergio Garcia (32 y.o.). They are currently in their RISE stage (as we can see in the historic overall score chart), but our model predicts that their curve will change their slopes to DECLINE. For those cases, our model is especially useful, since it can anticipate an early change in the player's skills trend. Thus, soccer scouts studying the acquisition of those players -at that time- should be very cautious about the potential signings!

Conclusion

In this article, we saw how machine learning systems can be leveraged to predict a sportsperson's performance which is influenced by multiple factors and is a complex problem to solve in real-life.

From choosing the right dataset to choosing the right machine learning model, we were heavily influenced by the best practices in big data for carrying out operations. We acquired our dataset from Kaggle, performed data wrangling, and checked for multicollinearity. After making sure we had structured data to work with for our Machine Learning (ML) models, we started our process of classification. We went through multiple ML models, determining along the way what worked and what didn't. We ended up reaching our goal with Random Forest Classifier which predicted, given a particular player and his age and skills, if they would be able to keep up/better their performance or decline.

This project can serve as a keystone idea for using ML models in the Sports arena. Not only the player's performance but various other field dynamics can be predicted this way. By gaining knowledge on the same, we can make interventions to guide us to obtain the desired outcome on the field.

Key Challenges of our Analysis & Next Steps

In our analysis, we encountered several challenges that can be addressed in the future. The core obstacles include:

- Timeframes limitations to our dataset. With only 2008 to 2016 included in our dataset, we were unable to track the complete careers of players to get a complete picture of a player's rise and decline.
- There are confounding variables to our analysis since a player's rise and decline are not always straightforward. A change of club or coach can have a dramatic effect on an overall rating. Due to our limited dataset, we were unable to fully account for these outside variables.
- Measuring a player's performance and contribution to a team is complex. Each player instance had 30 features, all of which were critical to understanding the overall rating. As a result, we were unable to perform feature reduction, since removing any feature could impact the accuracy of our predictive models.

With these challenges in mind, we suggest the following next steps to build upon our work:

- Include the changes in the skills as new features. E.g. `overall_score_diff_1` may represent the difference in overall score between the last period and the current period. That would be helpful for including past information as additional input. The drawback would be including so many new columns (it would be multiplicative in terms of the

number of periods we would like to add). In order to mitigate this effect, we could only include changes for a subset of the skills.

- Expand the dataset (past and future) to get the full careers of players. Doing additional data capture will allow us to have a fuller understanding of how players' skills develop over time. An additional approach would be to focus on selecting players who have retired and acquire data for them.
- Predict a date in time in which the peak will be reached. That would be a natural additional step to predicting the change in the slope (RISE or DECLINE).
- Quantify the decline or predict the age at which it will occur. We solved our problem with classification but we can also quantify our player's performance (currently categorizing) and use regression models.