Football Player Performance Rating Prediction based on Boosting Algorithms

Aditya Sarkar

Data Science For Sports – CS-IS - 3023

Computer Science, Ashoka University, ASP-23

Code Notebook

***Abstract* – This paper investigates the accuracy of different variations of Boosting models used to predict the Football Player Performance Ratings given by real-life football scouts along with various performance attributes in the form of physical, kinematical and tactical variables. The paper also analyses the uses of this method of rating prediction and its scope in the football industry.**

1 INTRODUCTION

The first evidence of analytics in sports is widely considered to be the brainchild of Henry Chadwick.[1] in 1859, he formulated the "box score"- basically a database where he recorded statistics like runs, hits, put outs, assists and errors for the Baseball team Brooklyn Excelsiors club. This paved a way for the people to not only study a game in great detail, but also compare two players' performance. The next big revolution in sports analytics came in 1977 when Bill James compiled a collection of Annual Baseball data on which he would develop analytical abstracts. He coined the term 'Sabermetrics' which he defined as the empirical analysis of baseball using statistics [2].

Technology has progressed since and like every other field, sports industry has taken full advantage. Data helps teams and organizations to track performance, make predictions and make smarter decisions both on and off the field. [3]. From opposition analysis, player recruitment, tactical analysis, performance analysis to various other spheres of numerous sports, big data has allowed players, coaches and organizations to improve their game and has also enhanced the quality of commentary and analysis for the viewers and fans.

One of the most important factors that leads a football club or rather any sports team to success is their recruitment. Recruitment in football clubs before the last few years were mostly done based on the initial reports given by football scouts or scouting agencies. A football scout would know what kind of a player the specific team and coaching staff wants. These agencies keep a track of a bunch of players depending on their scouting range by physically watching their games and analyzing their performances. Even though the availability of stats and data was not an issue, coaches and football clubs still used to depend on the naked eye of an experienced and reputed scout.

In Football, there are various problems which are being attempted to solve using data and analytics starting from comparing suitable players, analyzing performances, tactical analysis amongst a few.

The development in computational power and measurement technology has enabled the generation of data on the movements in various sports games for use in planning and evaluating the tactics and strategy. Companies like Focus Motion have made good use of the development of Micro-electromechanical systems (MEMS) which has reduced drastically in size, cost, and power consumption, while improving accuracy. The combination of different sensor technologies is considered a promising step in the monitoring of athletes. Those “wearables” enable the capturing of relevant physiological and tactical information in individual and team sports and thus replacing subjective, time-consuming and qualitative methods with objective, quantitative ones [4]. These devices generate precise physical and kinematical variables which is used by the data analysts in the football clubs to produce important observations to the coaching team. Thanks to the availability of massive data capturing all the events generated during a match, The problem of evaluating the performance of soccer players is attracting the interest of many companies and football clubs.

This paper attempts to make use of these physical, kinematical and tactical variables and use these attributes to predict the scout-made evaluation of players in the context of given matches using Ensemble Machine Learning Models, specifically some variations of boosting models like XGBoost, CatBoost and LightGBM. This paper tries to investigate the accuracy of the results along with the actual scout player rating evaluated by watching the player physically and also

discuss and predict the use of potential advantages and disadvantages of this method and its future in the footballing industry.

2 DATASET

The dataset used is a large scouting dataset containing the players, teams, and opponent's performance in terms of physical, kinematical, technical, and tactical variables.

In this dataset, each row corresponds row of the dataset corresponds to the evaluation of a player in a given match together with variables related to the player, match, team, and opponent and the rating of a player in a match given by a scout.

Every row of the dataset corresponds to the evaluation of a player in a given match together with variables related to the player, match, team, and opponent and the rating of a player in a match given by a scout.

Variables that encode player properties or performance are grouped according to the following scheme:

`player(offensive/defensive/positional/physical/general/other)(derived/raw/ratio)var#number`

- **player** shows if the variable refers to a player property
- (offensive/defensive/positional/physical/general/other) refers to the nature of the variable -(derived/raw/ratio):
- raw: variable empirically measured
- derived: variable defined based on two or more variables and not empirically measured
- ratio: variable defined as a ratio of two raw variables

Variables that encode team properties or performance are grouped according to the following scheme:

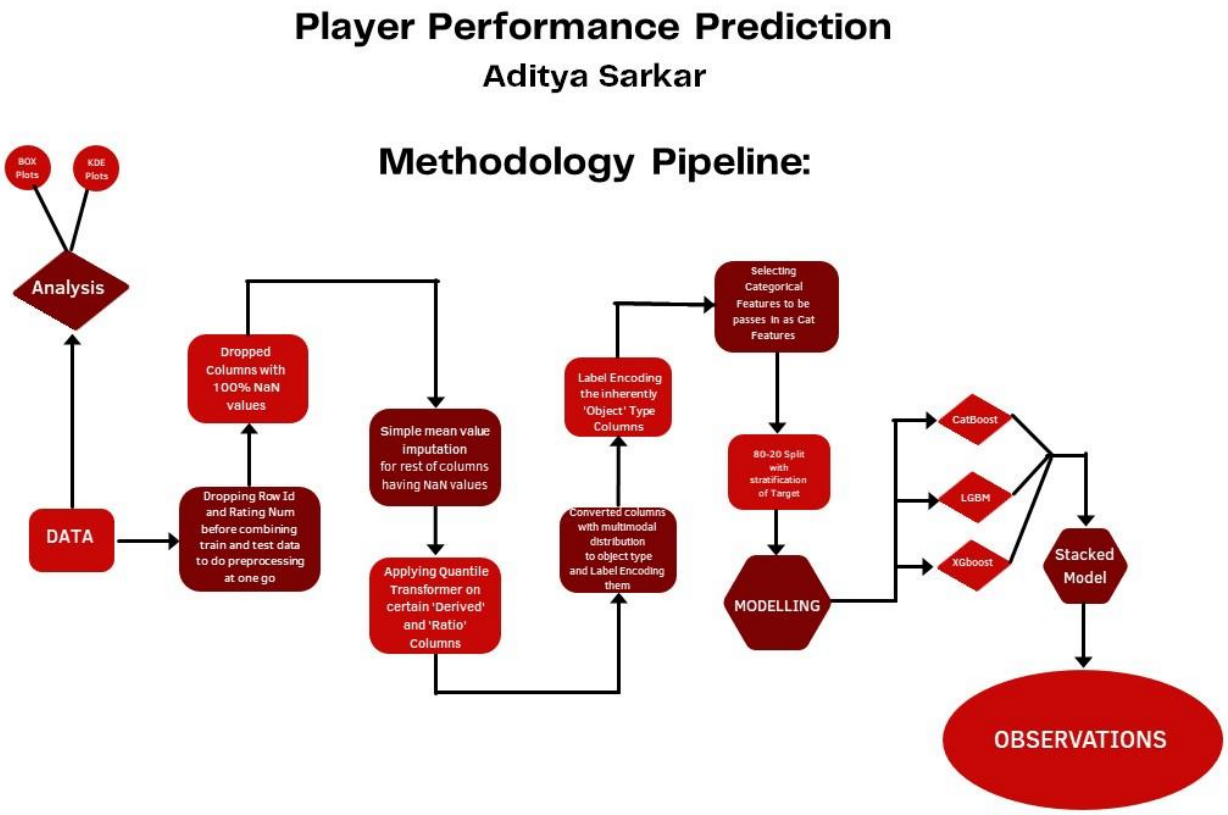
`team(1/2)(offensive/defensive/other)(derived/raw/ratio)var#number`

- team(1/2) shows if the variable refers to a team1 or team2 property

- (offensive/defensive/other) refers to the nature of the variable
- (derived/raw/ratio):
- raw: variable empirically measured
- derived: variable defined based on two or more variables and not empirically measured
- ratio: variable defined as a ratio of two raw variables

3 METHADOLOGY PIPELINE

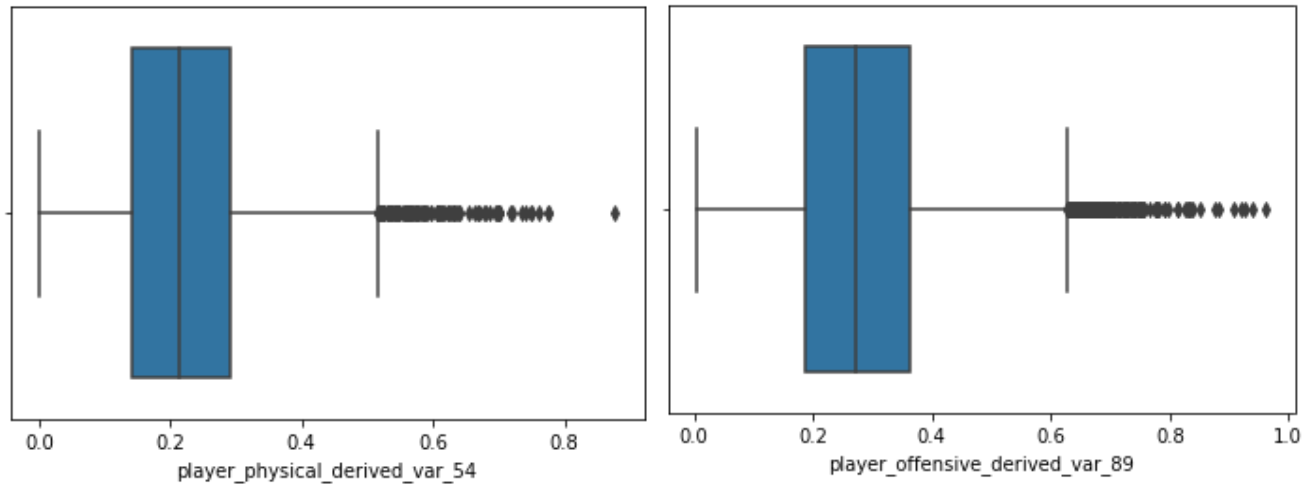
The methodology pipeline can be essentially broken down with the help of the pipeline flowchart below.



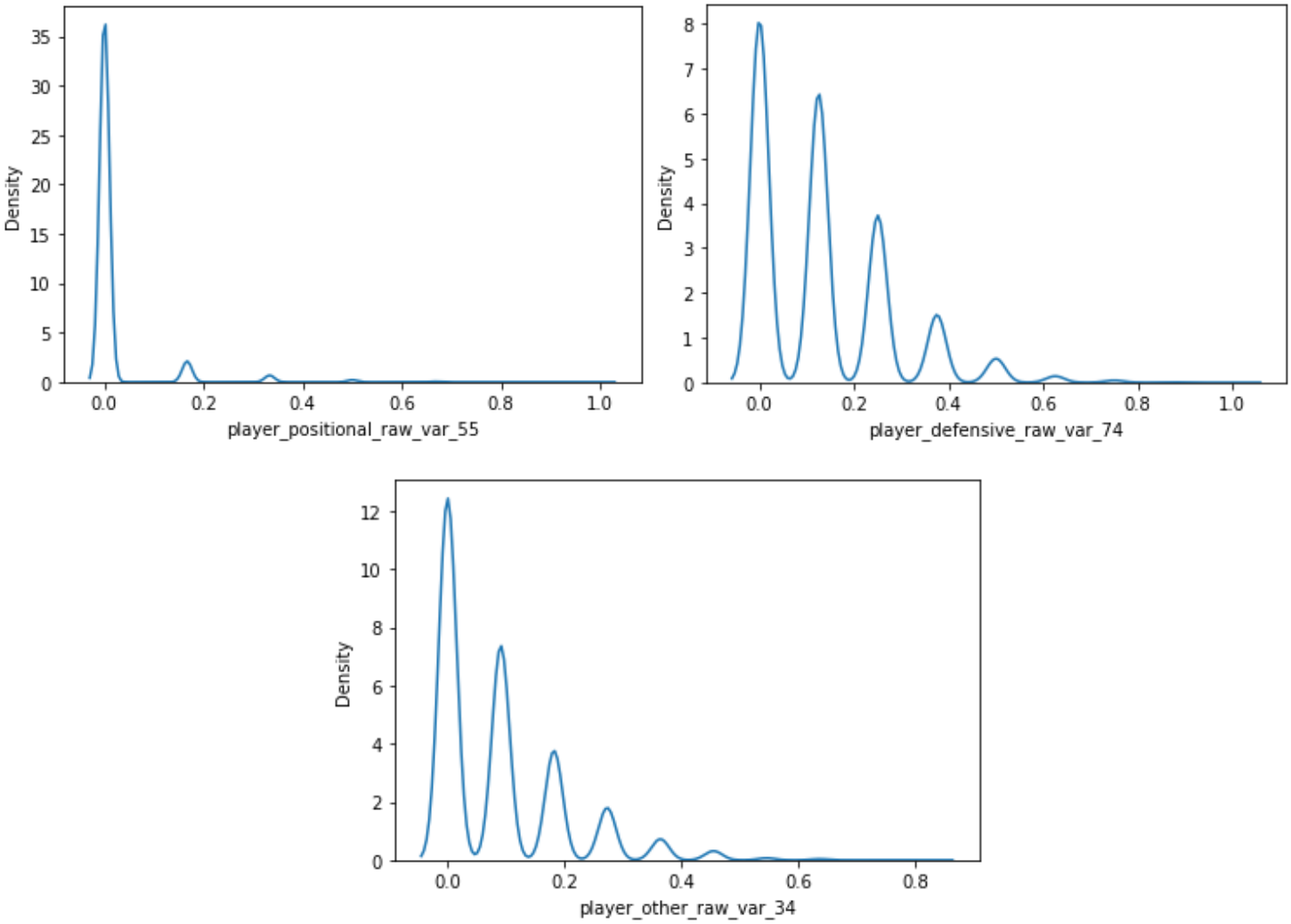
A Dataset Analysis

The dataset has a whopping 799 columns and 8774 rows with each row representing to a player's rating and statistical variables. However, around 50 columns have mostly null values in them. There are also a few columns having 100% null values.

With the help of **Box Plots**, we can see that a significant number of columns have a **skewed distribution**, especially the **derived and ratio columns**. Below are a few examples of those Box Plots



KDE Plots show that a lot of **raw variable** columns are not continuous and are rather **multimodal**. Below are a few examples of those KDE plots.



B Data Pre-Processing

Feature Engineering is a key step Generally, performance of a machine learning model relies on the quality of the feature set. Thus, irrelevant features may produce less accurate results. Extraction of discriminating features becomes crucial for better performance. Below are the pre-processing steps taken to enhance feature extraction.

Row ID and Rating_Num columns were removed before combining both the training data and the test data.

All columns with 100% NULL values were dropped from the dataset. For the rest of the columns with some NULL values, statistical **mean imputation** was used to replace the NULL values with the mean of that column. This technique helps in filling the empty data by preserving the mean of the data.

For the derived and ratio columns having a skewed distribution, **Quantile Transformation** was performed to transform the input variable having skewed distribution to have a Gaussian distribution, to be used as input to a predictive model. Many machine learning algorithms prefer or perform better when numerical input variables have a standard probability distribution, such as a Gaussian or a uniform distribution. [5]

In some specific raw variable columns having a multimodal distribution and other *object* type columns, **label encoding** was performed to convert the variables into model-understandable numerical data. Along with these, the *winner* and *team* column having [“winner, draw, loser”] variables and [“team1 and team2”] variables were converted to [1,0,-1] and [0,1] respectively.

Categorical Attributes chosen are `['scout_id','winner','team','team1_system_id','team2_system_id', 'competitionId','player_position_1','player_position_2']` along with the multimodal variables which are label encoded.

C Modelling

Boosting belongs to the class of ensemble learning methods. This method combines a set of weak learners into a strong learner sequentially to minimize training errors. In boosting, a random sample of data is selected from the training data and then fitted into the model. Then the second model is built which tries to correct the errors present in the first model. This procedure is continued and models are added until either the complete training data set is predicted correctly or the maximum number of models are added. With each iteration, the weak rules from each individual classifier are combined to form one, strong prediction rule along with reducing bias. [6].

One of the major types of Boosting is **Gradient Boosting**. It relies on the principle of Boosting that the best possible next model, when combined with previous models, minimizes the overall prediction error. **Gradient boosting** trains on the residual errors of the previous predictor. The name, gradient boosting, is used since it combines the gradient descent algorithm and boosting method. [6].

For this dataset, multiple variations of Gradient Boosting Methods have been used in an attempt to get more accurate results. The following are the models used:

- **CatBoost**: CatBoost is an open-source algorithm for gradient boosting on decision trees. It is developed by Yandex researchers and engineers, and is used for search, recommendation systems, personal assistant, self-driving cars, weather prediction and many other tasks. https://catboost.ai/.
- **XgBoost**: XgBoost stands for Extreme Gradient Boosting, which was proposed by the researchers at the University of Washington. It is a library written in C++ which optimizes the training for Gradient Boosting. https://github.com/dmlc/xgboost.
- **LightGBM**: LightGBM stands for Light Gradient Boosting Machine. It is a free and open source distributed gradient boosting framework for machine learning originally developed by Microsoft. It is based on decision tree algorithms and used for ranking, classification and other machine learning tasks. The development focus is on performance and scalability. https://github.com/microsoft/LightGBM.

4 OBSERVATIONS AND SCOPES OF IMPROVEMENT

The following are the specifications of the three different models for which the best R2 score and RMSE was observed.

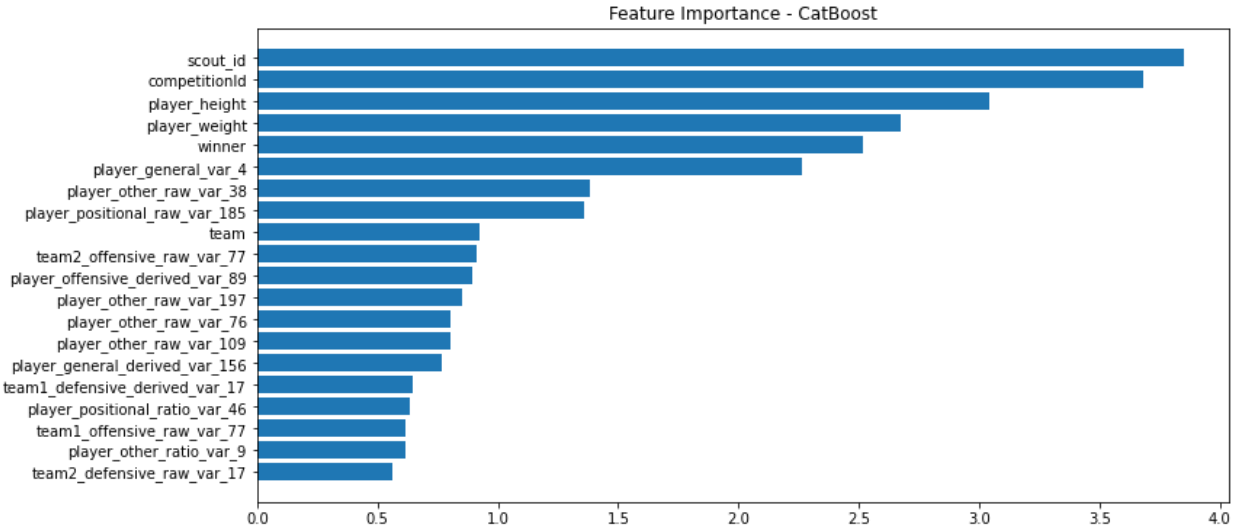
CatBoost: (with specified categorical features and default params)

Training Score	0.7296570243815041
Test Score	0.3672298937536369
RMSE	1.4597835310056915

CatBoost: (without specified categorical features and default params)

Training Score	0.6746432668659422
Test Score	0.33527824352407953
RMSE	1.4961854589252572

Feature Importance Graph:



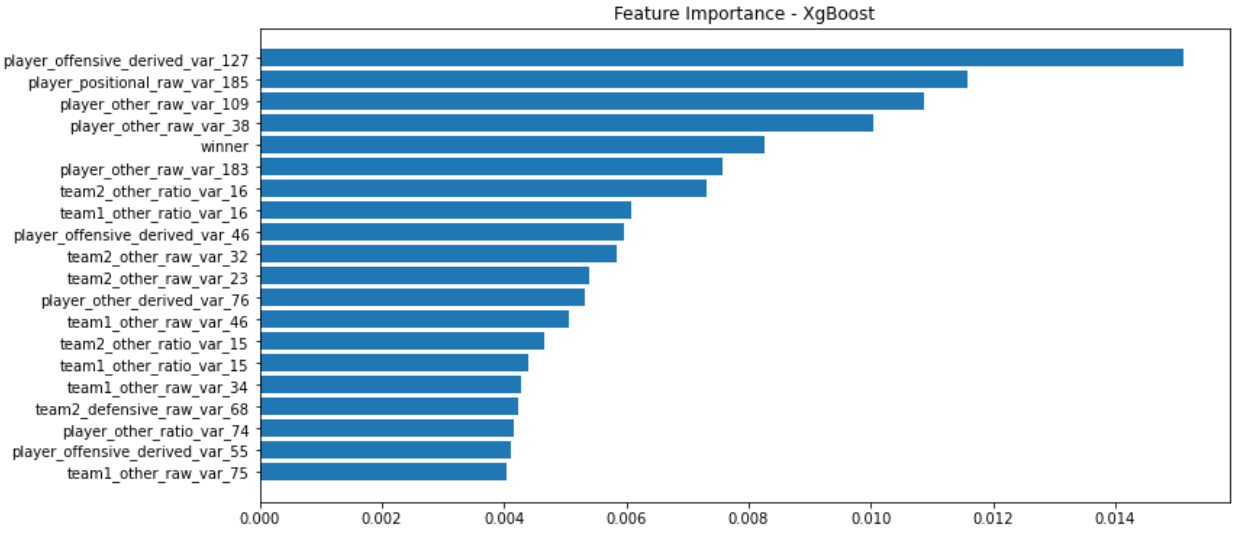
XgBoost: (1000 iterations, learning rate = 0.1)

Training Score	0.6482707292423282
Test Score	0.3152883601363955
RMSE	1.518515881023729

XgBoost: (1000 iterations, learning rate = 0.2)

Training Score	0.8050879305189585
Test Score	0.32229349814282393
RMSE	1.5107281058226947

Features Importance Graph:



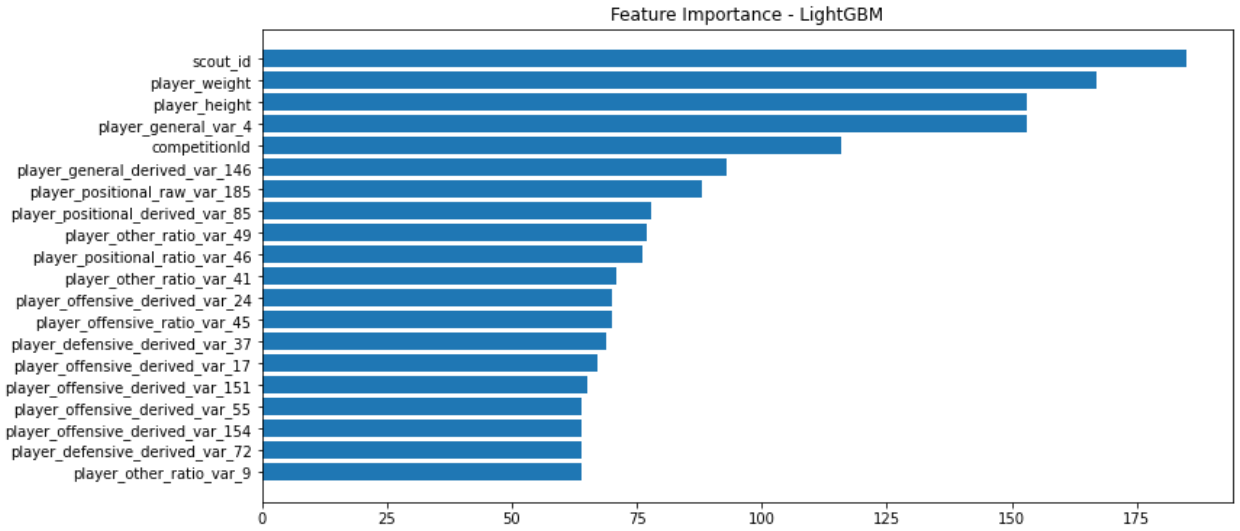
LightGBM: (1000 iterations, learning rate = 0.1)

Training Score	0.9834594655726425
Test Score	0.3879865383894856
RMSE	1.4356413885694483

LightGBM: (2000 iterations, learning rate = 0.2)

Training Score	0.9998864876708596
Test Score	0.3641706869682645
RMSE	1.4633080299145171

Feature Importance Graph:



Scope of Improvement

It is observed that LightGBM model with 1000 iterations and 0.1 learning rate has the best RMSE value and best R2 test score. However, it is to be noted that $R2_{train} >> R2_{test}$ which is not a great sign for the model. This might be because the model does not generalize well. Usually, high training score and low-test score indicates **over-fitting**. Overfitting means low bias and high variance. Thus, the training error is low but the testing error is high.

The individual test R_2 value is also not a very satisfactory indicator for the accuracy of the models.

One of the key reasons why the models are not accurate have overfit is the **sheer large number of attributes** in the dataset. 799 columns will be hard for any model to fit and will require very precise hyper-parameter specifications. However, this dataset was chosen as it very closely resembles a player performance dataset involving numerous variables produced by the sensor devices. A dataset just having some basic football statistics might be deceiving due to many trivial stats in football do not necessarily have a direct co-relation to performance.

Considering so many attributes, a major scope of improvement would be finding an optimal learning rate, minimum sum of weights of all observations required in a child rule, the optimum value of the maximum depth of the tree, maximum number of terminal nodes amongst a few parameters. Hence, **Fine Tuning the model** is one major scope of improvement.

Even though one of the key advantages of boosting models is that they do not get majorly affected by label encoding and presence of missing values or NULL data, Mean Imputation, Quantile Transformation and Label Encoding can potentially cause some issues. Mean imputation will obviously not preserve the relationship among variables and affect co-relation. Similarly in Quantile Transformation, the distribution is modified and therefore the true distance between data points will not be known to your model, thus affecting the relation between those variables. Nonetheless, **more meaningful feature selection and feature engineering** is a crucial scope of improvement.

Since there are so many attributes to consider, Gradient Boosting models also help in finding relative feature importance and thus helps in narrowing down hundreds of features into 10-15 by taking the top ranked features. These features can then be fed into a logistic regression model or other models. The gradient boosting model is thus useful for variable selection and give a better perspective and idea to the scouts on what attributes to look for.

Additionally, to all the above models, **Stacking Regression** was also attempted to enhance the accuracy by forming linear combinations of different predictors i.e. LGBM, CatBoost and XgBoost and make a new model to give improved prediction accuracy. Unfortunately, the model did not run due to CPU constraints. (Codebook Link)

5 SCOPE IN THE FOOTBALL INDUSTRY

Big data has become the most valuable player in the industry. The Scouting network of football clubs and scouting agencies have used big data extensively over the last few years in recruiting players. With the development of machine learning algorithms and data analysis techniques, whether the technology can gradually replace the roles of physical scouts is a very interesting debate.

Football is one of those sports where judging a player simply on data and statistics can be deceiving and inaccurate. Football is tactically a very vast game with different managers preferring specific styles of play, tactics and player specific roles on the field. Added to that, many aspects of the game on which has a high correlation with good performance can be have opposite meanings. Consider a midfield player having 90% passing accuracy and much higher number of passes played when compared to another player with a much lower passing accuracy and passes attempted. However, the player with low accuracy might take more risks during a match as they try to create more chances rather than playing simple passes and maintain possession of the ball. In this case, it is not possible to conclude which player is better but the specific coach that wants that player might have a tactical style that will suit the player. Football from the perspective of analytics, is a very visual sport where a scout would prefer to watch the player live in a match and during training sessions, understanding each and every aspect of the player's game which cannot be reflected by data and stats. Another very important factor while recruiting a player is the player's personality and mentality, which is crucial to ensure a perfect team chemistry making sure the player fits well and does not poorly affect the morale and character of the team. This is the reason why most top clubs and scouting agencies still depend on physical scouts and would potentially continue to do so.

However, prediction models can definitely aid the scouting process in many ways if not replace it completely. During the pandemic, when football was being played behind closed doors, it was becoming impossible for scouts to function normally for some time to come by going to watch players live [7]. For the most professionally run household-name clubs of European football, it's practically out of the question to sign a player who hasn't been properly vetted in several physical meetings. Not being able to meet in person with a player, their family and entourage is a huge

disadvantage. Such encounters give many important clues as regards to the player's personality and how keen they really are to join the club.[7] The maintenance and running of scouting networks is a very expensive process at the top level as a large number of scout travel around the world to physically see players play which obviously does not help the budget. Big data and especially finely tuned prediction models do provide a way of shortlisting players on the scouting database. Prediction models can also help analyze a team’s current players performance and also help establish interesting co-relations with some attributes which can lead to many developments in the sport. The accuracy of such models can be inaccurate and with the hope of more accurate regressors and data engineering, we expect these models helping the scouting process even more.

6 REFERENCES

1. "Henry Chadwick." *Baseball Hall of Fame*, https://baseballhall.org/hall-of-famers/chadwick-henry#:~:text=In%201859%2C%20Chadwick%20formulated%20his,them%20with%20the%20letter%20'K.
2. "A Guide to Sabermetric Research." *Society for American Baseball Research*, https://sabr.org/sabermetrics.
3. Schroer, Alyssa, et al. "How Sports Analytics Are Used Today, by Teams and Fans." *Built In*, https://builtin.com/big-data/big-data-companies-sports.
4. Lutz J, Memmert D, Raabe D, Dornberger R, Donath L. Wearables for Integrative Performance and Tactic Analyses: Opportunities, Challenges, and Future Directions. Int J Environ Res Public Health. 2019 Dec 19;17(1):59. doi: 10.3390/ijerph17010059. PMID: 31861754; PMCID: PMC6981928.
5. Brownlee, Jason. "How to Use Quantile Transforms for Machine Learning." *Machine Learning Mastery*, 27 Aug. 2020, https://machinelearningmastery.com/quantile-transforms-for-machine-learning/#:~:text=Quantile%20transforms%20are%20a%20technique,the%20performance%20of%20predictive%20models.
6. By: IBM Cloud Education. "What Is Boosting?" *IBM*, https://www.ibm.com/cloud/learn/boosting.
7. Karlsen, Tor-Kristian. "How Soccer Scouting, Budgets during Coronavirus Could Impact Transfer Market." *ESPN*, ESPN, 7 June 2020, https://www.espn.in/football/blog-espn-fc-united/story/4107274/how-soccer-scoutingbudgets-during-coronavirus-could-impact-transfer-market.
8.