Aug-Dec 2020

Project Report: Prediction of Player's Position in Football

~ Himanshu Gupta

Abstract

The Fédération Internationale de Football Association (FIFA) is a governing body of football (sometimes, especially in the USA, called soccer). FIFA is also a series of video games developed by EA Sports which faithfully reproduces the characteristics of real players. FIFA ratings of football players from the video game can be found at https://sofifa.com/. Data from this website for 2019 were scrapped and made available at the Kaggle webpage FIFA 20 complete player dataset. We will use the data to build a predictive model for the evaluation of a player's position. Subsequently, we will use the model exploration and explanation methods to better understand the model's performance, as well as which variables and how to influence a player's value.

1. Problem definition description

The project objective is to provide statistical insights to football clubs worldwide. The project involves using EA Sports' FIFA video game data as a prototype for providing insight into real football. The actual data and statistics are expensive to obtain and often kept secret by the respective clubs' analytics teams.

The main focus of this project will be:

- Suggesting Playing positions based on player's statistics eg. age, strength, speed, agility, etc.
- Analyze how attributes like pace, shooting, dribbling, passing, etc. generally vary with different positions of players on the field.
- Analyzing players' market values and wages based on the above statistics and trying to determine the features which are the most influential in players' transfer market value.
- Analyzing how age affects the physical and skill-based statistics of players.
- The best players with preferred positions and money value are arranged.
- Buyers can go through thier budget can can pick squad accordingly

2. Technology landscape assessment

Most of the major football clubs have their statistics teams which are getting more and more important every season. Hiring a top statistics firm is increasingly becoming some of the club's top priority just like hiring a world-class manager.

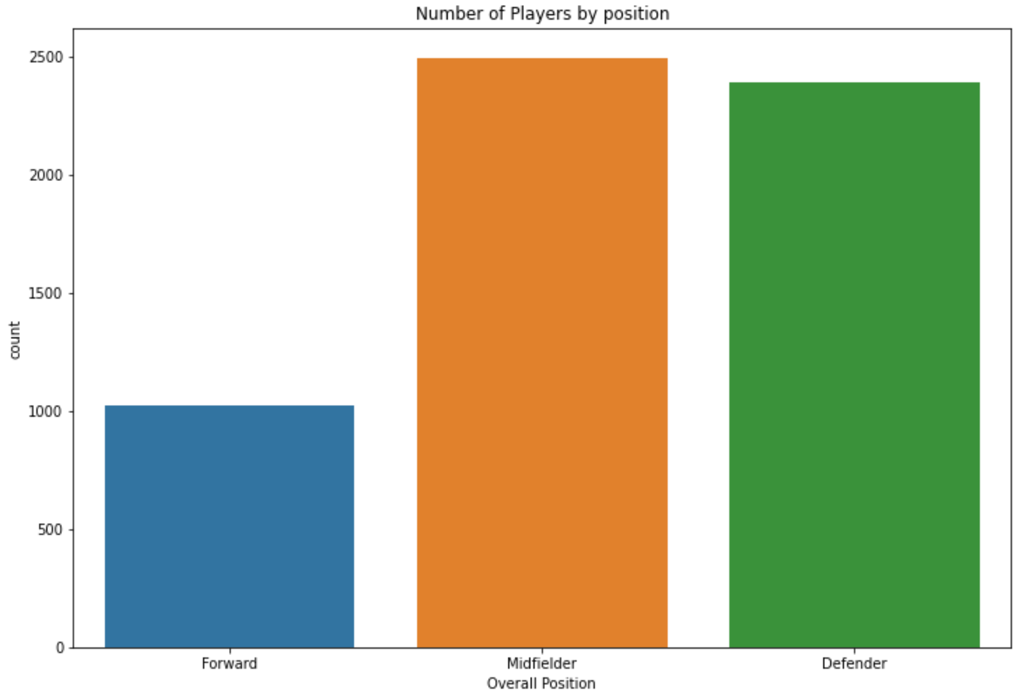
Despite the acceptance statistics are getting in football, most of the known use of statistics seems to be rudimentary and not in-depth. passes completed, shots were taken, average traveled distance are some examples of popular statistics which influence player's stature and those are obviously justified but data can offer much more intuition than that. Young players' positions and market values are still determined by experienced scouts' intuitions and extensive trial and error methods.

There are a few firms that cater to the club's needs for in-depth data analysis. Most of the other data analysis firms focus on fantasy football and betting predictions rather than actual involvement in squad building and player development. There is a growing market for such service providers.

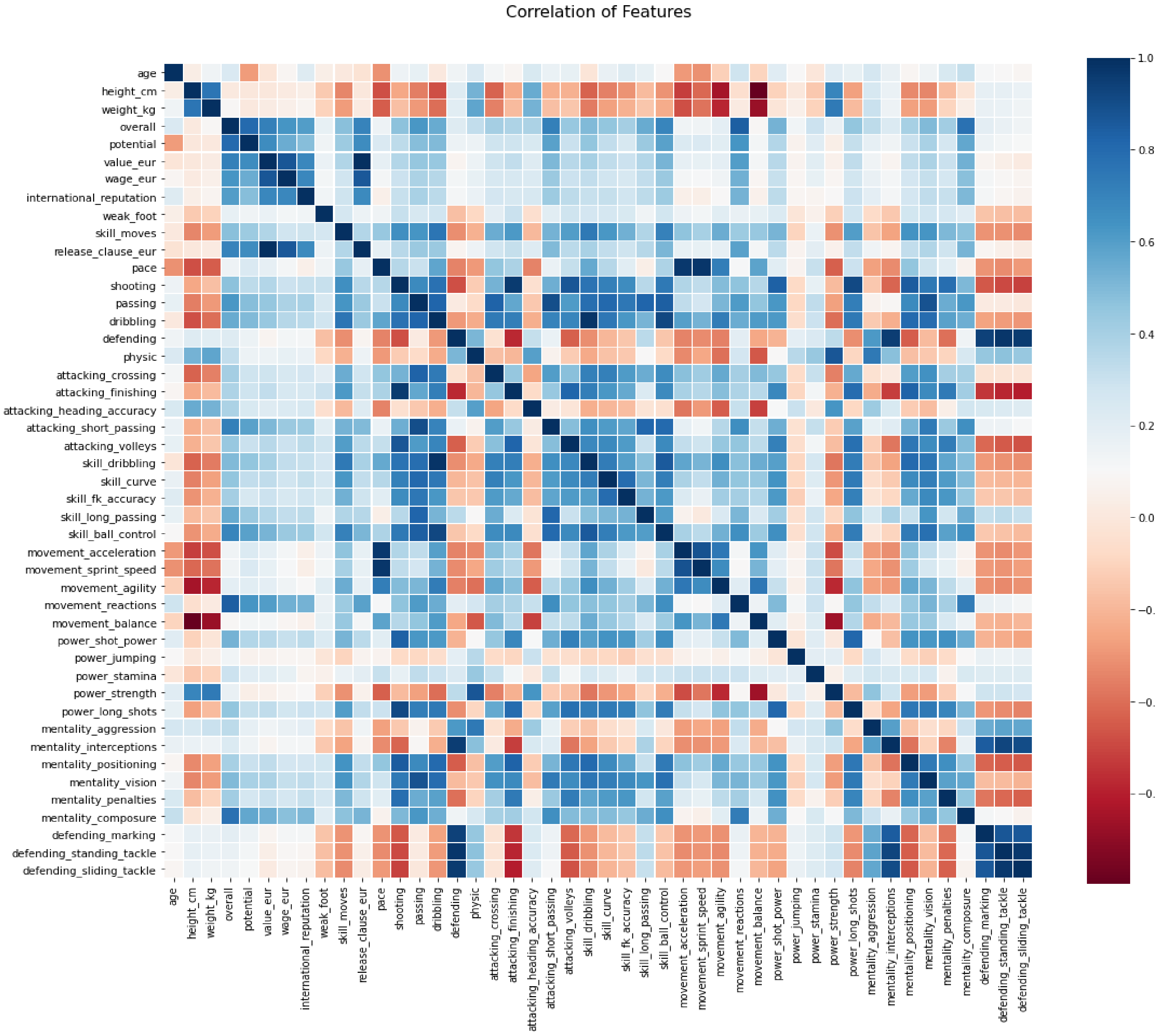
3. Data set Details

Every player available in FIFA 19 and FIFA 20 with 100+ attributes, URL of the scraped player, player positions, with his role in the club and in the national team, Player attributes with statistics as Attacking, Skills, Defense, Mentality, GK Skills, etc. Player personal data like Nationality, Club, DateOfBirth, Wage, Salary, height, weight, age, etc.

To build our model we had to remove the object type data like short name, ID, player URL, Last name, Player traits (playmaker, long-distance shooter, diver, leader, etc.), etc. The numerical data remained is thus used for the analysis. For calculating the position we used the 'team_position' attribute. All kinds of midfielders ie Attacking, right, left, defensive, etc. are given the position 'Midfielder' and similar for forwards and defenders.

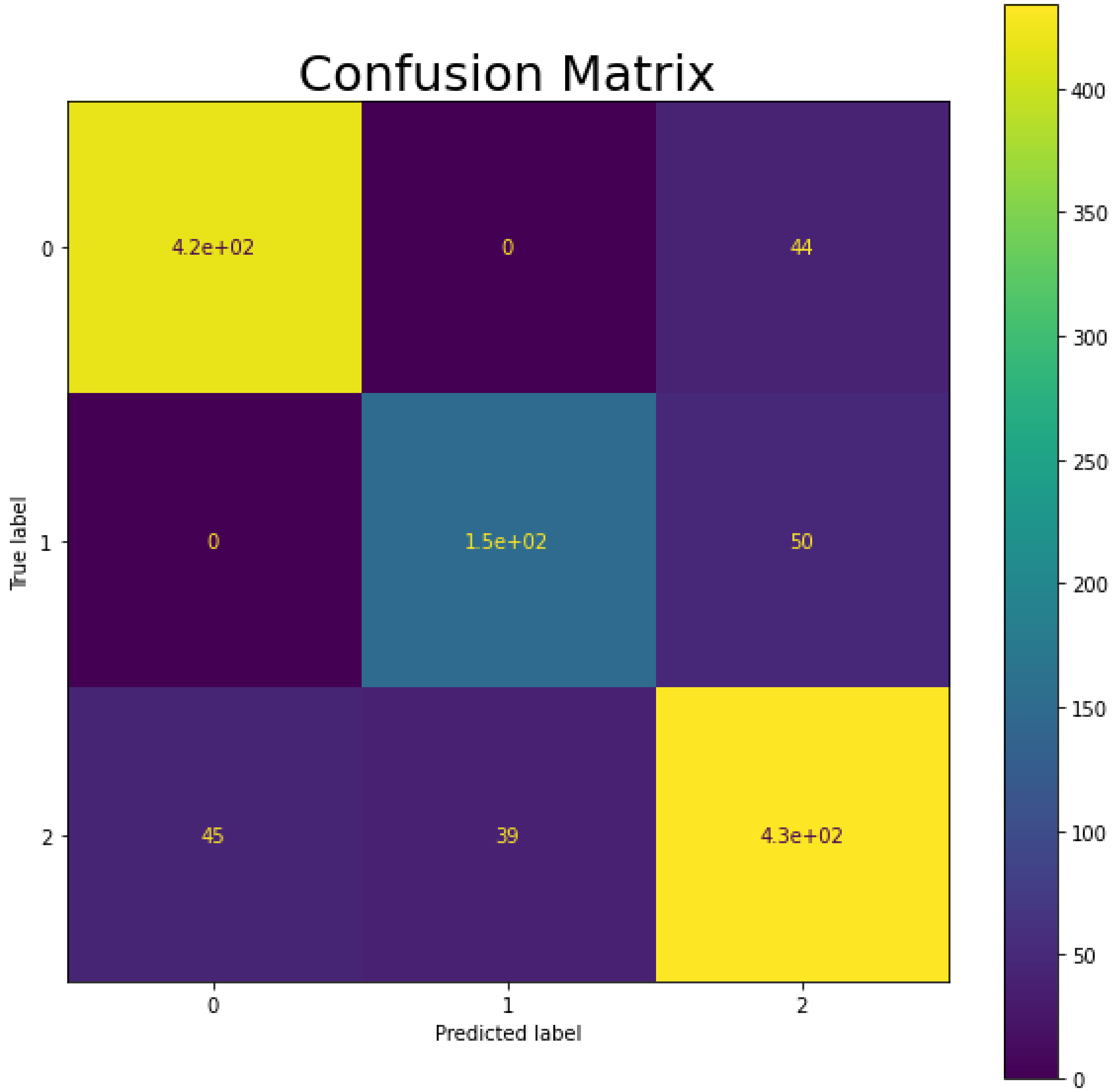


The data set also had ‘SUB’ and ‘RES’ (substitute and reserves) as ‘team_position’, due to which accuracy of any model will be very less because differentiating a midfielder substitute and a playing 11 midfielder has other aspects such as overall team statistics, the kind of players it contains, etc. We have not considered such attributes in our analysis and hence we have removed the substitutes and midfielders from our analysis which reduces the data set to nearly 5000-6000 players from 13000 players. After this, the data’s position information content is as shown in the above graph and the correlation between different attributes were like:

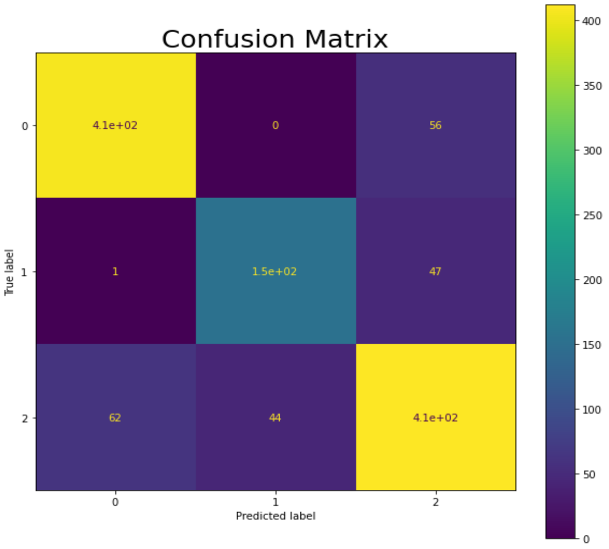


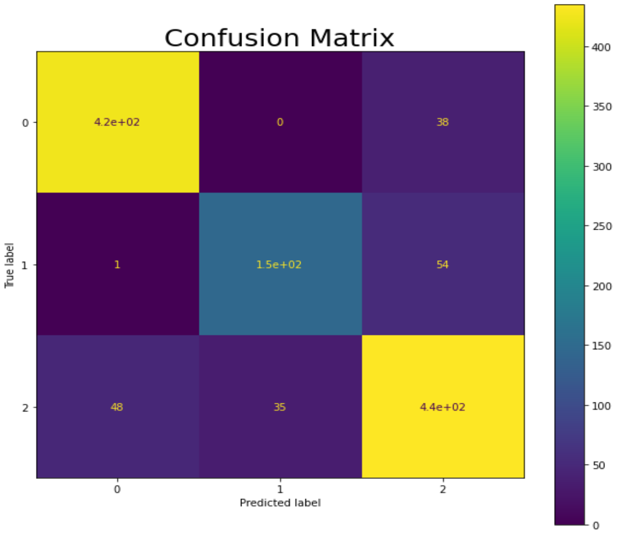
4. Methods and Approaches

4.1 Logistic Regression: We applied our first model Logistic regression without cross-validation and with cross-validation of 5 folds. Surprisingly we got the same accuracy of 0.85 with both the models. Below is the confusion matrix:



4.2 KNN: Then, we applied the KNN model without cross-validation and with grid search that applies to the K (number of neighbors) in a range from 1 to 25 with cross-validation of 5. KNN with cross-validation has higher accuracy of 0.85 over KNN without cross-validation of 0.82, as expected. Shown below is the confusion matrix of both the models:

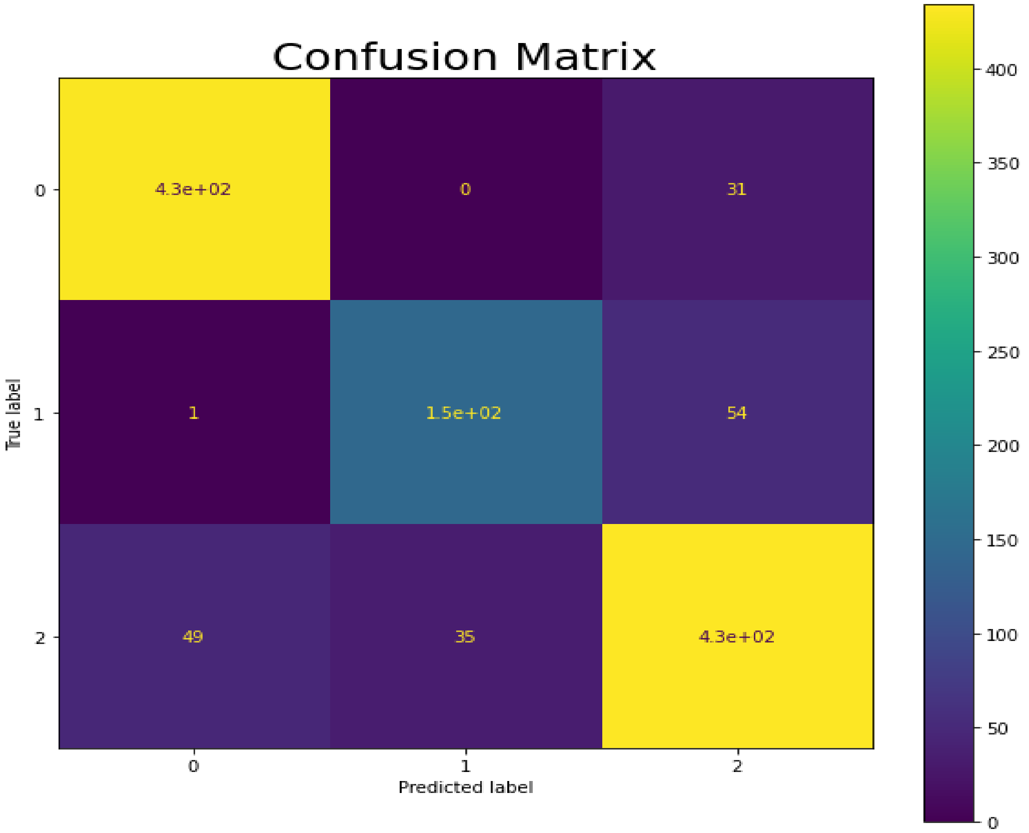




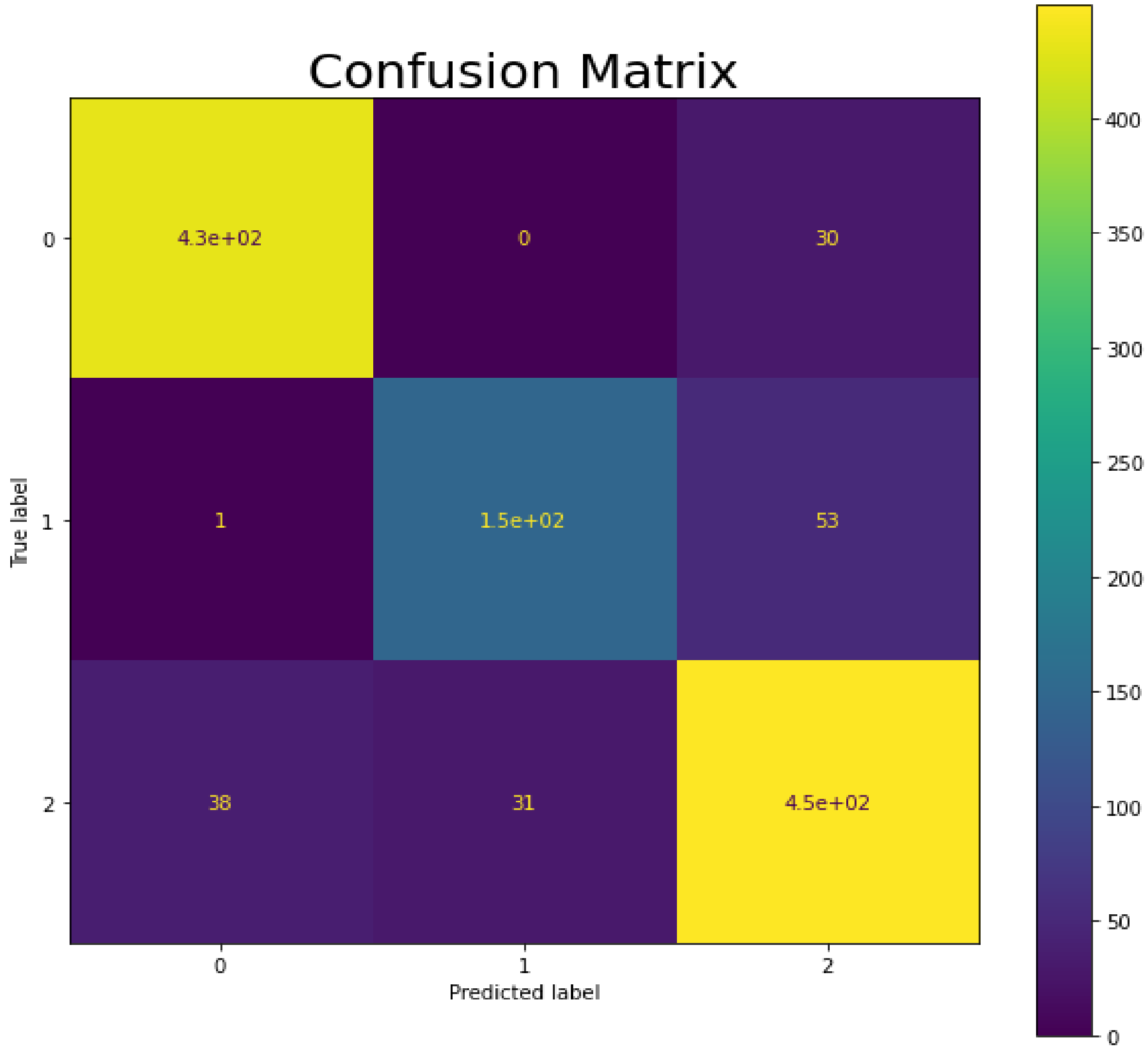
KNN without cross-validation

KNN with cross-validation

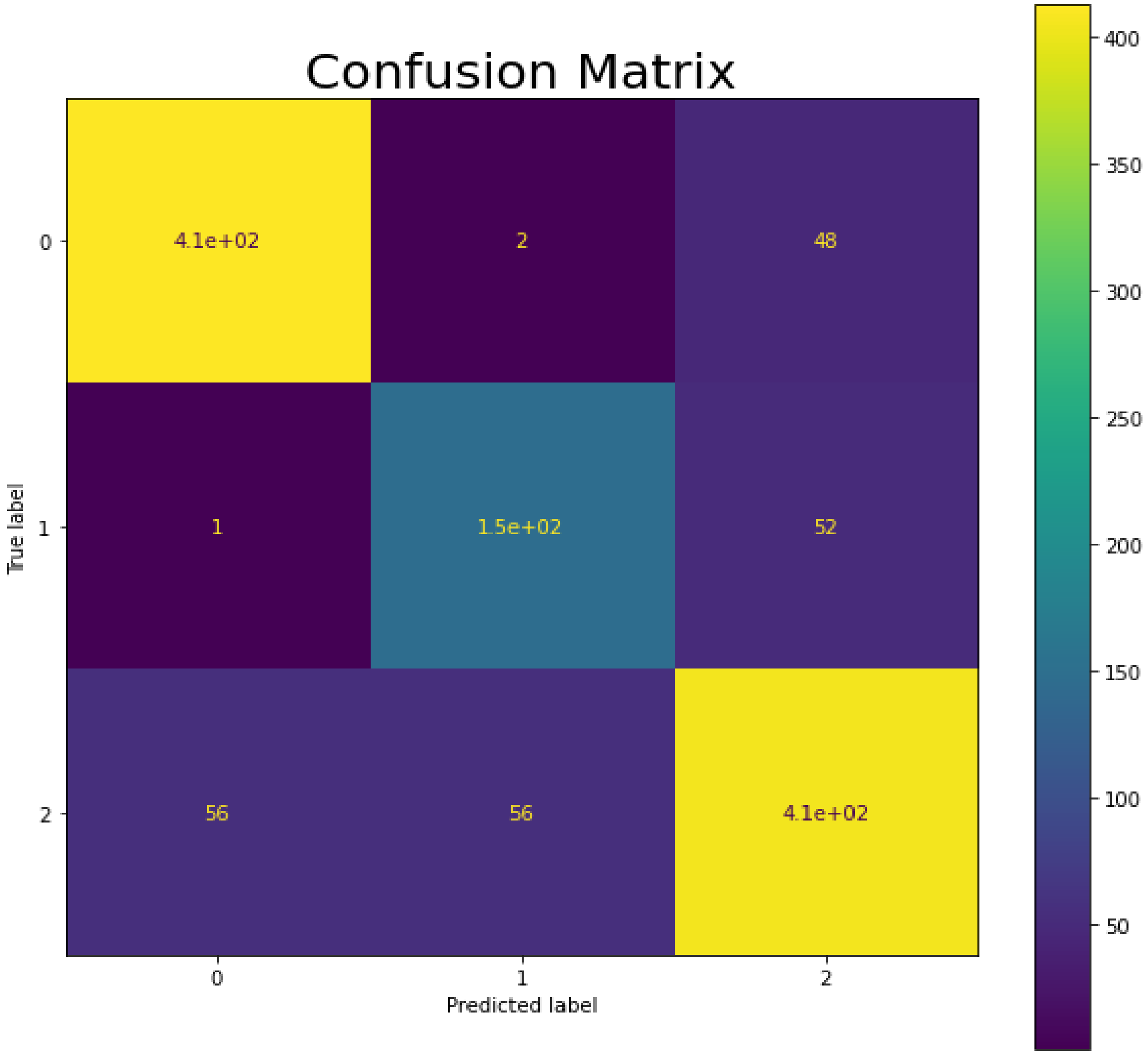
4.3 Random Forest Classifier: With random forest, we achieved an accuracy of 85.62%, a little better than the previous two models.



4.4 SVM Classifier: The accuracy of the model improved further to 87.06% using SVM Classifier. Its confusion matrix is as follow:

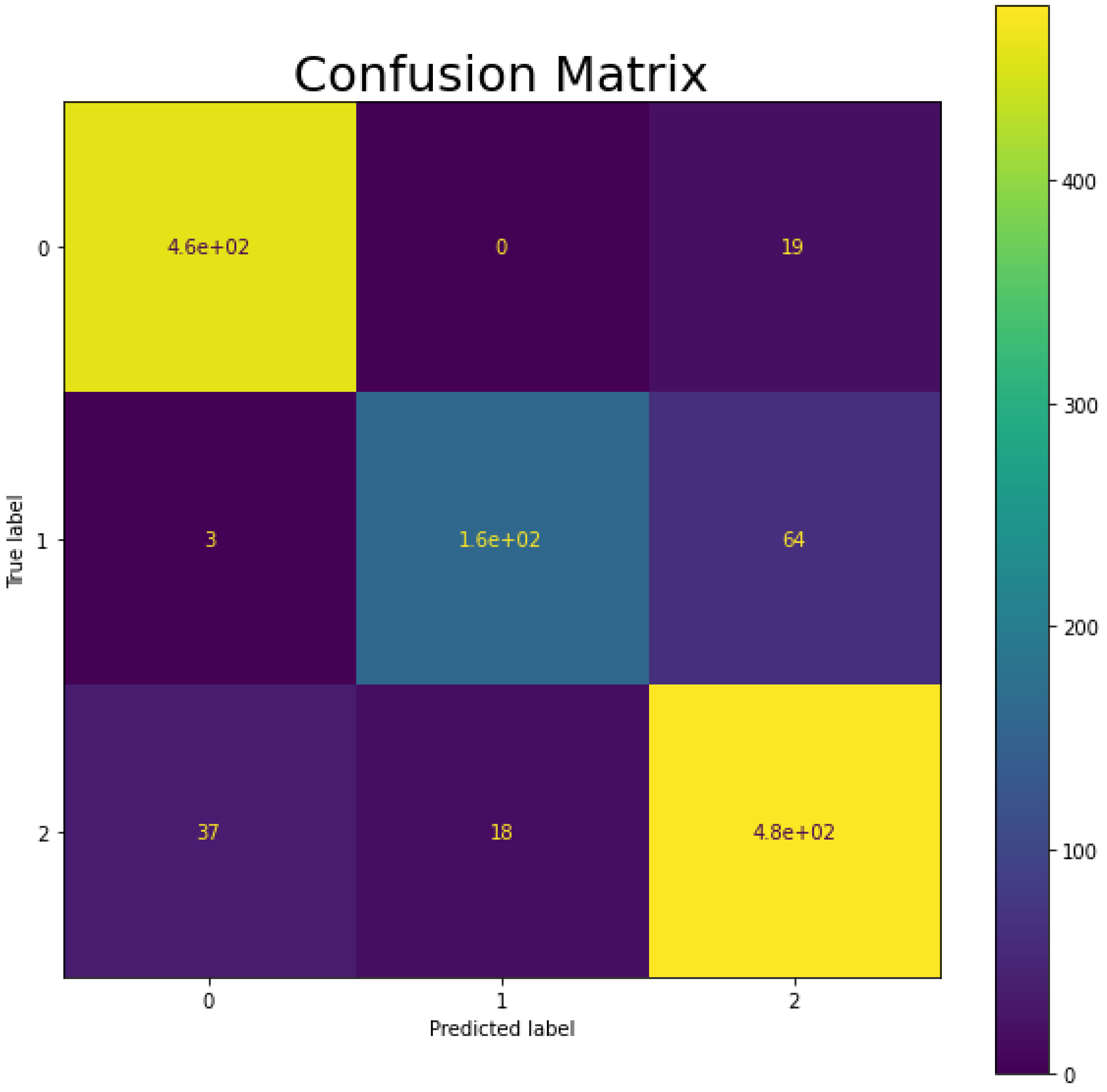


4.5 Neural Network: We used a neural network with different parameters of the number of hidden layers and the size. After tuning the parameters we find that the best option is 43,43,43 NN,i.e. 3 hidden layers of size 43 each. We got an accuracy of 81.82 %. Confusion Matrix is shown below:



5. Model Evaluation

According to the accuracy, we can see that almost all of the models get about the same score. The best model is the SVM classifier with an accuracy of 87.06%.Therefore,we used this model to predict positions of player on FIFA 20 datasets. We found an accuracy of 88.62 % which is very good and close to our training accuracy.Below is the confusion matrix on this test data (FIFA20):



Practical significance :

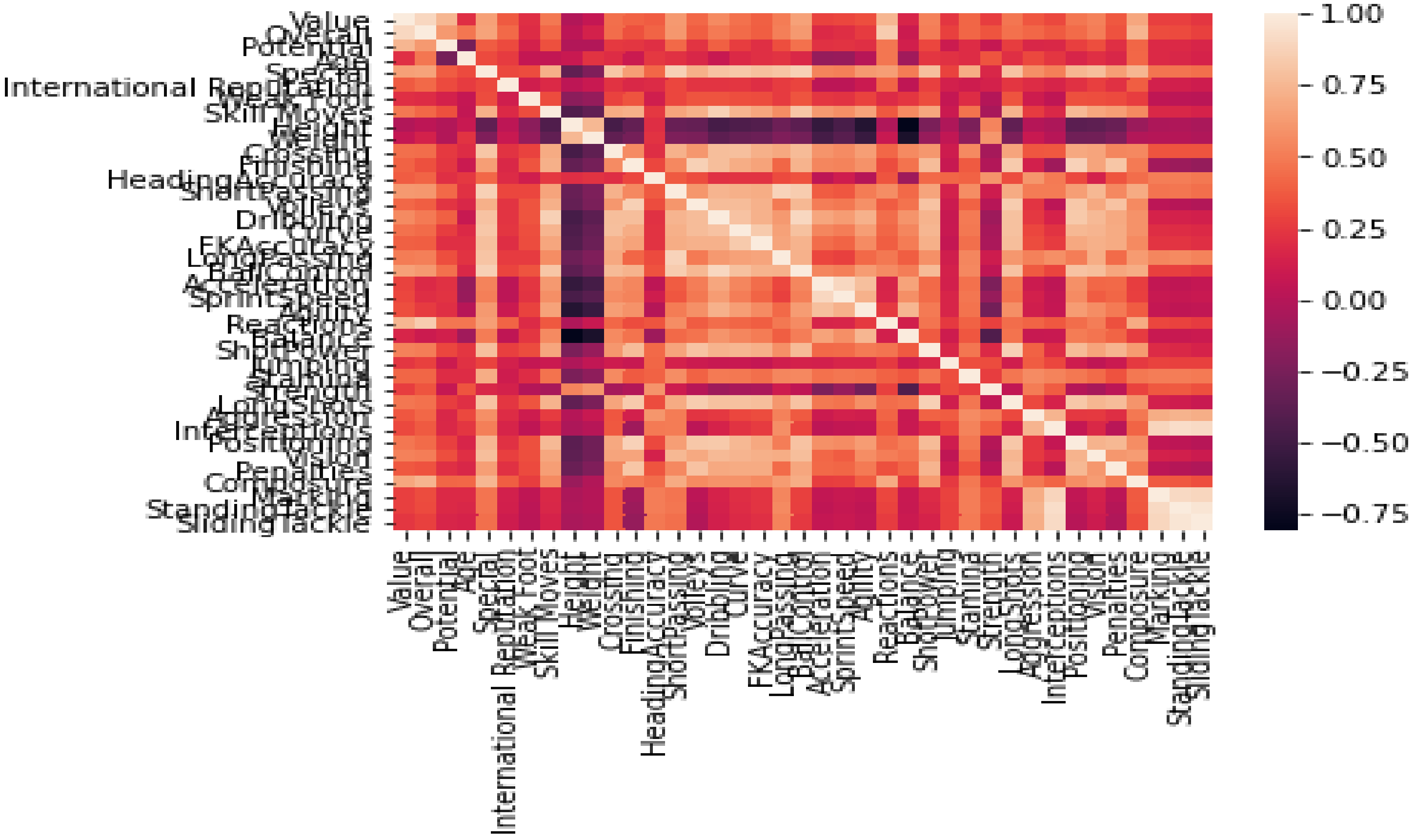
If you are an avid football fan, you might be aware of how a coach, The formation and The playing style can transform a player's career. There are countless examples of generational talents who never performed and Seemingly average players who turned into footballing icons just based on whether they had a coach who played them in a right position and systems. The proposed model with almost 85 % accuracy, gives an excellent way to judge a player's utility instead of using Trial and error or depending on the Coach's calibre.

PART II :- Player Market value prediction

- **Methods and Approaches :-**

Feature Selection

A correlation matrix of all the features with each other was plotted as shown below. The correlation indicator used was *Spearman Coefficient.*



Since our dependent variable is “Value”, The features having significantly high coefficient of correlation with Value were selected for training the proposed models.

The chosen variables are :- `['Special' , 'BallControl' , "Overall", "Potential", "Reactions", "Composure", 'ShortPassing']`

Train Test Split : 20% of the data i.e ~3300 entries were separated as a pure training set, The models will neither be trained or validated on the said set.

Following models were trained and Cross validated on the remaining 80% Data:-

1. Linear Regression : Linear regression showed a mere **44% R2 score** as demonstrated in the colab notebook and hence it was concluded to be insufficient to capture the information contained in the data about Players' market value.

2. Decision Tree Regressor :

   Decision Tree regressor showed a 100% R2 accuracy but resulted in high RMSE in 10 fold cross validation, This was a clear indicator of heavy overfitting and thus this approach was not used in the model.

3. RandomForestRegressor :

   Random Forest Regressor Produced R2 score on the training set and was seen to be resulting in significantly low validation set errors with minimum deviation thus it was selected as the best method to further fine tune.

Grid Search :-

To fine tune the random forest regressor Hyperparameters(Max_features and n_estimators), GridSearchCV method in python was used with multiple combinations of tunable hyperparameters. As demonstrated in the notebook, The best combination produced by Grid Search Cross validation was Max_features = 4, and n_estimators = 30. This model was then selected as The final model.

Performance on the Test Set : -

The Random Forest regressor with chosen parameters performance was recorded to be

- R2 Score : 95.54 % (Explained variance)
- RMSE = 1.16 Million Euros.

    - Note that the RMSE on training set is very encouraging since the players' value range from 1 million to 150 million with a normal distribution, Thus we can say that the model performance is very satisfactory

Practical significance :

With extremely competitive football leagues; both emotionally and economically football has ultimately become a game of thin margins. Optimising economic resources and building a high performing team is often credited by many to a competent analytics team of the club. This model provides a prototype of using data to scout talented players and be one step ahead in the transfer fee negotiations by knowing exactly how much money the player is worth.

7. Future Work

Keeping in mind how football clubs and virtuals games are growing, we can see a lot of different quality work that can be arranged out of this type of analysis and predictive models:-

- Probing whether the current year's video game data can provide insight into the future rating of players.
- Historical comparison between Messi and Ronaldo or between other stars. (what skill attributes changed the most during the time - compared to real-life stats).
- Ideal budget to create a competitive team (at the level of top-flight teams in Europe) and at which point the budget does not allow to buy significantly better players for the 11-men lineup.
- Sample analysis of top n% players (e.g. top 5% of the player) to see if some important attributes as Agility or BallControl or Strength have been popular or not across the FIFA versions. An example would be seeing that the top 5% of players of FIFA 20 are faster (higher Acceleration and Agility or pace) compared to FIFA 15. The trend of attributes is also an important indication of how some attributes are necessary for players to win games (a version with more top 5% players with high

BallControl stats would indicate that the game is more focused on the technique rather than the physical aspect).

8. Conclusion

Overall, all models did a very good job in predicting positions. SVM classifier beats all of them by a little margin, but we can use any of the above model to predict players positions. We used python and scikit-learn, starting with just one feature (value) and then adding some new features for training the models (age and finishing features). We noticed that by adding new features to the model, we would not always have better results and we mentioned common approaches to address this problem like feature selection, extraction and dimensionality reduction. As evaluation metrics for regression models, we applied two of the most commons: the mean square error (MSE) and the R^2 score.

References

1. FIFA 19 complete player dataset
2. How the spreadsheet-wielding geeks are taking over football