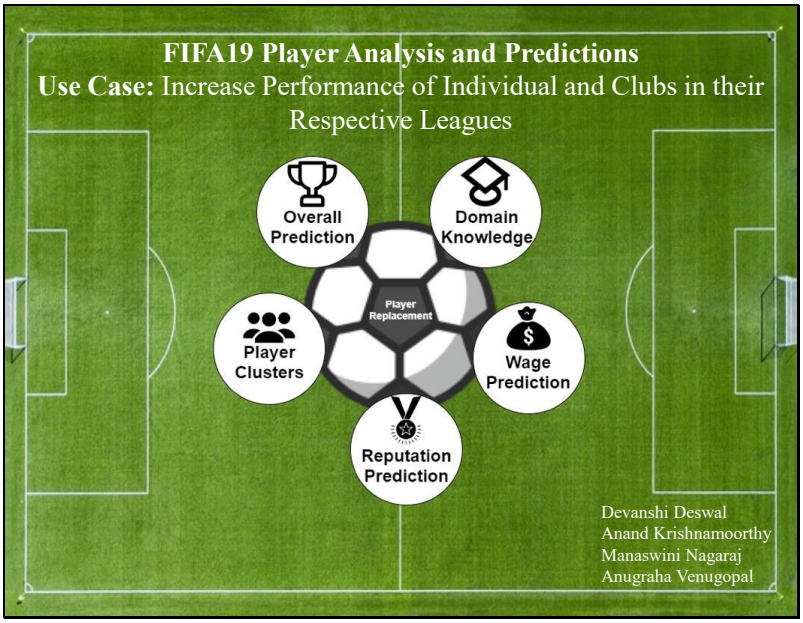


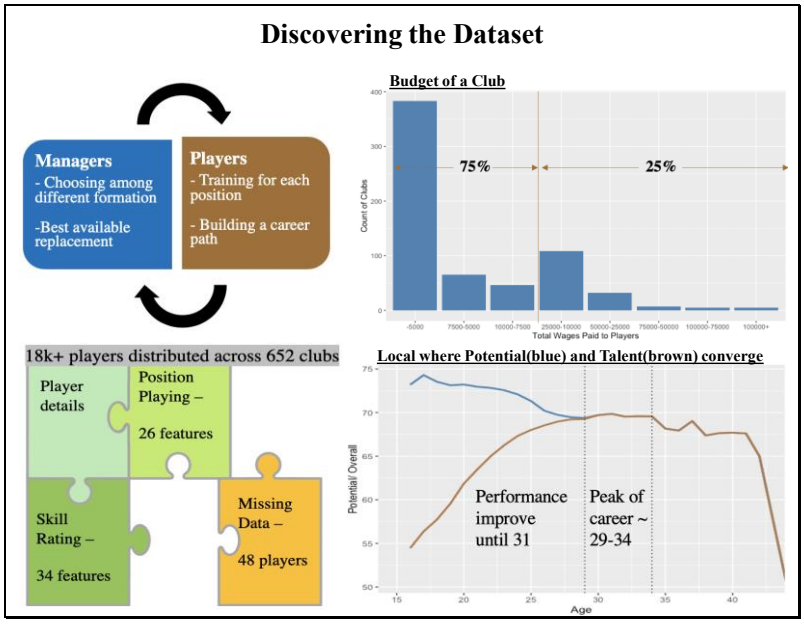
Introduction:

There is an avid interest in statistical methodology in sports. In the sports business, football is a lucrative industry associated with more than half of the global population. Understanding player skills is an integral part of the analysis of these methods in producing optimal results in terms of wins, team formation, and financial success.

The project 'FIFA19 Player Analysis and Predictions' aims towards Increasing Performance of Players and Clubs in their Respective Leagues by predicting the players' overall score, wage, reputation, and suggesting similar players.

The project consists of modules that both individually and in combination help achieve the goal.

- The overall prediction of the player helps determine the best strategy for team formation.
- The predicted wages of the players can help draw a fair contract between the player and club.
- The prediction of player reputation helps the team to maintain/strengthen the fan base.
- Clustering of similar players can help find an ideal alternative option for substitution during a match.
- These results and modules, along with expert knowledge, can help to make contracts with new clubs fulfilling the dynamics and the requirements of the club and the players.



Project Benefit:

The project focuses on helping the managers and players, both current and prospective. Managers seek to identify the skill set critical for each position. The project enables a selection of a team formation for upcoming league matches given player skills. Players, on the other hand, need to focus on how to properly train such that it makes them great in a position. Clubs can also focus on making ideal player replacement considering the team's budget and player requirements. A player can also have an informed estimation of the wage he might receive based on his current skills.

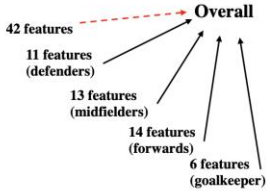
Dataset and EDA:

The dataset "FIFA 19 Complete Player Dataset" is publicly available on Kaggle and comprises over 18,207 players spanning over 89 attributes including personal details, evaluation of their finesse at 26 field positions, and quantification of 34 skill matrices. The dataset has 48 players with missing statistical values. On further research, it was observed that these players have played fewer than 100 matches in FIFA, and therefore accurate statistics of these players are not available, hence they were excluded from the analysis.

On exploring the data, an interesting finding came to light - more than 75% of football clubs are low budget, with less than 10k euros spent on players' wages. Another interesting analysis made is that the players "overall" generally improves until age 31 and then it starts to decline. The peak of their career ranges from 29-34 years, making young players more favourable over older players with similar skills.

Skill Matrix

BEST SUBSET SELECTION



Skill score - Goalkeepers

	Goal Keepers
Reactions	81.2167
GKDiving	81.08225
GKHandling	93.18445
GKKicking	30.63
GKPositioning	96.12477
GKReflexes	82.16312

Skill score - Outfielders

	Defender	Midfielders	Forwards
Age	0	10.949936	0
Crossing	0	8.267614	0
Finishing	0	0	9.969935
HeadingAccuracy	20.34838	10.074742	13.834306
ShortPassing	20.767777	30.003866	13.487955
Volleys	0	0	0
Dribbling	0	11.714423	8.846439
BallControl	15.75742	28.382948	21.524179
Acceleration	0	16.921453	7.423845
SprintSpeed	20.33502	0	5.423694
Reactions	21.48102	29.776679	26.049071
ShotPower	0	7.435741	15.35048
Stamina	14.37919	22.183033	0
Strength	21.1263	8.903207	13.523013
Aggression	18.70381	0	0
Composure	0	10.749109	8.173287
Marking	32.42094	0	5.973179
StandingTackle	21.87765	0	0
SlidingTackle	15.73232	0	7.216664
Positioning	0	0	14.756012
Vision	0	6.386257	0

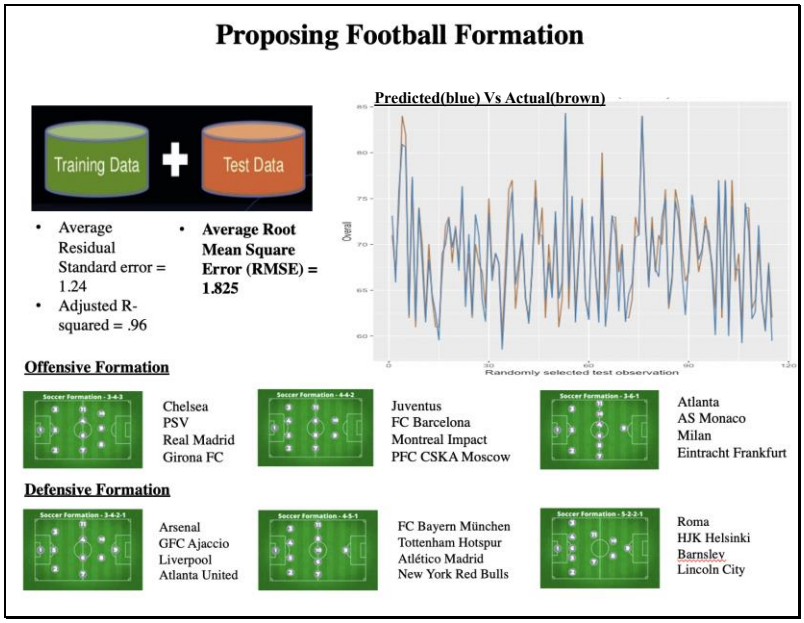
Module 1: Overall prediction

To identify the strengths and weaknesses of the clubs, the teams are evaluated by grouping players into these 4 positional buckets – forwards, defenders, mid-fields, and goalkeepers. The dimension of the dataset was reduced by 22 predictors, by grouping similar positions together into broader categories. Using the Best Subset Selection (BSS), the important skills at each position that contributes to building a strong model (and team) were identified. Using Cp and BIC graphs 11, 13, 14, 6 were identified as the optimum number of features required for each position respectively committing to the best bias-variance trade-off, since not much reduction in MSE is observed henceforth, against the minimum Cp and BIC values.

An overlap in the features is observed for players at different outfield positions. This is because players move forward and go back during the match; hence, there is no unique skill set that distinguishes the players at outfield positions, hence overlapping is bound to happen among skillset (apart from the goalkeeper).

Variable Importance:

The model-based approach was then used to find the variable importance for each skill at each position, considering the fact that although Short passing may be a major skill contributor for outfield players, its importance at each position tends to differ- more for a midfielder, less for forwards. Boosted trees, using the same approach as a single tree, but summing the importance's over each boosting iteration was used to find the variable importance. All measures of importance are scaled to have a maximum value of 100. Using the skill matrix, the player can set up the training schedule to focus on upgrading his skills for a specific position.

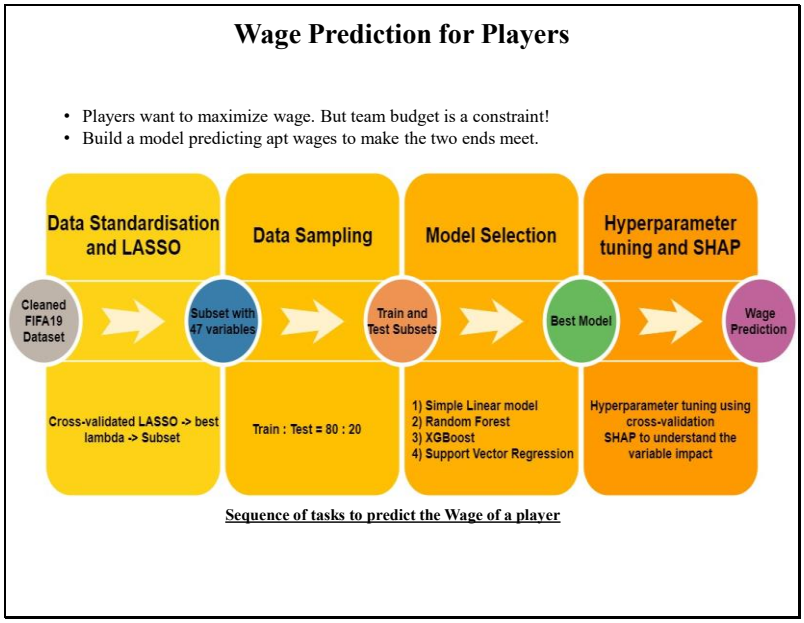


Regression model:

After the best subset regression was used for model selection to obtain a handful of models, residual analysis, transformation of the predictors and/or response, and addition of interaction terms are performed to arrive at the best regression model giving average residual error of 1.24 and R-squared of .96 averaged over each positional model indicating the data is fit well to the regression line. Upon testing the fitted model on test set Average Root Mean Square Error is observed to be 1.825 (overall positional variables) indicating the 'overall' values predicted by the model deviate from the actual values by 1.825.

Football formations are either offensive or defensive. A team's personnel make a big difference in how strong or how weak an offense/ defence can be. Using the prediction of the 'overall' of a player the best formation is suggested for a team among the 6 common formations. Each player in the team is evaluated for an overall score. Post which the top 11 players are identified among a club for each formation. The team strength is compared and evaluated using the top 11 players identified for each position. Each result is compared and evaluated to pick the best formation.

For instance, Juventus team's best forward players have a high overall score than their best defenders. Therefore, having a formation with more attacking players will benefit the team. Finally, 2-4-4 (2 defenders, 4 midfielders, and 4 forwards) formation was found to be best for the team.

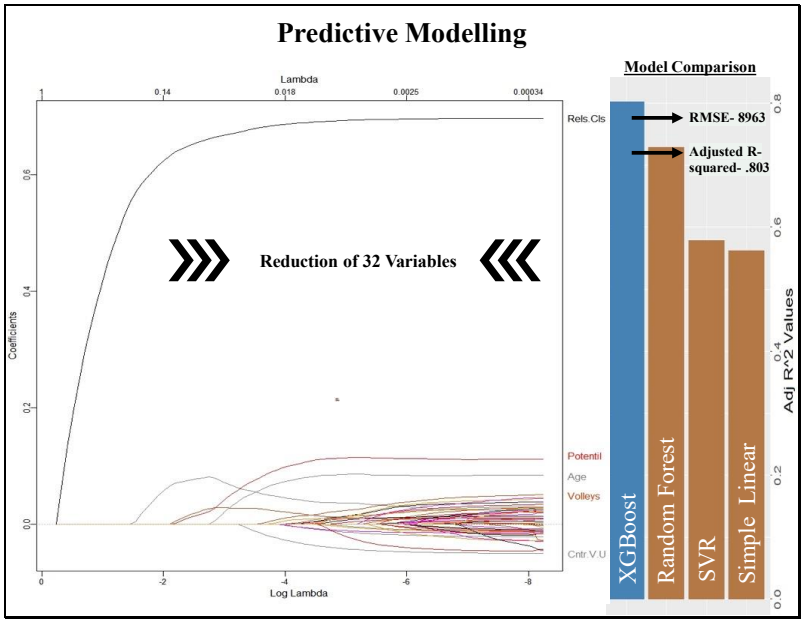


Module 2: Wage Prediction Overview

The prediction of wage helps the managers set a base line for the wage to be proposed to a new player based on his skillset. The managers can also consult with the team sponsors to understand the team budget to finalize the wage. Similarly, players can also have an informed expectation of how much they might receive when they transfer or join a new club.

The model for "Wage" prediction starts by applying LASSO over the cleaned and standardized dataset as a method for feature selection. Once a subset of features was obtained, the data was split into train and test subset with a 80:20 ratio respectively. After the data split, different models are applied to the training subset to see which model fits the data better, with the adjusted R-squared value as the evaluation metric to compare models. The model with the highest adjusted R-squared value was selected as the best one and its hyper-parameters were tuned using cross-validation to get the best out of it. Once the model is ready, it was tested using the test subset kept aside to obtain the RMSE value.

While the model predicts the wages that a player should get based on his skills, the managers, on the other hand, would want to know what are the exact features that contribute towards a player's wage, and how it affects the wage. Therefore, to interpret what the model now predicts, a SHAP model is implemented. SHAP explains what features are considered, in their order of importance and how those features affect the players' wages. Using this interpretation, the managers may decide to offer the player an extra incentive if needed.



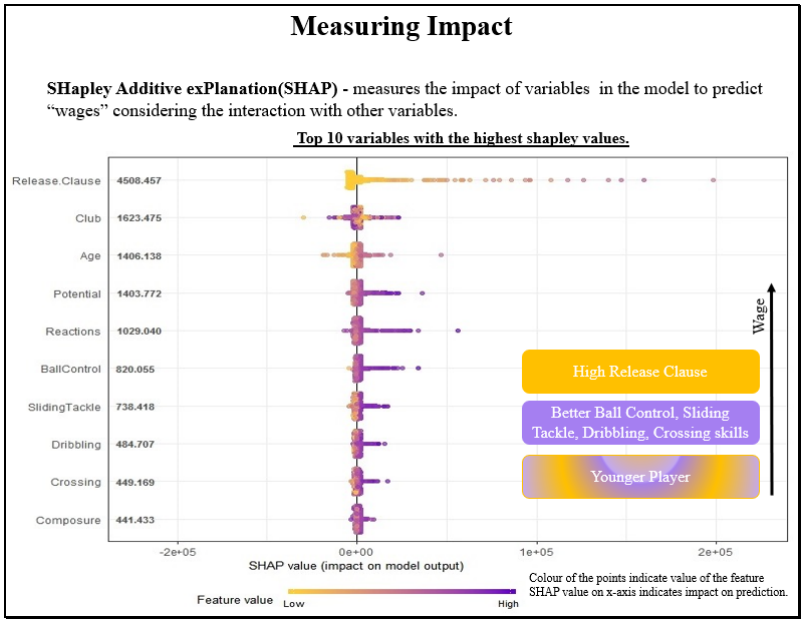
Subset Selection:

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that also performs feature selection along with regularization to enhance the interpretability of a model. Lasso uses an L1 penalty to perform subset selection, it has the effect of forcing some of the coefficient estimates to be zero when the tuning parameter lambda is sufficiently large. In this project, LASSO was applied over the standardized dataset of 79 features, and a cross-validation approach is used to find the best value of lambda, resulting in the cutting down of 32 variables from the original set. As observed in the Log-Lambda plot above, the variables that are most important shrink the last. It is observed that the Release Clause is the most important variable for the model.

Model fitting:

With the new subset consisting of 47 features, four different models- Simple linear model, Random Forest, XGBoost and Support Vector Regressor- were applied to it. Among these models, the tree-based algorithms- Random Forest and XGBoost showed good performance. And ultimately, XGBoost performs the best with an adjusted R-squared value of 0.8.

XGBoost is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Boosting is an ensemble technique that corrects the errors made by previous models while building new models. It is called gradient boosting when the boosting technique uses a gradient descent algorithm to minimize the loss. The models built were sequentially added until no further improvement was seen, to obtain one final model. The tree-based approach for XGBoost first built a decision tree with weak learners, then a new model was built to predict the errors of the previous model. This process was done stage-wise and the final prediction model was in the form of an ensemble of weak prediction models. In this project, after selecting the best model, it's hyperparameters were tuned using 10-fold cross-validation. After tuning, the model gives an RMSE of 8963 and an adjusted R-squared value of 0.8.



Interpretation:

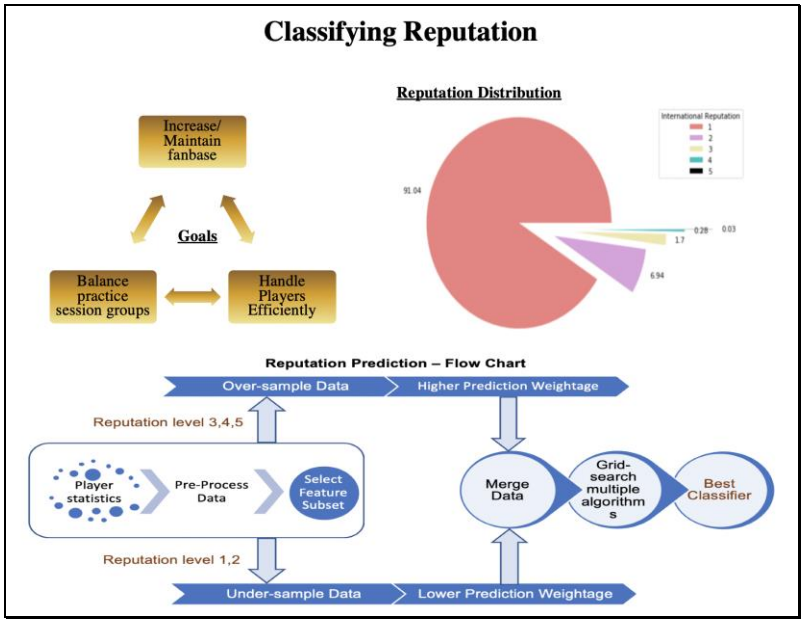
SHAP- SHapley Additive exPlanations, explains the prediction of an instance X by computing the contributions of each of the features to the predictions. To do this, SHAP computes Shapley values, which is a method from coalitional game theory, that tells us how to fairly distribute the "payout" among the features, essentially it is the average contribution of a feature value to the prediction in different coalitions.

SHAP is useful in scenarios where "right to explanations" is involved, for example, if a player invokes the "right to explanations" to know why he is being offered that particular wage, then the SHAP results can be used to deliver a full explanation on what features influenced the predicted value by how much.

The SHAP summary plot displays the features that affect the prediction in descending order of their effect along the y-axis and the Shapley values along the x-axis. The color of the points represents the value of the feature and its effect on the outcome.

It is observed that the highest contribution to the prediction of wage is contributed by the "Release Clause", lower the value of the "Release Clause" lower is the player's wage. Similarly, better the player is in Ball Control, Sliding Tackle, Dribbling and Crossing skills, positive is its effect on his wage thus increasing the value.

It is also interesting to note that on average, older the player is more negative the effect is on his wage.



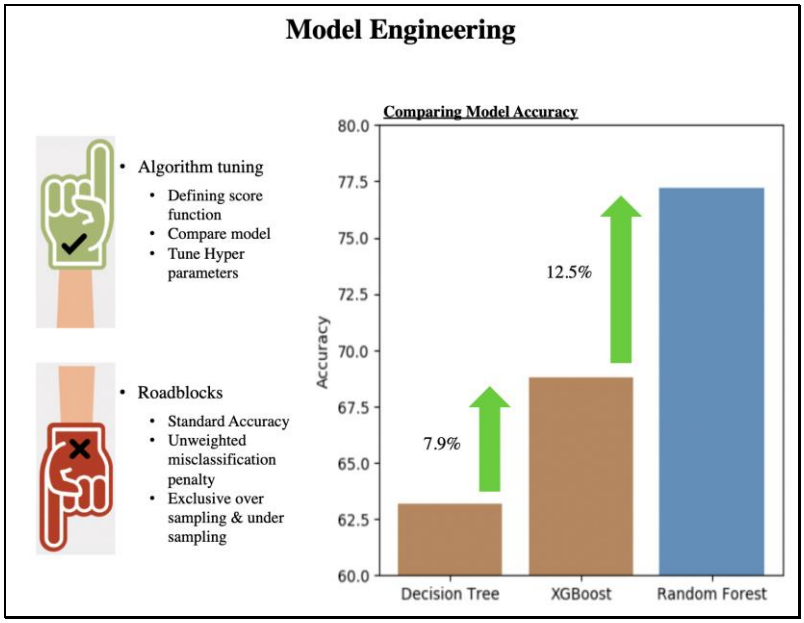
Module 3: Classification of Reputation Overview

The goal of this module is to predict the International reputation of players so that the team managers can handle player needs efficiently. The highly reputed players are essential in increasing/maintaining fanbase of the club. It is also essential to arrange team practice sessions with equal players of reputation in each group, therefore accurate prediction of reputation is very essential for a team.

EDA analysis:

There are 5 levels of reputation, therefore prediction of reputation is a multiclass classification problem. It is a highly imbalanced dataset with only 0.3% of players making it into level 4 and 5 reputations. The pie chart visually represents this class imbalance.

Level 1, and 2 players were filtered and identified as, set 1. Similarly, level 3,4, and 5 players were grouped and identified as, set 2. To create balance among classes, the data in set 2 was increased by oversampling and the data from set 1 was down-sampled. Both the data were then merged. A scoring function was then defined which applies more penalty for misclassification of set 2 players compared to of set 1 player. Various models were compared, and the best model was selected using grid search and random search techniques.



Approach:

A subset of attributes was identified as the best subset using the feature importance scores from the Extra Trees method. This subset was used for model building.

Steps in the model building:

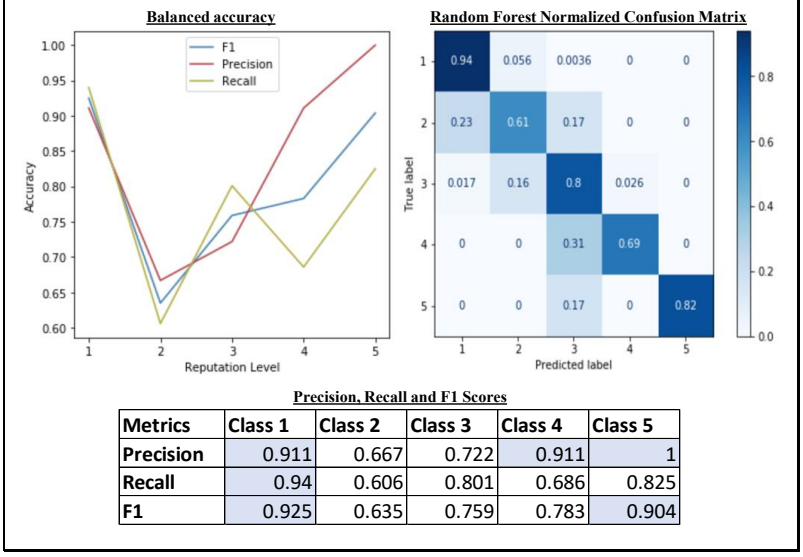
Initially, a Scoring function was defined which penalizes misclassification of higher reputed players. Then various algorithms like Logistic Regression, Naive Bayes, Decision tree, Random Forest, XGBoost were run against the scoring function to shortlist the best set of algorithms. Decision trees, Random Forest and XGBoost performed well. The hyperparameters of the shortlisted algorithms were then tuned using a randomized search technique. The bar plot compares the model accuracies. Random Forest was observed to predict the highest balanced accuracy of ~ 77%.

RoadBlocks:

Since the data is imbalanced almost all the data points were predicted as level 1 by the machine learning model. Therefore, Oversampling and under sampling of the data points were considered. Exclusive oversampling the set 2(level 3,4, and 5) data points causes the same data points to be multiplied several times, which produces poor results. Similarly, exclusive under sampling of set 1 (level 1,2 players) caused the number of data points to reduce drastically and lower the algorithm's performance. Therefore, to tackle these problems a combination of both oversampling and under sampling techniques was suggested.

Next tuning algorithm using default scoring methods, i.e. accuracy, and ROC which did not produce the desired result. Therefore, a customized soring function was defined which penalizes misclassification of set 2 data points more than the misclassification of set 1 data points. Additional accuracy measures like precision, recall and F1 scores were considered to evaluate the models.

Model Evaluation



Precision, Recall and F1 Scores

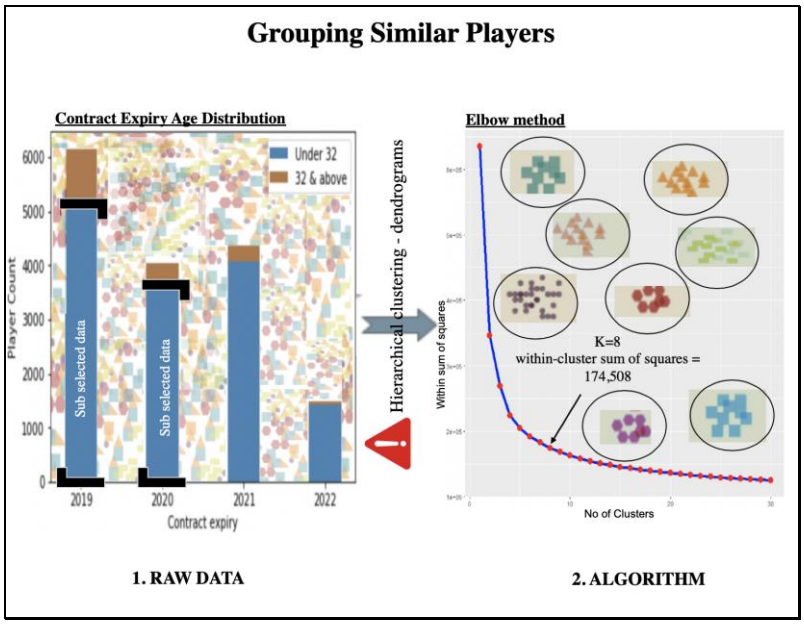
Metrics	Class 1	Class 2	Class 3	Class 4	Class 5
Precision	0.911	0.667	0.722	0.911	1
Recall	0.94	0.606	0.801	0.686	0.825
F1	0.925	0.635	0.759	0.783	0.904

Evaluation:

Post tuning the hyperparameter of the models, the Random forest produced the maximum accuracy. The confusion matrix of the Random Forest model was evaluated. Diagonals from top left to the bottom represent the model accuracy. Ideally, all the diagonal elements should have value 1 and the rest of the elements should have value 0. The balanced accuracy of the best model (Random Forest) is ~77%. High accuracy of 69% and 82% is achieved for reputation level 4 and level 5 respectively. These levels are quite high compared to the initial accuracy of 15% obtained for levels 4 and 5.

For an imbalanced dataset, evaluation measures like Precision, Recall, and F1 scores are ideal. Precision is the proportion of the prediction made by the model corresponding to a particular class that also appears in the actual dataset. The Recall is the proportion of the actual number of examples in a particular class identified by the model. F1 score combines these two scores to provide an average score.

As expected, because of the higher datapoints high precision, recall and F1 score of over 90% were obtained for reputation level 1. The graph shows a high F1 score of over 70% for reputation levels 3,4 and 5.



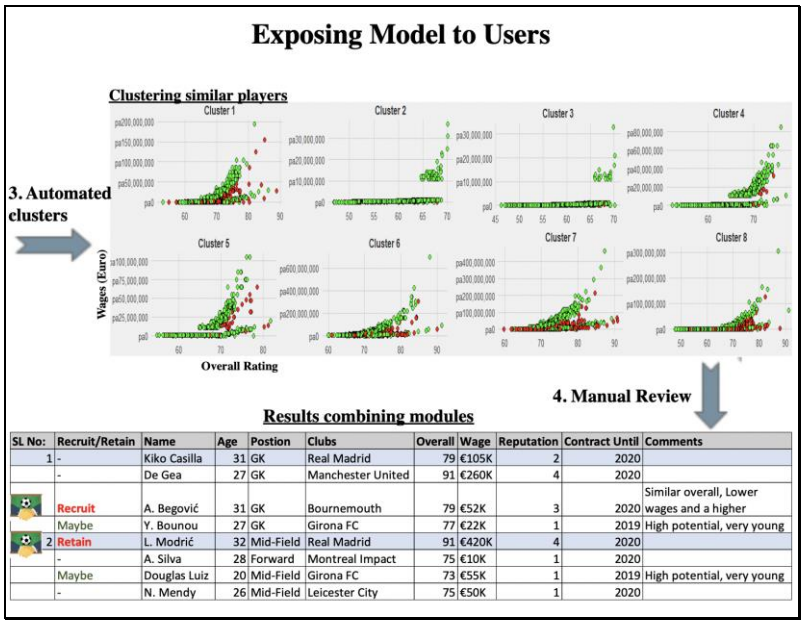
Module 4: Identify similar players

To provide a list of choices for replacement of a player who has a similar skillset for a particular position, we group players based on the similarity of their ratings for various skills and positions. The reasoning behind this is that a player in a particular position on the team is best replaced by another with a similar skill set, who would be of value in that position since different positions on the field require expertise in different skills. For a player needing a replacement, it is easy for the manager/coach to choose one from the same cluster, meeting the team's constraints in terms of the team dynamics as well as financial considerations.

First, the relevant features, i.e., the ratings of the players for various skills, were scaled. Next, the elbow method was chosen first to find the optimal number of clusters (value of k) for k-means clustering. The 'elbow' of the plot occurred at k = 8, with a total WSS (within-cluster sum of squares) value of 174,580.1.

Failures:

Since there are more than 18,000 observations in the data set, the dendrograms, as well as the clusters in them, were highly cluttered, making readability and interpretability difficult. Uncertainty also arose as to whether Euclidean distance or Correlation-based distance is a more suitable choice of dissimilarity measure. K-means clustering was used which gives more accurate and interpretable results.



SL No:	Recruit/Retain	Name	Age	Postion	Clubs	Overall	Wage	Reputation	Contract Until	Comments
1	-	Kiko Casilla	31	GK	Real Madrid	79	€105K	2	2020	
	-	De Gea	27	GK	Manchester United	91	€260K	4	2020	
	Recruit	A. Begović	31	GK	Bournemouth	79	€52K	3	2020	Similar overall, Lower wages and a higher
	Maybe	Y. Bounou	27	GK	Girona FC	77	€22K	1	2019	High potential, very young
2	Retain	L. Modrić	32	Mid-Field	Real Madrid	91	€420K	4	2020	
	-	A. Silva	28	Forward	Montreal Impact	75	€10K	1	2020	
	Maybe	Douglas Luiz	20	Mid-Field	Girona FC	73	€55K	1	2019	High potential, very young
	-	N. Mendy	26	Mid-Field	Leicester City	75	€50K	1	2020	

Clustering Results:

The plot depicts the grouping of players into 8 clusters based on the similarity in the skill and positional values. The green points depict players below the age of 32, and the red points depict those above the age of 32. For a similar player cluster, the relationships between the players' overall ratings (x-axis) and current market value in Euro (€) (y-axis) are captured for easy comparison.

The algorithm allows managers/coaches to input a player and it generates a pool of players for replacement. Players with age over 32, and players whose contract expire in 2021 and beyond are filtered out, and the rest of the players are displayed.

Ideal Replacement – combining models

First, the clustering algorithm was run to identify players of similar skills. The player and the list of potential replacements were obtained along with their predicted wages, reputation and overall.

The results obtained are sent to experts who make the final decision to retain or recruit a new player. As an example, players similar to Kiko Casilla as obtained from the model were Dea Gea, A.Begovic, and Y. Bounou. The overall score predicted for A.Begovic was the same as Kiko Casilla's overall score. But A.Begovic's wage as predicted by the model was much lesser and the reputation was much higher. Therefore, A.Begovic is suggested as the replacement player for Kiko Casilla. Another example performed on the model showed that retaining L.Modric was ideal for the team.