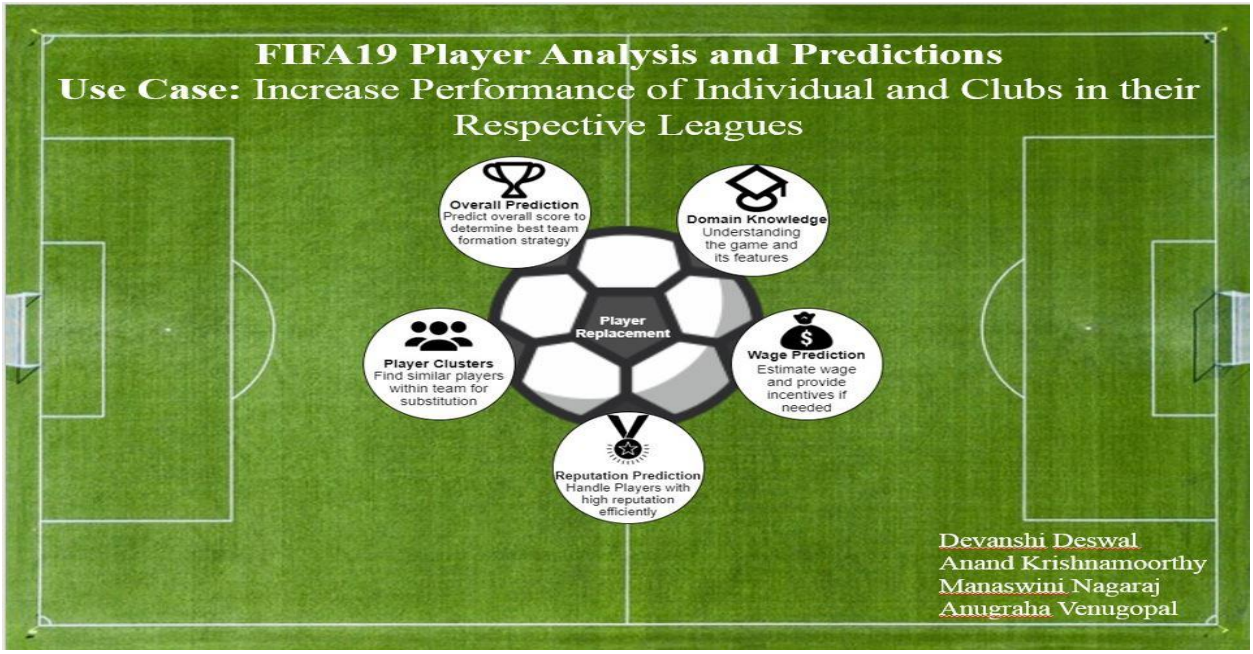


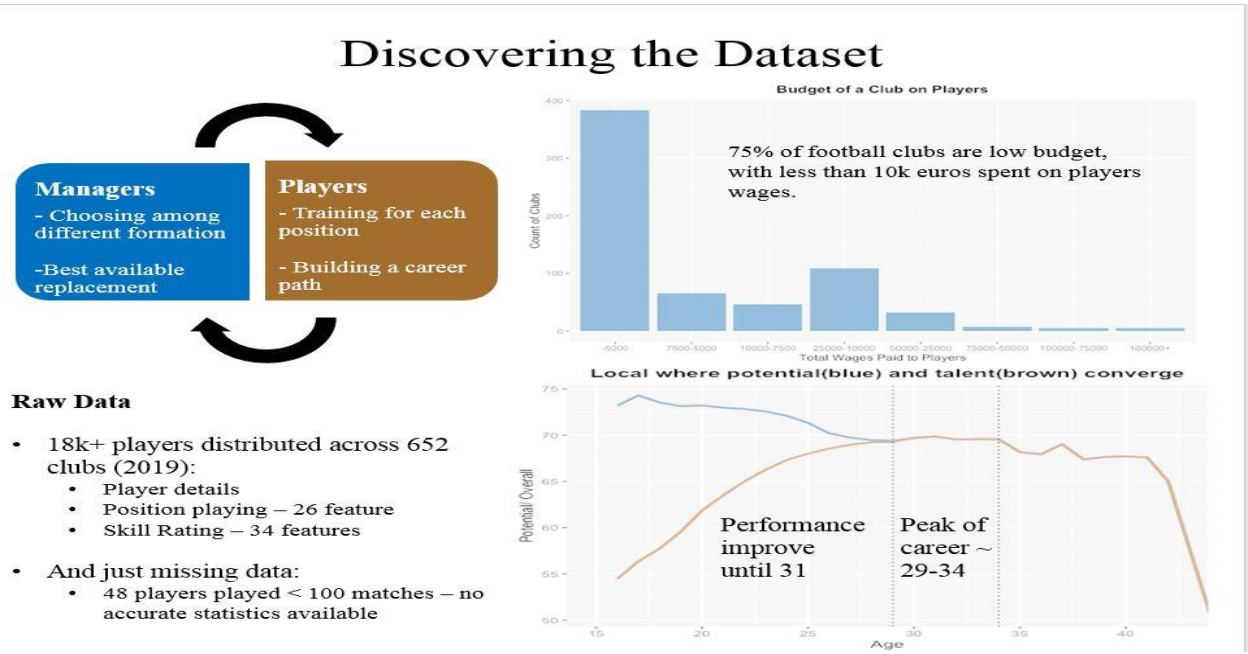
There is an avid interest in statistical methodology in sports. Within the sports business, football is a lucrative industry associated with more than half of the global population. Understanding player skills is an integral part of the analysis of these methods.

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

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

- The overall prediction of the player help determines the best team formation strategy.
- The wages for the players can be predicted for a fair contract between the player and the club
- High reputation players require additional attention and care because of their large based fan followers. Clustering of similar players can help find an ideal player replacement for players for substitution during a match.

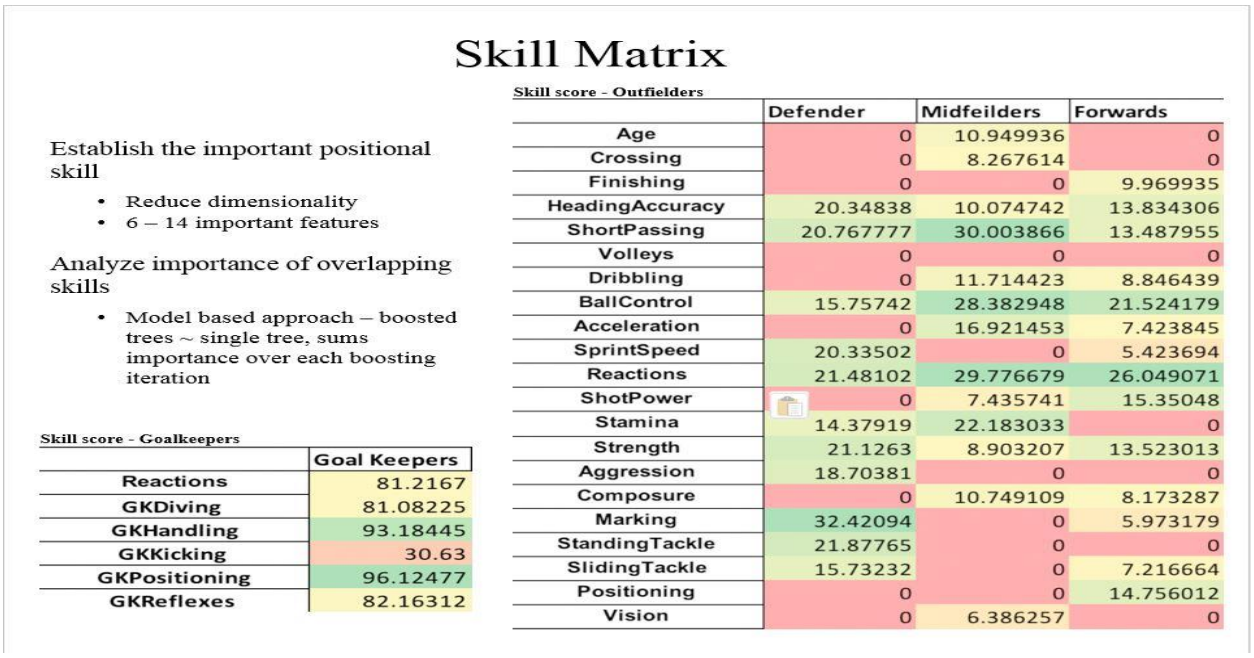
Using these predictions together along with field experts can help to make contracts with new clubs fulfilling the dynamics and the requirements of the club and the players.



The project focuses on helping the players and managers with both these current and prospects. Managers seek to identify the skill set critical for each position. It 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 their football position. For prospects, to keep the competition up and running the club focuses on making ideal player replacement considering the team's budget and requirement. While a player who is focused on building a career path for the long term would prepare/accept deals to the best of his interest

The dataset "FIFA 19 Complete Player Dataset" is publicly available on Kaggle and dwells 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 is found that these are players who have played less than 100 matches in FIFA, therefore they are removed from the analysis since proper statistics of these players could not be obtained

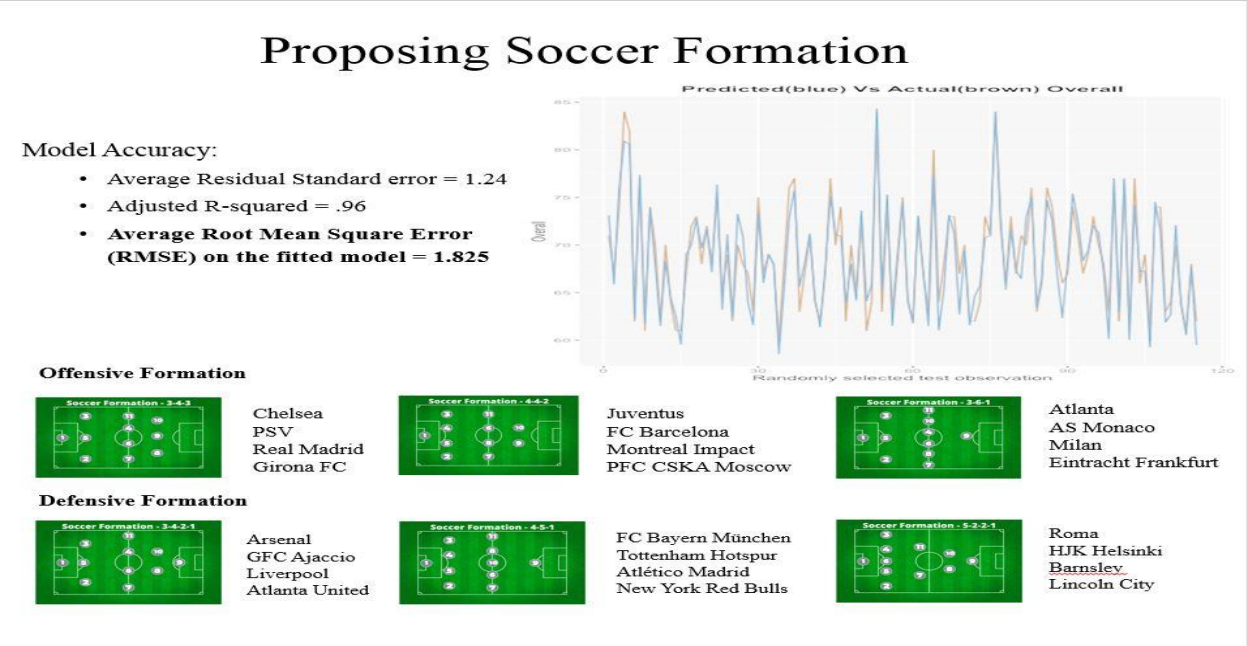
Using exploratory analysis, an interesting analysis lets us conclude that more than 75% of football clubs are low budget, with less than 10k euros spent on players' wages. Hence, the model can be greatly useful for such a club. Another interesting finding manifests the players generally improve until age 31 and then they start to decline in overall ratings. The peak of their career ranges from 29-34 years, making young players favorable over old players with similar skills.



To identify the strengths and weaknesses of the clubs to assist group predictions, the defense, mid, attack, and goalkeeping of the teams are evaluated by grouping players into these 4 positional buckets. The dimension of the dataset is reduced by 22 predictors, by grouping similar positions together in broad categories. Using the Best Subset Selection (BSS), the important skills at each position that contributes to building a strong model (and team) are identified. Using Cp and BIC graphs 11, 13, 14, 6 optimal features are identified 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).

The model-based approach is then used to find the variable importance for each skill at each position, considering the fact although Short passing may be a major skill contributor for outfield players, however, 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 is 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.

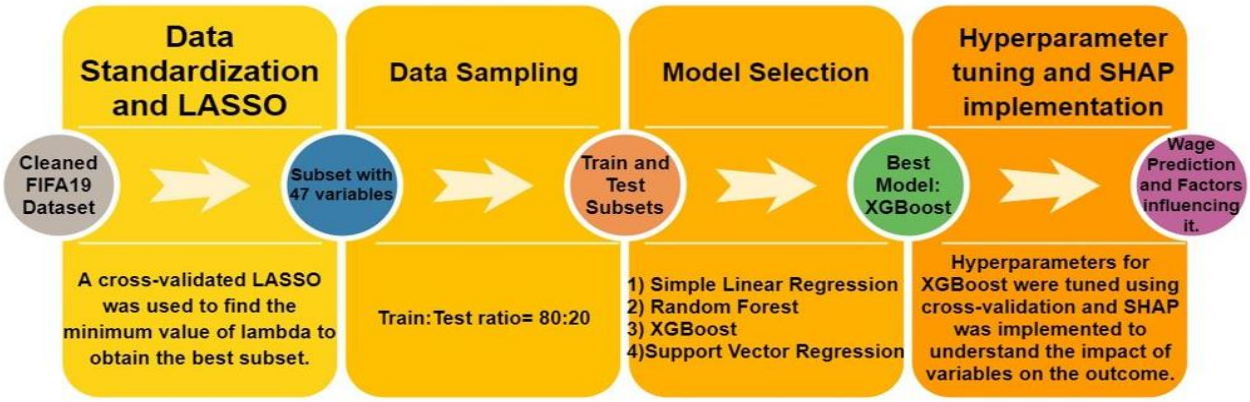


Post the best subsets regression as a screening tool to reduce a large number of the possible regression model to just a handful of models that are further evaluated to derive at one final model. Residual analysis, transforming the predictors and/or response, adding interaction terms, are performed to 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, it 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/ defense 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.

Wage Prediction for Players

- Players want to maximize wage. But team budget is a constraint!
- Build a model predicting apt wages to make the two ends meet.

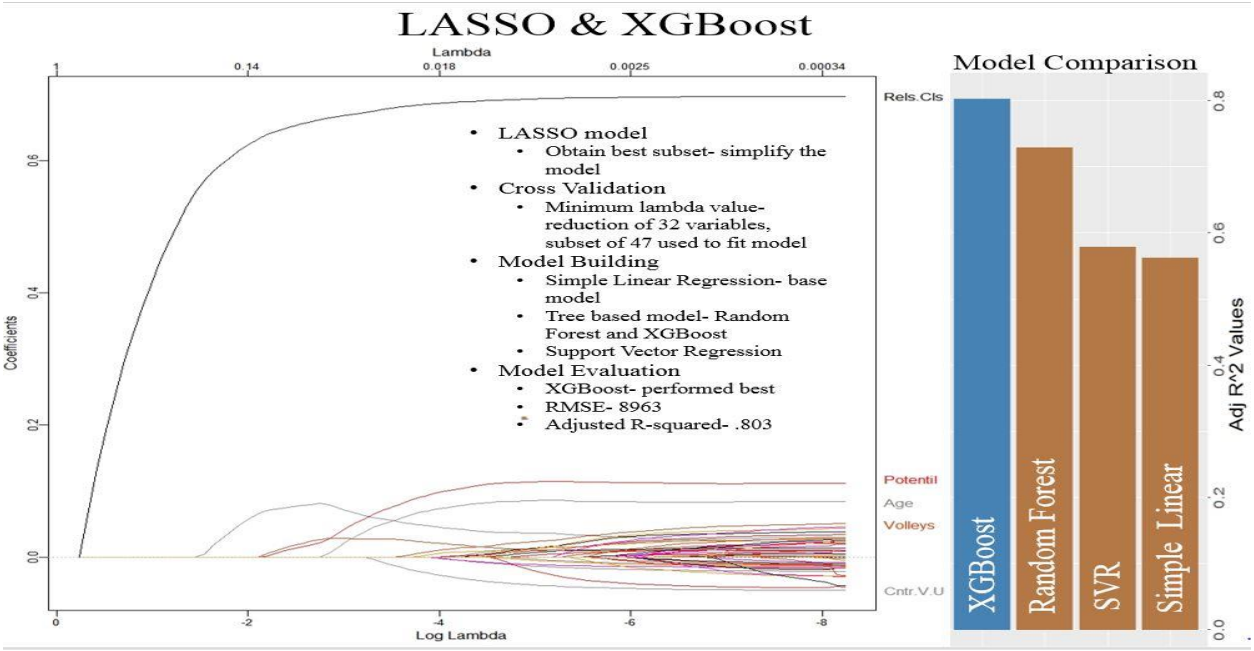


Sequence of tasks to predict the Wage of a player

The prediction of wage helps the managers set an expectation 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 is obtained it is split into train and test subset with an 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 is selected as the best one and its hyper-parameters are tuned using cross-validation to get the best out of it. Once the model is ready, it is 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.

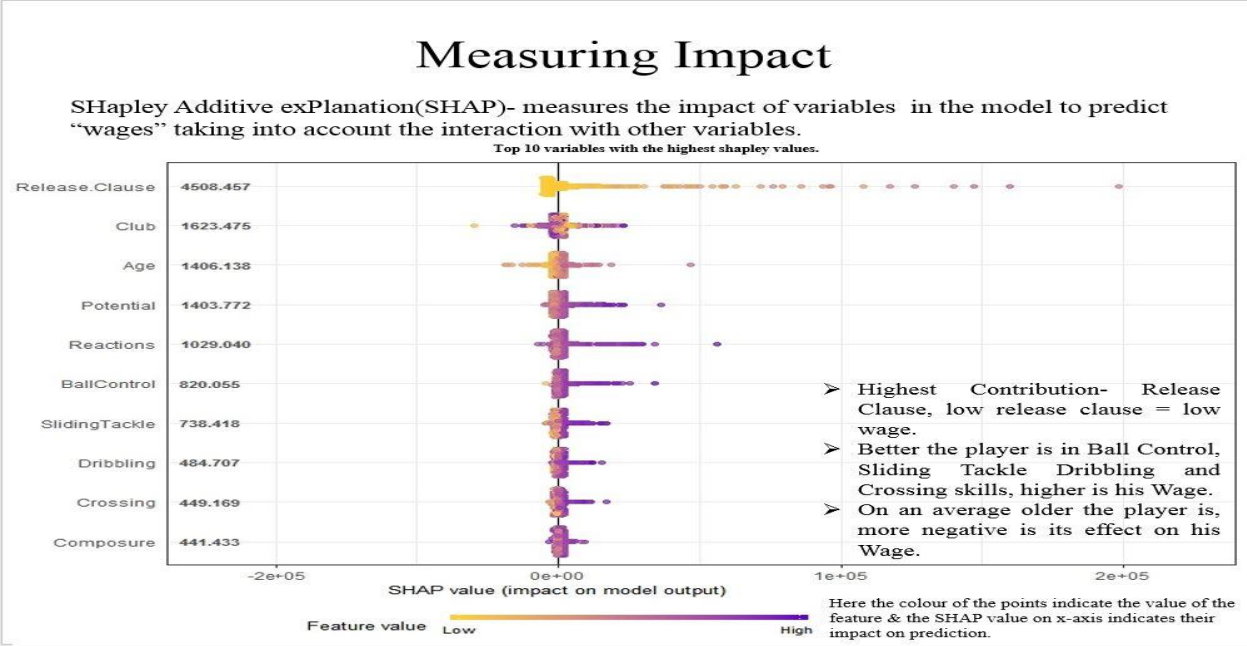


LASSO (Least Absolute Shrinkage and Selection Operator) is a regression analysis method that also performs feature selection along with regularization to enhance the interpretability of a model. Lasso uses an L1 penalty to perform subset selection, it has the effect of forcing some of the coefficient estimates to be zero when the tuning parameter lambda is sufficiently large. In this project, LASSO is 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.

With the new subset consisting of 47 features, four different models- Simple linear model, Random Forest, XGBoost and Support Vector Regressor- are applied to it. Among these models, the tree-based algorithms- Random Forest and XGBoost show 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 the new models, it is called gradient boosting when the boosting technique uses a gradient descent algorithm to minimize the loss. The models built are sequentially added until no further improvement is seen, to obtain one final model. The tree-based approach for XGBoost first builds a decision tree with weak learners, then a new model is built to predict the errors of the previous model. This process is done stage-wise and the final prediction model is in the form of an ensemble of weak prediction models.

In this project, after selecting the best model, it's hyperparameters are tuned using 10-fold cross-validation. After tuning, the model gives an RMSE of 8963 and an adjusted R-squared value of 0.8.



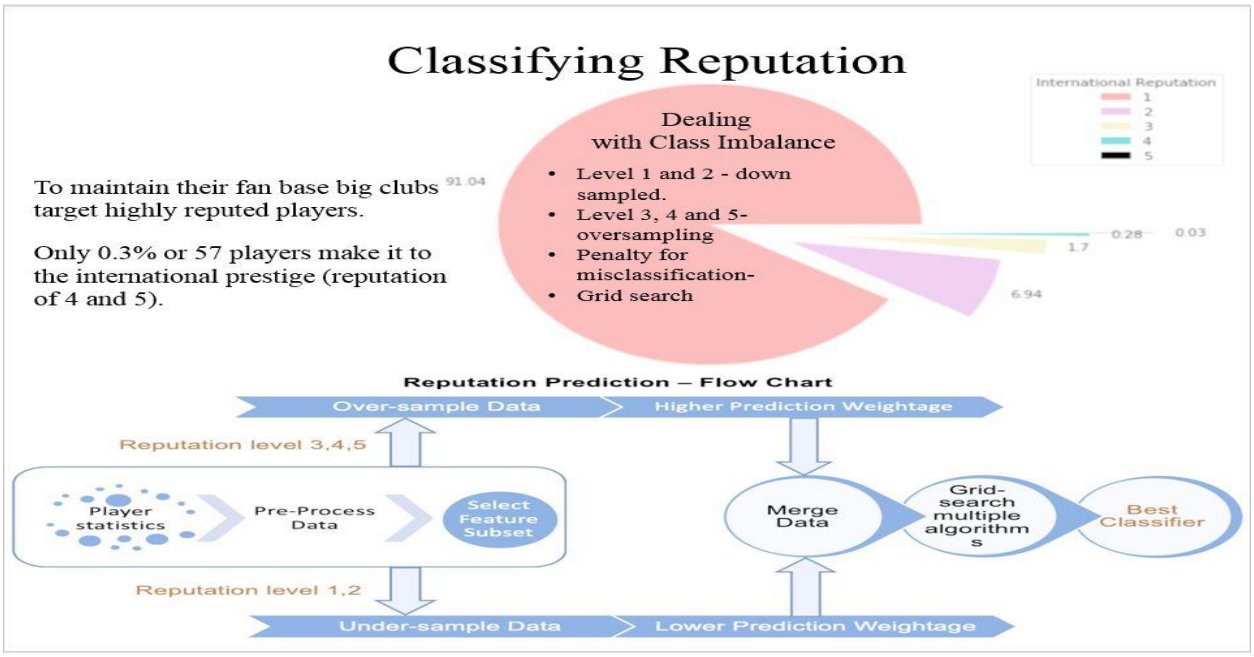
SHAP- SHapley Additive exPlanations, explains the prediction of an instance X by computing the contributions of each of the features to the predictions. To do this, SHAP computes Shapley values, which is a method from coalitional game theory, that tells us how to fairly distribute the "payout" among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

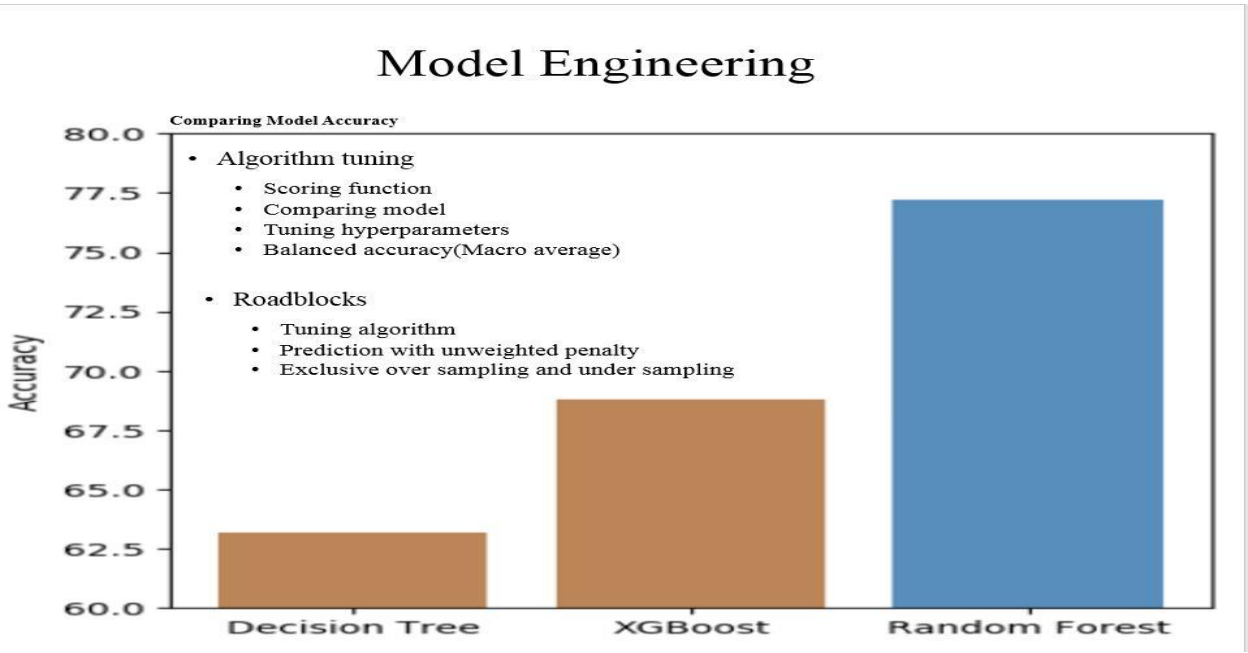
It is also interesting to note that on average, older the player is more negative the effect is on his wage.



The goal of this module is to predict the International reputation of players so that the team manager can handle player needs efficiently. The prediction can help clubs target highly reputed players which could help increase/maintain fanbase of the club.

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



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

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

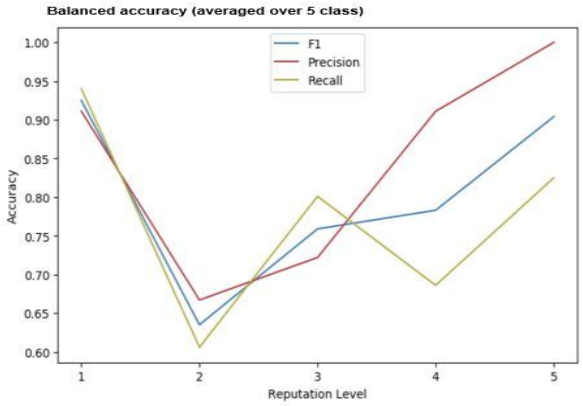
Since the data is imbalanced almost all the data points are predicted as level 1. Therefore, Oversampling and undersampling of the data points are considered. Exclusive oversampling the set 2(level 3,4, and5) data points causes the same data points to be mutated several times, which produces poor results. Similarly, exclusive Under sampling of set 1 (level 1,2 players) causes 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 undersampling is suggested.

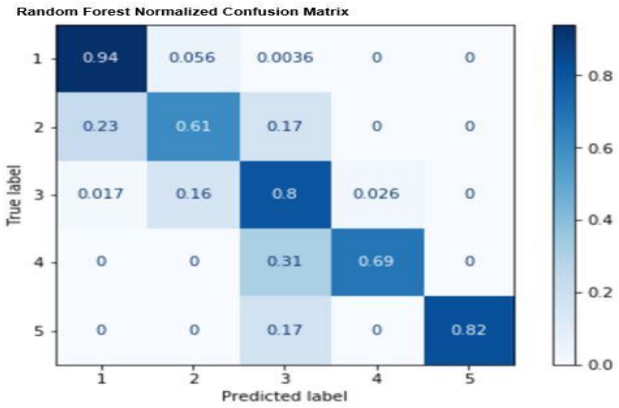
Next tuning algorithm using default scoring methods, i.e. accuracy, ROC did not produce the desired result. Therefore, a customized soring function is defined which penalized misclassification of set 2 data points more than the misclassification of set 1 data points. Additional accuracy measures like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

Model Evaluation

Analyze importance of overlapping skills for each position

- Confusion Matrix- Random forest (best model)
- Correct predictions ~ diagonals; overall accuracy of 77.2%
- Precision, Recall & F1 averages to 80.6%

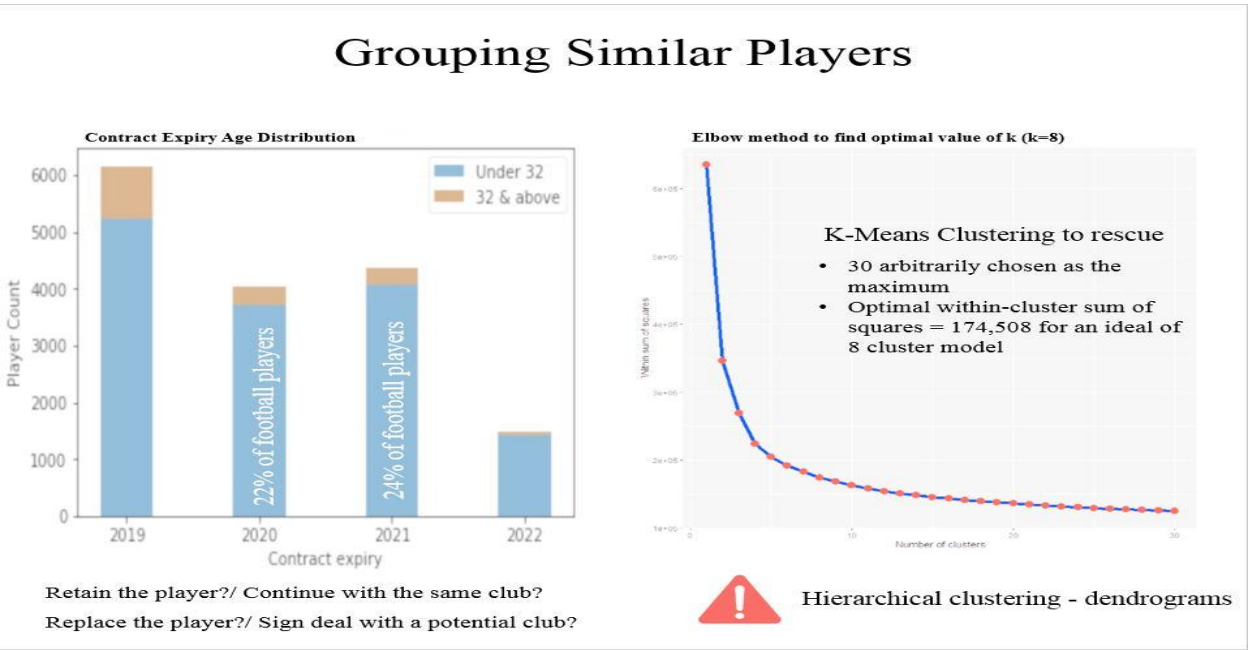




Post tuning the hyperparameter of the models, the Random forest produces the maximum accuracy. The confusion matrix of the Random Forest model is evaluated. Diagonals from top left to the bottom represent the model accuracy. Ideally, all the diagonal elements should have value 1 and the rest of the elements 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 imbalanced dataset evaluation measures like Precision, Recall and F1 scores are ideal. Precision which measures the proportion of the prediction that appears corresponding to a particular class in model output and the actual dataset while recall is the proportion of the actual number of examples in a particular class identified by the model corresponding to the particular class is plotted and evaluated for each class. F1 score combing these two scores is plotted alongside to summarize class results.

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



To provide a list of choice for replacement of a player who has matching skillset at a similar position we cluster players in similar groups based on the similarity in 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.

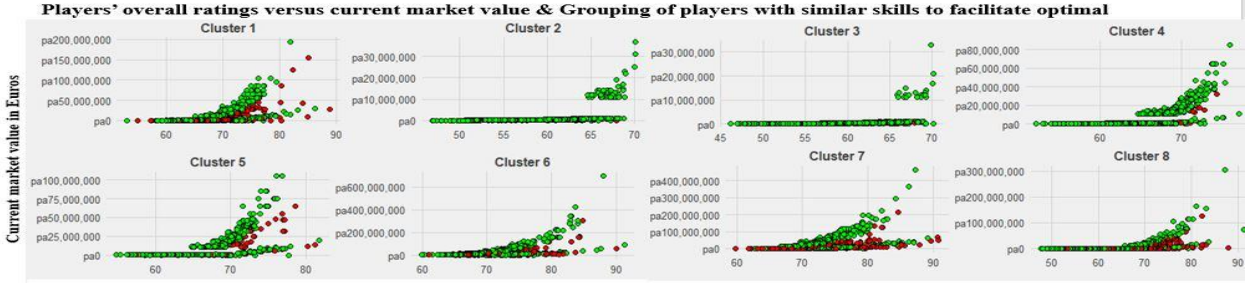
First, the relevant features, i.e., the ratings of the players for various skills, are scaled. Next, the elbow method is chosen first to find the optimal number of clusters (value of k) for k-means clustering. The 'elbow' of the plot occurs at k = 8, with a total WSS (within-cluster sum of squares) value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

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

Exposing Model to Users

Pool of replacement player - Contract valid until current year (+/-1) & Age < 33 (green)

Players' overall ratings versus current market value & Grouping of players with similar skills to facilitate optimal



Results combining modules

SL No:	Ideal Replacement	Name	Age	Postion	Clubs	Overall	Wage	Rep	Contract Valid	Comments
1	-	Kiko Casilla	31	GK	Real Madrid	79	€105K	2	2020	
	N	De Gea	27	GK	Manchester United	91	€260K	4	2020	
	Y	A. Begović	31	GK	Bournemouth	79	€52K	3	2020	Similar overall, Lower wages and a higher reputation. So good replacement
	Maybe	Y. Bounou	27	GK	Girona FC	77	€22K	1	2019	High potential, very young and athletic
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 and athletic
	N	N. Mendy	26	Mid-Field	Leicester City	75	€50K	1	2020	

The plot depicts the clustering 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 less than 32 are considered among the players whose contract expires the current or immediate years for ideal replacement due to an observation of declining skills post 34 years. The clustering algorithm is run to identify players of similar skills. The player and the list of potential replacement are obtained along with their predicted wages, reputation and overall.

The data of the player are compared by the experts to suggest the final decision to retain or recruit a new player. As examples, the model results suggest the replacement of Kiko Casilla with A.Begovic, while retaining L.Modric.