DETERMINING PLAYER POSITIONS IN SOCCER USING PLAYER ATTRIBUTES

BAX 452 Project Report

Group C

Jeet Patel

Keshore Suryanarayanan

Ravi Kiran Bachu

Executive Summary

Position determination is instrumental in defining the success of a player and how quickly a player makes it out of the youth ranks into the senior roster. Often, players lose their way in their careers because either their strongest position is difficult to determine, or they are constantly played out of position. While concretely assigning players to their strongest position is important, it should align with the rapidly evolving tactical trends in soccer.

While coaches have strong intuitions which stem from their experience and expertise, even their intuition could contain biases of what they think is the best for a player. These biases could be cognitive ("I have seen this work, so it works") and cultural ("In my country, the job of defenders is only to defend"), which could be a problematic subjectivity to have.

The project aims to provide a supplementary source of player position determination for coaches, one that either corroborates or challenges their intuition with data. We have trained multinomial and random forest classification models, using the FIFA 20 dataset, to capture the trade-offs among physical, mental and technical attributes and accurately determine a player's best position. The FIFA dataset is one of the most easily accessible quality datasets for player attributes, and is also up-to-date. The strategy of using it to predict player positions is economical and effective. We observed a satisfactory out-of-sample accuracy in both the models and the inaccurate predictions serve as useful information for coaches about the versatility of the players for whom the predictions were inaccurate.

The model is intended to take the coaches' evaluation of their players across the 38 features as input and provide predictions/classifications of position for the player. The credibility of predictions could be improved by adding match data such as win % that would reflect the success of the player in that position.

Background

One of the key responsibilities of coaches at the youth level is to determine the perfect position for their players. Position determination is a trade-off among three kinds of player characteristics - physical, mental, and technical. Players who are tall are typically coached in one of the central positions - goalkeeper, center back, central midfielder, or center forward - since their height gives them a significant advantage to thrive there. Players who exhibit positional discipline along with technical security from a very young age often end up as central midfielders. Those with excellent pace and dribbling become wingers, while those with remarkable finishing and movement become strikers.

While there is much more nuance to a player's success than simply their positional fit (role in the team, coaching philosophy, etc.), position is the first real identity of that player in the world of soccer. This is why it is important to employ scientific methods and objectivity, in augmentation with the expert eye of highly experienced coaches, to determine player position.

Traditional Approach

The ideal position and role of a player are collectively determined by several youth coaches and scouts who go through hours of video footage to determine the strengths and weaknesses of the player across the three characteristics.

The subjectivity involved at the decision-making stage poses a problem because the opinion of these coaches could be based on past stereotypes which might not be aligned with the modern developments in soccer. In the last 5 years alone, soccer has seen so many changes in its norms that had not been seen in the previous 15 years and beyond. With evolving tactics and strife to achieve perfection, a lot more is expected of players in every position than what was previously deemed enough to succeed there. It takes time for youth coaches to unlearn and relearn this, so they could benefit with an objective model that is trained with data that captures the current trends.

Research Strategy

The dataset of the most popular soccer video game, FIFA, is one such dataset in this context. It captures the current rating of each player across many physical, mental and technical metrics, while also having information about the position in which the player plays for their current club. The stakes in top-tier professional soccer mean that top-tier clubs need to stay ahead of the curve, so it is safe to assume that the players' current positions are aligned with the latest soccer trends, which could, in turn, act as a reference point for youth coaches to determine player positions.

To ensure the capture of the latest trends in the trade-off among the metrics, businesses (professional soccer clubs) could employ classification models trained with data from the latest FIFA game. When productionized, these models could iteratively train on newer versions of FIFA, thereby providing classifications that are closely aligned with the recent trends.

Analysis

Data

We have considered the FIFA dataset from Kaggle for our analysis, which has been scraped and consolidated from the publicly available website https://sofifa.com/. Our dataset consists of 18,483 rows and 106 columns. The data is at a player level, with their position in their current team as our outcome variable. This problem statement falls under multi-class classification.

Preprocessing and Feature Engineering

We have filtered a subset of the 38 independent variables out of 105 based on the compositions of physical, mental and technical characteristics of a player. Out of 18,483 rows, we observed that 10,778 rows are for position *SUB* and *RES,* which denote 'Substitute' and 'Reserve,' respectively. Since there is a limit of 11 starting players and 25 players in the squad for every club, we observe these high counts. We considered these values as 'missing data'

since they do not indicate where the player should be playing on the field. We have imputed the position for these players from a separate column that denotes the possible positions a player can play at (‘player_positions’). We have *assumed* here that the first position present in this column is the player’s strongest position.

Another *assumption* we are making is that the direction where the player plays does not matter for a given position. E.g., ‘LDM’ (Left Defensive Midfielder), ‘CDM’ (Central Defensive Midfielders) & ‘RDM’ (Right Defensive Midfielders) are all considered equal under a blanket position ‘DM’ (Defensive Midfielder). The reason for making this assumption is twofold. First, these positions in themselves are homogenous, and the skill-set required to play in them are roughly the same, irrespective of a player playing at LDM, CDM, or RDM. Second, the position of the players at being Right, Left, or Center is primarily dictated by their dominant foot and team’s formation. Hence, if a player is predicted to be a DM, they can be assigned to play at LDM, CDM, or RDM. We have ignored Goalkeeper from our analysis because goalkeepers possess very distinctive characteristics from the outfield players and in the real world, once a player is positioned as a goalkeeper, it typically does not change over time.

After following the above steps, we have reduced the number of positions of a player from 27 to 9. In the subset of 38 independent variables, we have 37 continuous variables and one categorical variable. There were no missing values in the explanatory variables, hence no further imputation was required. We have converted the categorical variable denoting preferred/dominant foot into an indicator variable. Finally, we have Label Encoded the outcome variables. We have followed an 80-20 split of the data, with 20% of the data being kept as our test set to evaluate the OOS performance. We further scale all the continuous variables into their corresponding z-values by standardizing them. We have used training data to calculate the mean and variance of the continuous variables and have used the same to standardize the test set to avoid information leakage and maintain consistency in standardization.

Model

We started the analysis with a **Multinomial Logistic Regression**, which is a special case of logistic regression where we have multiple classes as outcome variables. This is suitable for the player position as the outcome variable. Using the multinomial logistic regression, we can get probabilities of a player belonging to or being suited for a particular position. We ideally want a scenario where the model would give a good prediction for newer players (test dataset) and therefore avoiding overfit is crucial. This was made sure by using cross validation along with multinomial logistic regression. The training dataset was divided into k stratified folds and parameters were trained. Once we trained the model, we evaluated its out-of-sample performance on the test dataset.

The multinomial classification also gave us the feature importance for each class, for e.g., what it takes for an Attacking Midfielder to be one. This is a helpful insight for scouting and coaching players.

Another technique we have implemented is the **Random Forest Classifier**. Random forest is a decision tree-based ensemble learning algorithm that uses a bagging approach and majority voting to determine the class for the given data. An advantage of using random forest over other classification algorithms is that it uses different subsets of columns having bootstrapped data to predict the class of the data. This helps us in eliminating any bias that can occur in the prediction of the data by chance and can generalize well on unseen data.

We have validated the results by performing 5 fold cross-validation. To find the best hyperparameters, we have performed a grid search by experimenting with the *number of estimators*, *max_depth,* and *max_features,* further validating them by doing a 3-fold cross-validation on each combination.

Since this is a multiclass classification problem we have used metrics such as accuracy, weighted precision, weighted recall, and weighted F-1 score to evaluate our models.

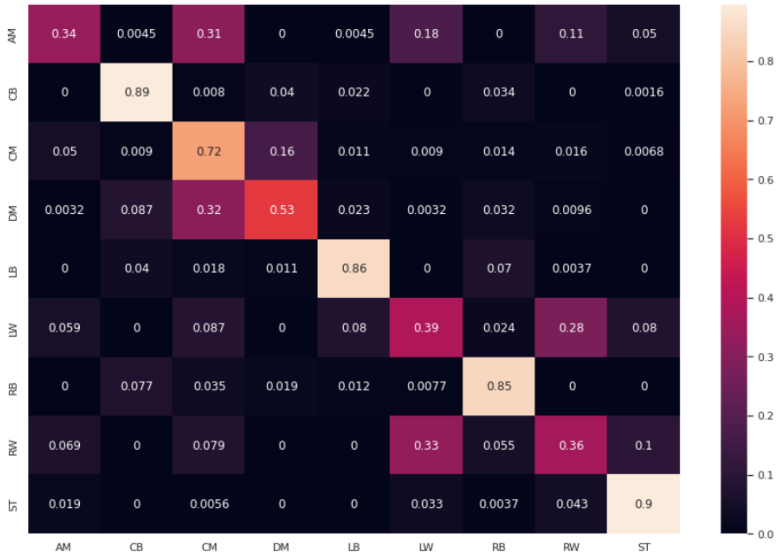
Results

Multinomial classification

Out-of-sample performance:

Performance metric	**Value**
Accuracy	69.9%
Precision	69.2%
Recall	69.9%
F1 score	69.2%

Confusion Matrix:



Based on the confusion matrix, we can observe that the model is able to predict *'ST', 'RB', 'LB'* and *'CB'* with over 85% accuracy. The model performs most poorly in predicting *'AM'*, *'DM', 'LW'* and *'RW'*. The model is unable to effectively predict *'AM'*, often confusing it with *'CM'*, *'LW'* and *'RW'*. The model also has difficulty in differentiating *'DM'* often confusing it with *'CM'*. Another difficulty the model has is in differentiating *'LW'* from *'RW'* and vice versa.

Feature Importance:

	TopVar1	TopVar2	TopVar3	TopVar4	TopVar5
AM	mentality_vision	passing	power_long_shots	attacking_short_passing	skill_ball_control
CB	defending_marking	defending	attacking_heading_accuracy	mentality_interceptions	height_cm
CM	skill_long_passing	mentality_vision	attacking_short_passing	skill_ball_control	power_stamina
DM	mentality_interceptions	skill_long_passing	defending_standing_tackle	attacking_short_passing	defending
LB	attacking_crossing	defending_sliding_tackle	defending	defending_standing_tackle	movement_sprint_speed
LW	attacking_crossing	skill_dribbling	mentality_positioning	attacking_finishing	dribbling
RB	attacking_crossing	preferred_foot	defending_sliding_tackle	defending	defending_standing_tackle
RW	attacking_crossing	mentality_positioning	dribbling	movement_sprint_speed	skill_dribbling
ST	attacking_finishing	mentality_positioning	shooting	attacking_volleys	attacking_short_passing

Since Multinomial Regression performs a logistic regression for every other category with the common base category, we are able to obtain feature importance for every position. The feature importances returned by the model align with domain theory. Vision and passing are the most defining attributes for an attacking midfielder; marking and defending are key for a center back; long passing, vision and short passing distinguish a central midfielder while the same features bar interceptions instead of vision for a defensive midfielder. Full backs (LB and RB) and wingers (LW and RW) have more and more chance creation responsibilities in recent times, which is the intuition behind crossing being the determining attribute for them. As there are more right-footed left backs than left-footed right backs, preferred foot is a key determining factor for a right back but not so for a left back. Finally, as expected, finishing and positioning are the most important attributes for a striker.

Random Forest Classification

Baseline: {n_estimators: 1000}

Fine-tuned: {n_estimators: 1750, max_depth: 56, max_features: ‘sqrt’}

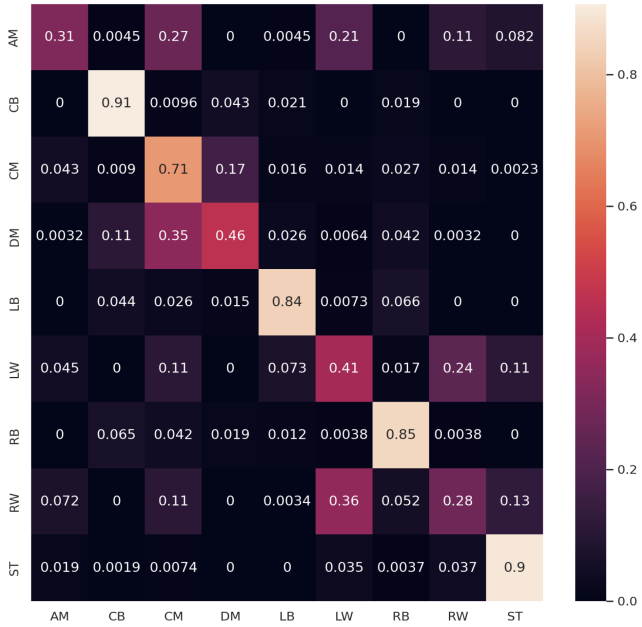
Out of Sample Performance:

Metric	Base Configuration	Base (Cross-Validated)	Fine-tuned	Fine-tuned (Cross-Validated)
Accuracy	68.4%	69.1%	69.0%	69.2%

Precision (Weighted)	67.3%	67.8%	67.9%	67.9%
Recall (Weighted)	68.4%	69.13%	69.0%	69.2%
F1-Score (Weighted)	67.3%	68.1%	67.9%	68.2%

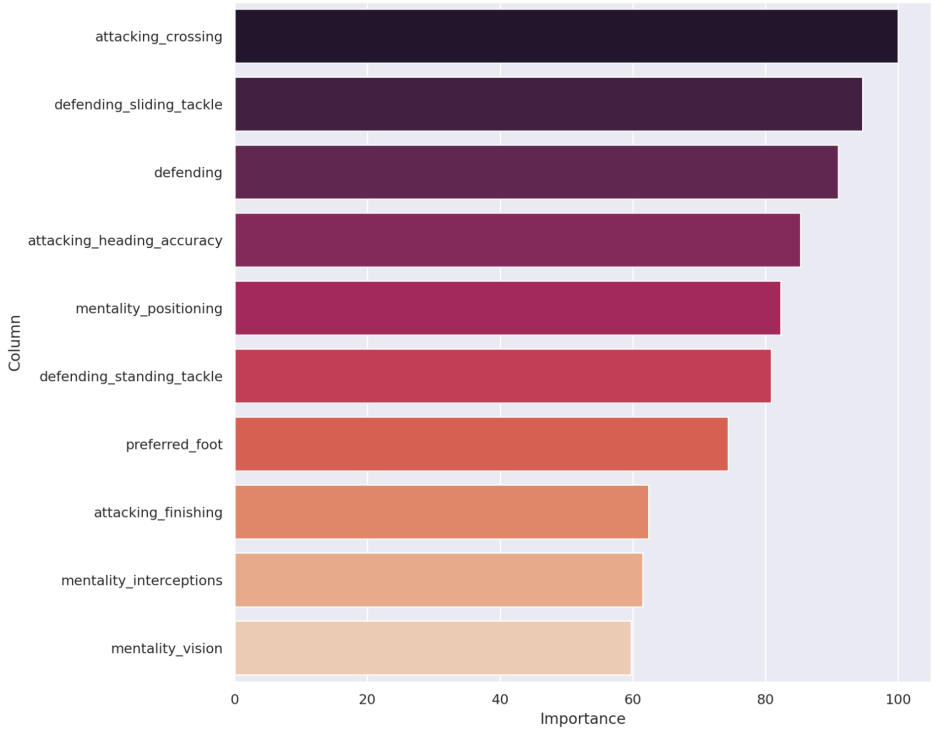
We can see from the evaluation metrics that the cross-validated baseline Random Forest model with no hyperparameter tuning is able to perform as well as the cross-validated Random Forest model fine-tined for *n_estimator, max_depth*, and *max_features.* This is in line with the expected behavior of Random forest being able to generalize well because of using multiple trees, reducing the variance in the Out-of-Sample Predictions.

Confusion Matrix:



As observed in Multinomial regression, Random Forest is also able to identify ‘LB’, ‘RB’, ‘ST’ and ‘CB’ with over 85% accuracy. Similar to Multinomial, Random Forest comes up short in distinctly identifying ‘LW’, ‘RW’, ‘AM’, ‘CM’ and ‘DM’.

Feature Importance (Relative):



Applying soccer theory, the ‘confusions’ observed in both the cases of Multinomial regression as well as Random Forest are perfectly alright to live with because attacking midfielders often double up as a central midfielder or a winger due to possessing the required characteristics to succeed in those positions. Similarly, defensive midfielders operate as central midfielders depending upon the system and philosophy employed by the club. Finally, it is common practice for wingers to interchange between LW and RW even within a game, let alone across games. Since the model is intended to be a supplementary tool for coaches, they can easily identify these nuances and derive value out of even the seemingly inaccurate predictions.

Recommendations

At the moment, coaches determine player positions through intuition from watching the players play in multiple positions and rewatching them again through video footage. It is recommended that the coaches input their evaluations of players across all the 38 attributes to the model, so that they receive suggestions for ideal positions to play these players in. The

model could eliminate any subjectivity and historical biases in position determination. Functionally, the model could at best be decisive and at worst be directive, which is still a win for the relatively less-privileged youth setups that nonetheless produce the best of talents.

The model could be further improved by adding match data, such as win percentages of players when they played in the specified team position (normalized by the team's win percentages to level the playing ground for all players), so that there is an element of how successful the player is by playing in that position. This would validate the trade-offs implemented and add further credibility to the predicted positions.

Conclusion

To achieve the objective, we used the FIFA 20 dataset and trained classification models using the multinomial classifier and the random forest classifier. Both the classifiers have an out-of-sample accuracy of 69%. We also did not observe an uptick in the model performance of the random forest classifier after hyperparameter tuning. The out-of-sample accuracy of both the classifiers, through their respective confusion matrices, is observed to be negatively skewed by inaccurate predictions for wingers, central midfielders, defensive midfielders and attacking midfielders. However, these inaccuracies are healthy for the model to have because it means it is capturing the commonalities in the decisive attributes for these positions, which can also help determine the versatility of a player.

Overall, the strategy of building multinomial or random forest classifiers trained with the data from the latest version of FIFA is a cost-effective way to determine player positions for coaches especially in the youth setups of underprivileged clubs.