**DATS 6101**

Project 2

FIFA Player Analysis

**TEAM:**

David Amankwah

Lydia Jin

Avijeet Kartikay

Nhung Nguyen

**I. Introduction**

There have been numerous analyses performed on online games, as some games are based on real sports, including FIFA 18. This report will explore specific playing abilities in each individual, to evaluate whether these features can predict the probability of an individual becoming a particular position in a soccer match: attacker, defense, mid-field, or goalkeeper. To create a suitable model to predict soccer positions, this report will utilize different approaches, which are KNN algorithm and logit models, and compare the benefits and drawbacks of each approach.

**II. Literature Review**

David Lao uses different Python packages to predict players’ position in the FIFA 18 online game. In order to deal with the numerous features that the game reported for each individual player, Lao focused on a majority of features, and then normalized the data to deal with outliers (Lao, 2018). Lao utilized different approaches, including logistic regression, PCA, and neural network machine learning algorithms to predict the player positions (Lao, 2018).

There have also been multiple analyses on how to predict player position in other sports besides soccer. In “Predicting Player Position for Talent Identification in Association Football,” Razali, Mustapha, Ahmad Yatim, & Ab Aziz (2017) utilized “Bayesian Networks, Decision Trees, and K-Nearest Neighbor” to evaluate how individuals can fit into specific positions “based on players' individual qualities; physical, mental, and technical” (Razali et al., 2017).

**III. Methodology**

For the prediction of the player’s position, we use logistic regression, a binary classifier, to predict the target label for each class using four individual models. Creating four models for each class is computationally exhaustive, hence, we also use the One vs. All classification approach. According to One vs. All, one position is predicted at a time. So the samples with predicted class labels that are not the target class, will then be used to predict the next target class, and so on.

As there is an imbalance among the classes, the model can predict the highest density label for all samples and still produce results with a high accuracy. We consider predicting all the class labels at once using a multi-class classifier, that is, KNN classifier. As the player attributes are closely related and the player position groups exhibit similarity in groups, i.e, attacking players are required to have similar attacking qualities , goalkeepers have similar goalkeeping qualities and so on, hence, the KNN Classification algorithm is good choice as it identifies clusters based on the similarity of sample’s attributes. Consequently it was observed that KNN displays a high accuracy and is our choice for this classification problem.

**IV. Dataset**

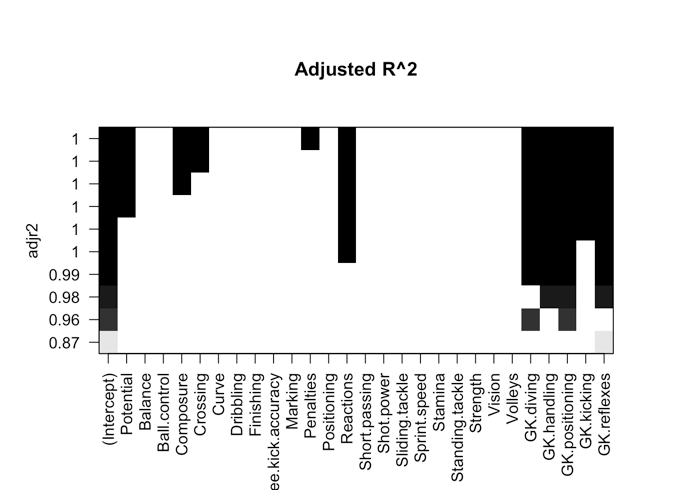
The FIFA 18 Complete Player dataset was collected by Aman Shrivastava, and made available on the Kaggle website. The dataset describes the various game-related features of a soccer player and player’s basic information including physical strength, players’ style, and nationality. The dataset consists of 17,981 observations and 74 attributes.

There are few limitations to our dataset. The player style statistics in the dataset include character type data (e.g. “75+2”). When converting these attribute values to numerical type data, they become NaN values, and are excluded from the summary statistic - mean, median, standard deviation - calculations. Since there are more than 1000 observations containing this type of character data, rather than dropping these observations altogether, we simply keep the leading number (e.g. 75).

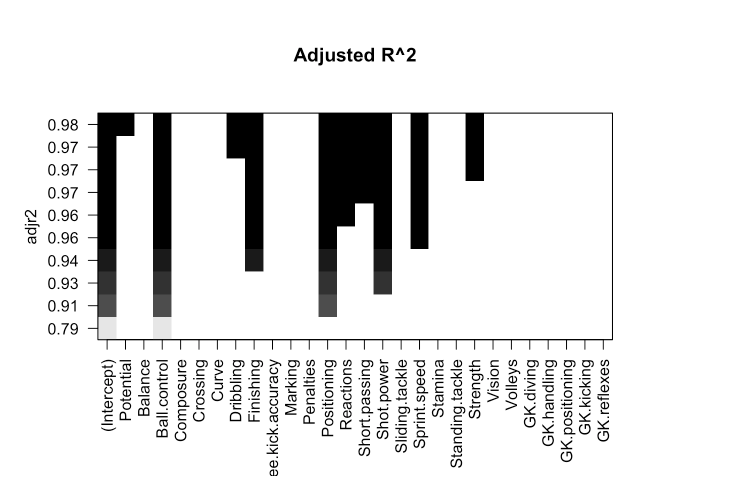
**V. Result & Interpretation**

**Logistic regression Modelling results predicting for Attack & Goalkeeper**

Using feature selection we find the most important variables for each position, that is, for Attack and Goalkeeper. For goalkeeping position we narrow down the feature to 10 features and use a logistic regression model. With a Macfadden statistic of 0.92, we notice that the model explains 92% of the data.



Using the same approach, feature selection renders 10 features for the Attack position data model. We again use a logistic regression model that has a Macfadden statistic of 0.86, which means that the model explains 86% of the data. Though these accuracies are good, we can not rely on these values as there is a Class Imbalance problem.

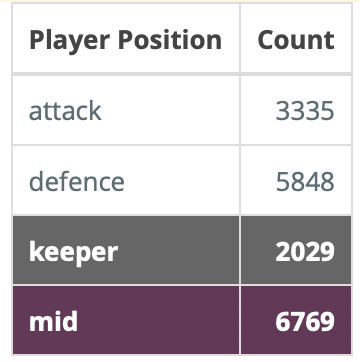


The class imbalance problem is the situation when the model is trained to predominantly predict one value, that it only predicts that label. With the mid and defence labels prominent in the data, a model with a high accuracy predicts the wrong label for the minority classes, especially keeper. We use a different approach to solve this problem. The One Vs. All multi-class classification uses a binary classifier to predict more than 2 classes.

**Logistic Regression Modeling results predicting all four positions**

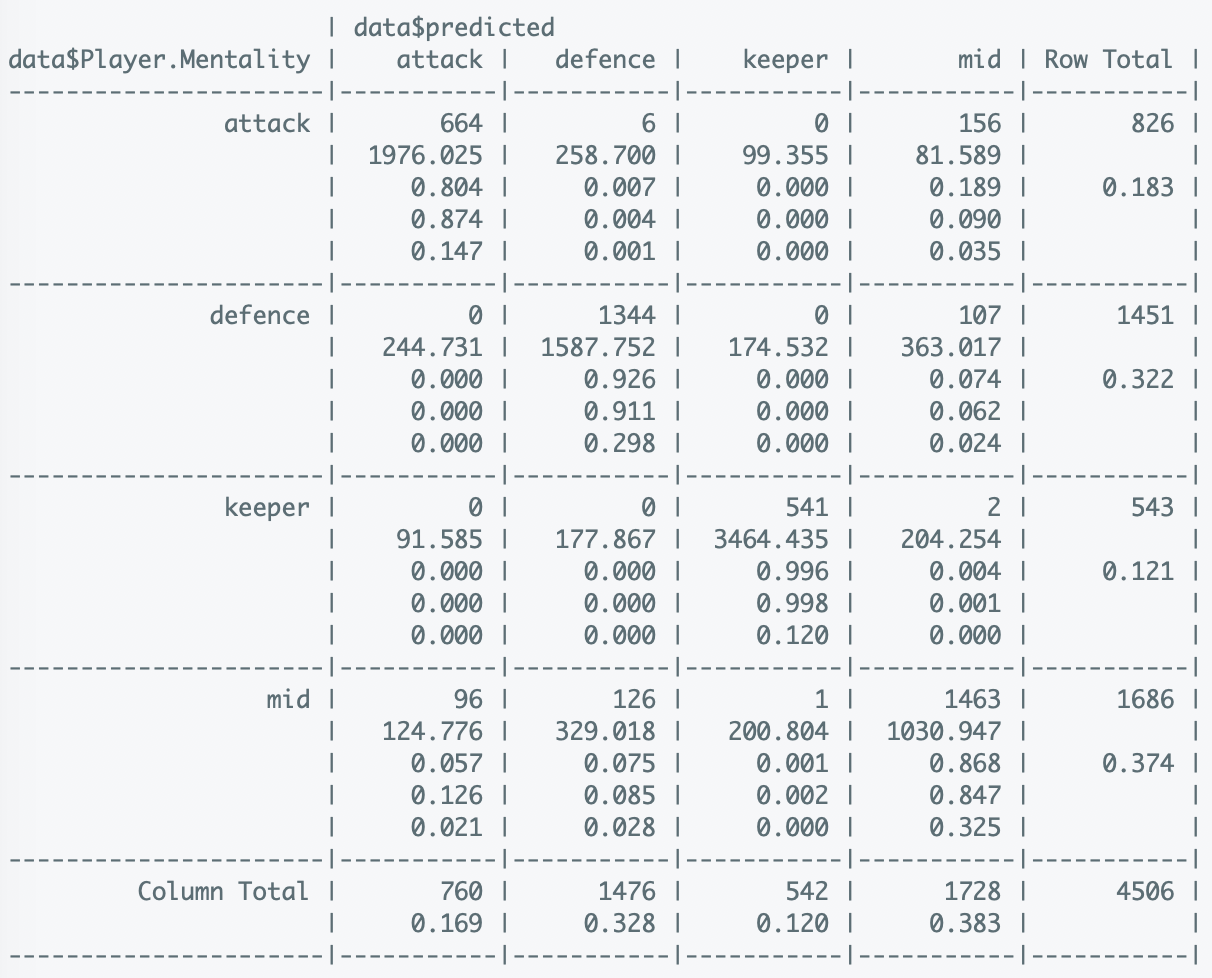
For comparison purposes, we turned to logistic regression modeling, but realized that predicting one class label against three others greatly skewed our data. In the case where we predicted goalkeeper or other, there were only approximately 2,000 out of 17,000 players for a goalkeeper position. Accordingly, the results of the logistic regression modeling we found are not dependable. To address this problem, we recursively ran logistic regression and feature selection.

During individual logistic regression, we found that goalkeeper had the highest predictive accuracy, attack and defence had relatively similar accuracy, and midfield had the lowest accuracy.



The table above, gives us the number of players for each position. We can see that there are much less goalkeepers, while there are many midfielders. Goalkeepers have more specific predicting features, so in our complete logistic model, it makes sense that we start with keeper. On the other hand, mid has the greatest number of players, and they share many of the same predicting features with the attack and defence positions, so mid is the least accurate. The order in which we iteratively performed logistic regression was keeper, defence, attack, and then mid.

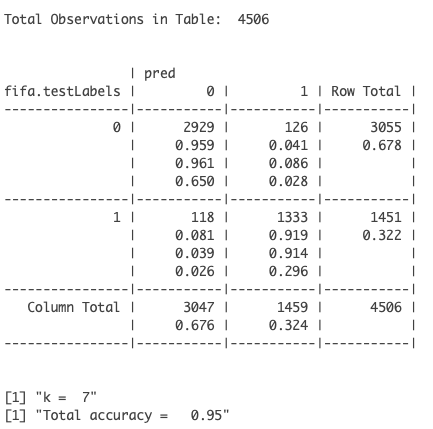
After splitting our data into a training and testing datasets, we performed feature selection and created a logistic model using the training dataset, before predicting goalkeeper or ‘other’ from the testing dataset. Then taking a subset of the observations found to be ‘other’, we use that as a new training dataset to perform feature selection and create another logistic model, before predicting defence or ‘other’ from the testing dataset. After one more iteration of feature selection and logistic regression modeling for attack or ‘mid’, we can compare the predicted results with the actual player positions of the testing set. The accuracy we received for this type of recursive logistic regression model is approximately 89%. A cursory glance at our predicted against actual class labels shows that most incorrect predictions occurred as a result of the midfield position. This can be seen even more clearly in the following CrossTable.



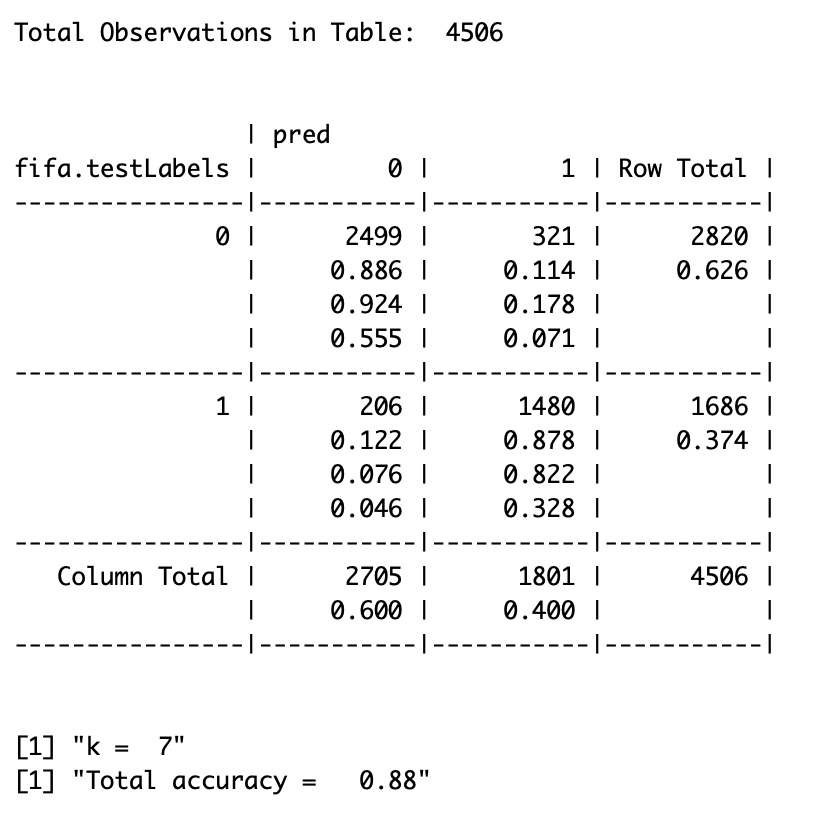
This Cross Table allows us to make comparisons to our KNN results. Two important aspects to note, are that keeper has the largest prediction accuracy. For all the samples predicted keeper, only one sample was incorrect. Next, we can see that the majority of the remaining incorrect predictions are mostly related to the midfield position. As expected, mid is the weakest predictor of logistic regression.

**K-NN results predicting defence and midfield position**

Next are the results from the KNN approach to predict the positions, defence and midfield. First, regarding the table below, the numeric values 0 and 1 represent the other and defence positions respectively. With the KNN predictions of defence, 1333 out of 1451 are correctly predicted, using the original predetermined positions as a comparison. With a k-value of 7, this prediction approach has a total accuracy of 95%.

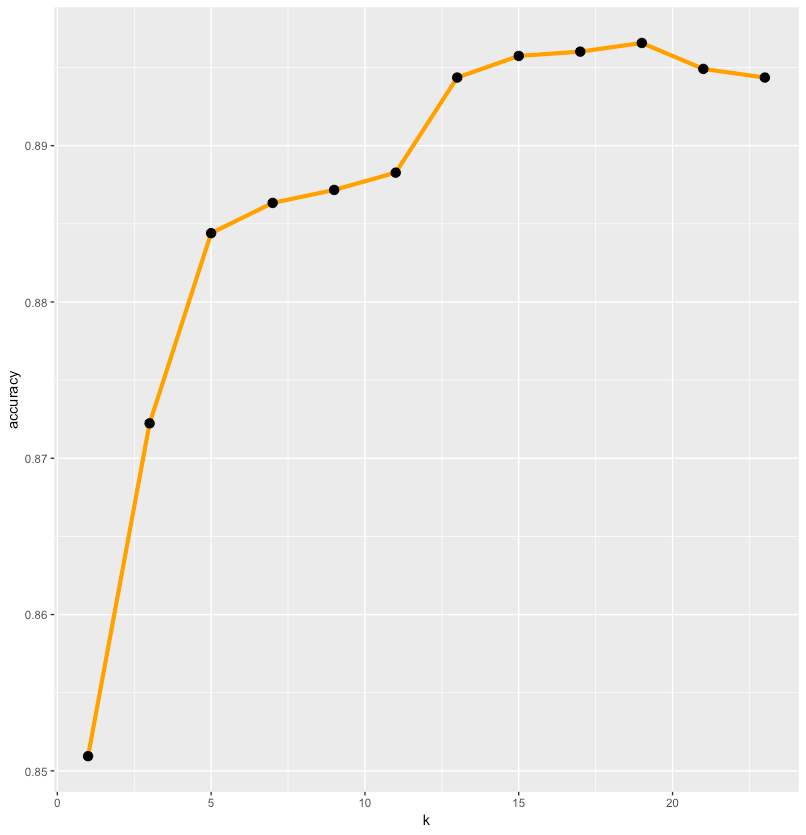


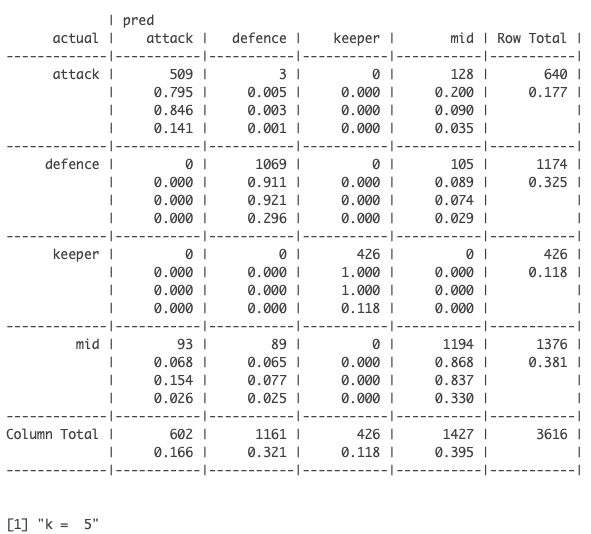
Below is the table with the prediction of the position midfield where the numeric values 0 and 1 represent the other and midfield positions respectively. The other position is a collection of all the alternative positions which are not midfield. From the results, 1480 out of 1686 are correctly predicted to be in the midfield position, in comparison to the predetermined positions of the soccer players in the original data, taking into account their individual attributes. The total accuracy for the prediction of this position is 88%, with a k-value of 7.



**K-NN results predicting all four positions**

Besides using KNN approach to predict individual position, this analysis also try to use the same approach to predict all four positions. Below is the accuracy of the classification for each specific k. The accuracy increases as the k-value increases, until k is 15. After k=15, the accuracy is improved by very little. After k=19, the accuracy starts to decrease, instead of increase.



Based on the result from the Cross table (which is showed below) with the focus on minimizing the number of false predictions and choosing a relatively high accuracy, a k of 5 is chosen. As a result, the KNN approach leads to choosing a classification with 5 neighbors. With k =5, the algorithm has an accuracy of 88.5 %. This result means that the k-NN algorithm is based on the chosen 5 closest neighbor’s labels of defence, attack, mid-field, and goalkeeper to assign a label to a specific point. Below is a detailed table showing the predicted result from the algorithm against the actual from the test dataset. 

Using k =5, the algorithm can correctly predict the followings:

1. 80% for the attack position
2. 91% for the defender position
3. 100% for the goalkeeper position
4. 87% for the mid-field position

**VI. Conclusion**

To re-summarize our process, we began by using KNN for the individual player positions. Then we proceeded to individual logistic regression, a recursively applied logistic regression, then finally a complete KNN for predicting 4 class labels at once. We have decided to select the KNN with k=5 as the more reliable classification method over the slightly more accurate logistic regression model. KNN allows us to classify data points based on feature similarity for groups, which fits our data, where we hope to group based on player position. Also, as KNN is a learning algorithm, we find this to be more robust than a more rigid logistic regression model.

**References**

Lao, D. (2018, July 27). FIFA 18 - Predict Player’s Positions. Retrieved October 20, 2018 from

<https://www.kaggle.com/laowingkin/fifa-18-predict-player-s-positions>

Razali, N. , Mustapha, A., Ahmad Yatim, F. & Ab Aziz, R. (2017, August). Predicting

Player Position for Talent Identification in Association Football. IOP Conference Series:

Materials Science and Engineering. 226. 012087. 10.1088/1757-899X/226/1/012087.

Original dataset: <https://www.kaggle.com/thec03u5/fifa-18-demo-player-dataset>