Predicting the FIFA World Cup

[1]Taher B Dossaji, [2]Ujwal KV

[1, 2]*Department of Computer Science and Engineering, PES University, RR Campus.*

[1]taherbdossaji1102@gmail.com

[2]ujwalkv@gmail.com

***Abstract*— Football is the biggest sport on this planet. It is spectated globally on the largest scale, with millions of people tuning in to support their favourite teams. The FIFA World Cup is the biggest spectacle this game has to offer, and by far one of the largest sporting events held worldwide. In this project we aim to analyze and predict the outcomes of the world's biggest, and most unpredictable tournament. Predicting the outcome of a sport is close to impossible, simply because of the monumental number of variables that could affect the outcome of any given game. It takes only a fraction of a second for a team that is considered a favourite to win to produce an unfathomable error. Taking into account these uncertainties, it is hard to get any machine learning algorithm that can predict any given game with an extremely high accuracy.**

The solution we have found to attain the required prediction includes a boosting and random forest model which works on data collated on the EA sports FIFA 22 game as well as the history of every previous international game played to achieve the required prediction.

***Keywords*— Data Analytics, FIFA World Cup, Ensemble Model, Boosting, Football Analytics**

I. Introduction

The world of sports is today more data-driven than ever before. The field of studying football related data including but not limited to the study of players, tactics, opponents and training has grown immensely in the last few years. The rise of technology has resulted in every modern-era football decision being taken based on data acquired from past games. It is important to realize the extreme advantage a good analytics department in a particular team can make to a particular team's overall performance.

The FIFA World Cup is the biggest single-sport competition in the world. After a preliminary competition, the 32 qualified men's national football teams compete to become world champions in a final competition staged during one month in a host country selected by FIFA.

It is anticipated that over one million spectators will attend the tournament's 64 matches, and the competition will reach a global in-home television audience of over 3 billion people[1], with more than one billion fans tuning in to watch the final match. In addition to the matches, there are a host of other official competition-related events, including draws, team and referee seminars and workshops, opening and closing ceremonies, award ceremonies, cultural events, press conferences and launch events.

The scale of this tournament makes it a feast for data analytics, and predicting it is one of the most daunting and challenging tasks possible. In this paper we will aim to analyze and predict the entirety of the world cup, right from the group stages onwards and aims to give an insight to the common football fan the strengths and weaknesses of each team arriving at the tournament and the possible line-ups each team could produce based on their current players. We also aim to find similar players to a certain given player based on how similar their attributes and positions are to find the best players that play a particular position or role on the football pitch.

The paper contains the following upcoming sections in order: II. Related Works; Literature survey corresponding to our problem statement, III. Proposed Methodology; Dataset description, preprocessing, descriptive analysis, training and testing, details on model building and methodologies, IV. Results and Conclusion; Experimental results of models and inferences made along with concluding remarks have been detailed.

II. Related Works

To create a solution for this problem statement, we do not have access to any existing publications that have predicted the FIFA World Cup yet, but we can use a case-study of other prediction models from other sports such as cricket and the Olympics.

The first task faced with any data analytics project involves acquiring the right data and ensuring it is cleaned. To ensure the sanctity of the data being used in the models, a stringent procedure to make our data usable is carried out and the data is transformed based on its ordinal and numeric features to fit the model in use. We follow an enhanced data cleaning technique as specified in a publication[2] to ensure the data is capable of providing accurate results once cleaned and scaled.

Understanding the various factors that can play into affecting the outcome of a football match is an inherently complex and challenging task. To carry out this task effectively and with higher accuracy, we need to resort to

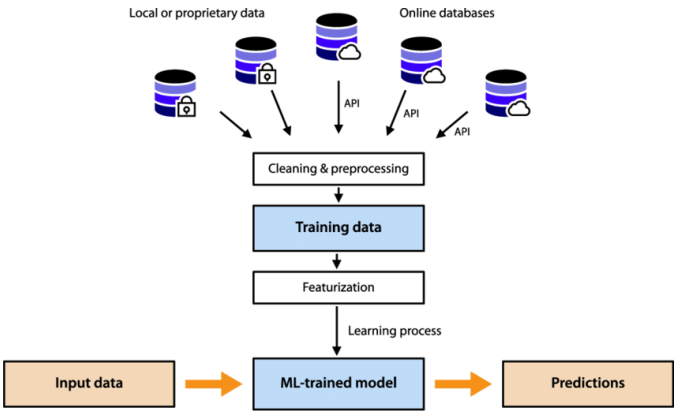
complex deep-learning algorithms and some supervised learning algorithms such as Naïve-Bayes algorithm[3].

J. Hucaljuk and A. Rakipovid have developed a software system that will be able to anticipate the result of the Champions League with about 60 percent precision. To address this, they have first opted for feature selection which is the most important step for predictions. They have selected the following features: the current form of teams shown based on results obtained in the last six months, the result of the previous game-playing squad meeting, the latest ranking place, the number of disabled players on the first squad, the total number of goals earned and earned per season. They performed a series of tests to find the optimal mixture of sets [4].

A generalized statistical software for forecasting the outcome of the English Premier League has been demonstrated by Baboota, Rahul & Kaur, Harleen. (2018). Using software development and exploratory data processing, they have created a software collection to evaluate the most significant variables for forecasting the outcome of a football match and, as a result, create a highly accurate predictive algorithm using machine learning [5]. Their best gradient-boosting model achieved a result of 0.2156 on the graded likelihood score (RPS) metric for Game Weeks 6 to 38 for the EPL aggregated over two seasons (2014–2015 and 2015–2016), while the betting organizations we find (Bet365 and Pinnacle Sports) achieved an RPS rating of 0.2012 for the same period.

III. Proposed Methodology

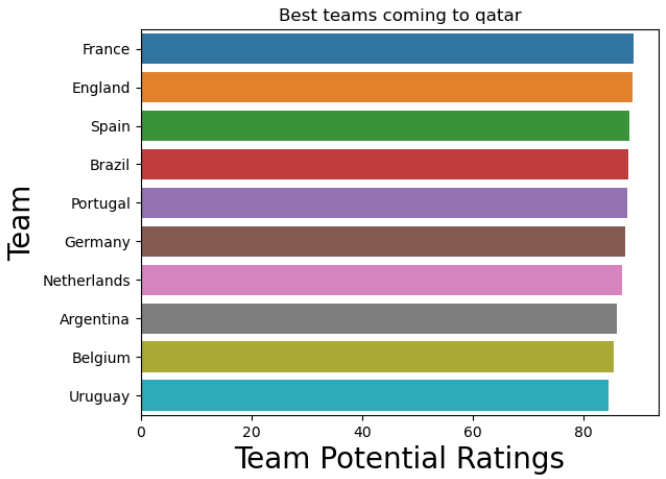
Football matches are predetermined in length, with a team either winning, losing, or drawing. The aims of the teams are less clear today since they adopt a defensive strategy, but in earlier models, algorithms had been developed to determine each team's potential with the help of which outcomes had been obtained. [6]. The impact of the player on the game, or whether the team's best player is in the starting lineup, can also be a deciding factor. The position of each player and how that player may impact the team's overall performance and alter the outcome are discussed in this paper along with the team outcomes.

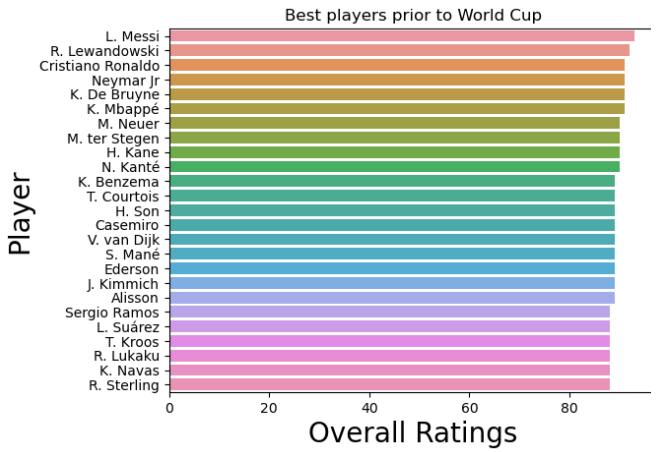


A. The Datasets

To carry out the analysis required and create our models, we will be using three datasets as input mainly. These are the FIFA 22 players [9] and international matches from 1872 to 2022 [8] datasets as well as a simple dataset created of the teams at the world cup grouped by their respective world cup groups decided at the FIFA World Cup draw conducted several months prior to the tournament.

The Players_22 dataset contains 110 features of each player in the FIFA 22 game. To clean this data it is required of us to select the most important attributes to us. Our initial cleaning is decided by choosing the most important features such as the player's overall rating and their market value to decide the best players going to the world cup. This data is further extrapolated to collate the players and group them by nationality to produce the best possible team going to the world cup this winter. There is also an alternate method of producing best possible teams which involves creating the database of best possible teams by looking at the players potential rating in the future. Assuming that some players tend to step up at the biggest of stages, several players may perform better than expected and hence, an 11 consisting of the players best potential is generated.





Analyzing this data, it is observed that France and Brazil are two of the strongest teams heading into the tournament purely based on the ability of the squad they possess. This rating for each team generated will be a vital input to the models we will cover in the future sections.

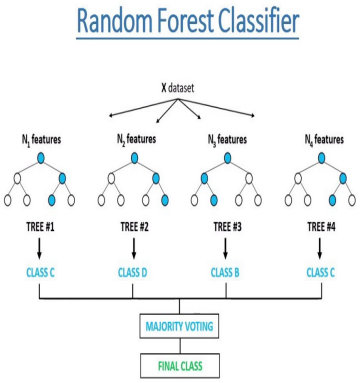
The second dataset used will be the history of all international games played prior to the world cup, dating back to 1872. Studying the history gives a rich insight to the footballing heritage some of the teams possess, arriving at Qatar. This data from the international matches dataset is used to visualize the form of teams, their ability in each zone of the pitch based on their squads abilities and rank the teams based on their FIFA ranking. The win percentages of each team at the world cup is analyzed and processed based on their home and away games to understand if a team performs better when they have the home advantage. An offense, defense and midfield score is generated for all teams.

B. Training and Testing

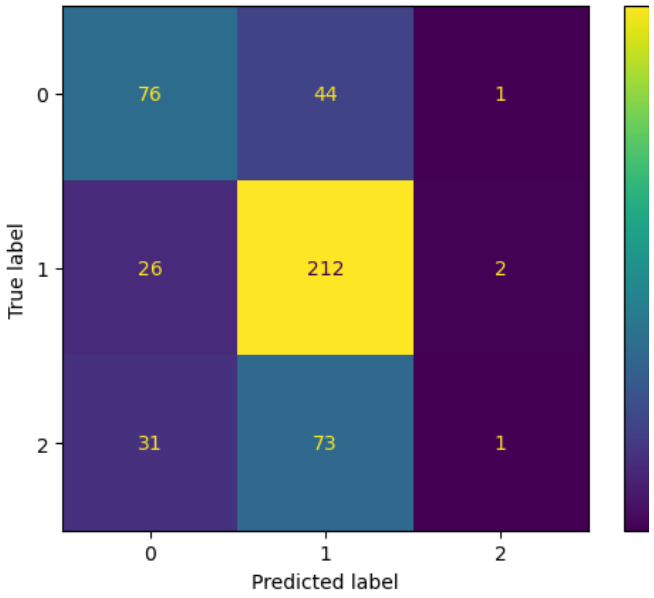
The cleaned and preprocessed data was used to train and predict using four machine learning algorithms that were intuitively implemented. These algorithms had to classify two sets of predictions; one was to predict the outcome of the group stages which is a system of each team playing every other team in their respective groups once and winners accumulate points. The group stage allowed for the possibility of a draw between the teams if the teams were similarly matched based on the data we had generated earlier. Hence, classifying algorithms had 3 classes to predict for the group stages, whether the match was won, lost or drawn. This was a different scenario in the knockout stages which followed a 1v1 knockout format and the outcomes of these were only winning or losing. The classification models were trained separately for the group stages and knockout stages based on the number of classes they had to separate the incoming data into.

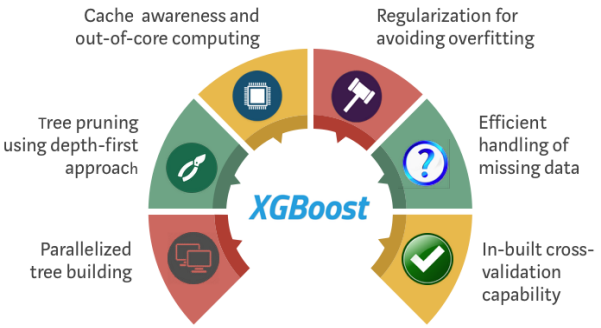
The first model we aimed to classify the group stages was the Random Forest Classifier. A random forest is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting. The sub-sample size is controlled with the max_samples parameter if bootstrap=True (default), otherwise the whole dataset is used to build each tree [7]. The random forest provided a very good accuracy of our predictions in the knockout stages but failed to provide a high accuracy for the group stages. This was assumed to be caused by the higher number of classes we required to classify the given data.

We resorted to a boosting algorithm after the Random Forest algorithm. The first model used was the AdaBoost algorithm to try and predict the outcome of the group stages. AdaBoost is frequently used to combine weak base learners (like decision stumps), but it has been demonstrated that it can also combine strong base learners (like deep decision trees) well, leading to an even more precise model [8]. The output produced was a good accuracy, but we assumed a better accuracy could have been obtained by training and testing using an XGBoost algorithm that could have produced a better accuracy.



The best accuracy for the group stages and knockout stages was produced by the XGBoost algorithm implemented. A gradient boosted decision tree implementation is called XGBoost. Instead of fitting the current hypothesis $h_m(x)$ on the residuals, XGBoost fits it on the gradient of the loss function, where m specifies the iteration number. This method minimises the loss function that is predicted using the Taylor expansion. The greedy method is used to build a tree instead of exploring every conceivable tree, and regularisation is used to penalise too-deep trees in order to prevent overfitting. Every $h_m(x)$ is multiplied by, which explains the variation in the effects of the split's many branches.





C. Predicting and Simulating the World Cup:

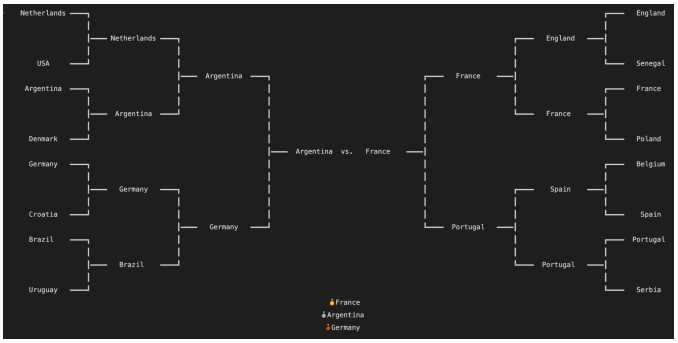
In a world-cup there are 8 groups and each group has 4 teams. Each team in a group plays each other once and the top two teams of the group get qualified for the round of 16, followed by quarter-finals , followed by semi-finals,followed by the final. First we predict the outcomes in the group stage of the world cup using an XGBoost model with optimal hyperparameters to predict the outcomes of the group stage games. A probability score is calculated which gives us the winning, losing, drawing probability of each game between two given teams .The knockout stages are predicted using a random forest classifier model (as it gave the best score). The target variable here takes only two values because of the knockout stages. Both these models are evaluated using a confusion matrix, XGBoost giving the best score for group stages and Random Forest giving the best score for knockout stages.

D. Predicting player rating and Grouping Similar Players

Here we collected, cleaned and preprocessed a dataset containing name,club,nationality,net-worth,football skills and attributes of many famous players .We created new attributes by combining various other attributes of players thereby reducing the number of dimensions in our dataset . Using a random forest and an XGboost model we could predict the ratings of players with a very high accuracy .We also grouped similar players to each other using their traits and ratings .Here we first converted our data into a vector space using the tf-idf vectorizer model of the "sckit-learn" library , and then we found the cosine similarity of the player traits .In conclusion,the players having the highest cosine similarity are very similar to each other.

IV. Results and Conclusion

After training and testing the data successfully, and making our predictions, we created a bracket style format for all teams with an output for results for both group stages as well as the knockout rounds. Each game is analyzed, and tested through the model taking the group data from the simple dataset created of the teams sorted by their respective world cup groups. The group stage XGBoost model predicts the outcomes of each group and which teams are likely to qualify. Thisdata is then stored and passed to our knockout stage model with the random forest classification. The random forest model works on the ladder based bracket and classifies the winner for each round of qualification i.e. the round of 16, quarter-finals, semi-finals and then finally the world cup final. At the end of our predictions, it is judges that the winner of the tournament would be France, who were favourites from the squad strength analysis we had conducted before.



This research paper can be concluded with a predicted winner for this year's world cup. This long-drawn tournament went through an exhaustive process of data analysis and model tuning which hence, declared our winner just going by the data at possession. The only thing left now is to see who actually ends up winning once the tournament gets rolling and checking if the predictions we have made are in fact correct.

Acknowledgment

We would like to thank the institution and management for giving us an opportunity to learn and gain knowledge on various topics with respect to the course(data analytics) . We would like to thank our professor for giving us various insights and inputs with respect to many details of this project and this course. We would like to thank our friends for helping us in many technical and non technical aspects in our project , mainly the frontend . We would like to thank all our well-wishers for giving us the confidence to go about this project

References

[1] https://publications.fifa.com/en/sustainability-report/sustainability-at-the-fifa-world-cup/profile-of-the-fifa-world-cup-qatar-2022/
[2] https://dl.acm.org/doi/10.1145/2882903.2912574
[3] http://positifreview.com/gallery/38-june2022.pdf
[4] https://ieeexplore.ieee.org/docum
[5] https://www.sciencedirect.com/science/article/abs/pii/S0169207018300116
[6] https://www.semanticscholar.org/paper/Effects-of-expertise-on-football-betting-Khazaal-Chatton/fd130aa25dcc3d2fe4771ad9d36ba7f1c612a4d0
[7] https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[8] https://www.kaggle.com/datasets/martj42/international-football-results-from-1872-to-2017
[9] https://www.kaggle.com/datasets/stefanoleone992/fifa-22-complete-player-dataset?select=players_22.csv