Statement of Work - Project-Modelling

Title**:FIFA World Cup Prediction**

**Rationale Statement:**

Sport prediction is one of the growing areas that requires high predictive accuracy because of the enormous monetary stakes involved in betting. Furthermore, club managers and owners also look for classification models to help them understand and design winning strategies. These models are based on a variety of game-related factors, including past match results, player performance indicators, and opponent information.

Football is one of the famous sports and making a prediction analyser for the game is an interesting topic. As football is a famous sport worldwide and the FIFA world cup is also approaching, everyone is curious about the winner of the world cup .The goal of this project is to predict who is going to win the next FIFA World Cup 2023 using the past datasets.

**Objective:**

1. To create a machine learning model capable of predicting the outcomes of football games in the 2023 FIFA World Cup.
2. Predict which team is going to win the finals.

**Data Sources:**

In this project, we are using multiple datasets which are interlinked to each other, two of them having more than 40 thousand records.This dataset contains all available FIFA men's international football rankings from August 1993 to April 2018 and international football results. It is a huge collection of data on matches spanning more than a century of international football games. Currently, the following datasets are being used from Kaggle. Additionally, more data can be collected from official football association websites.

1. FIFA football Rankings from 1993 to 2018

2. International football results from 1870 to 2021

3. World cup 2018

**Software, packages and tools used :**

1) We would be executing our code using the Jupyter Notebook or Google collab, as they provide the best performance for our code executions.

2) We would be using the following packages and libraries :

* pandas
* numpy
* seaborn
* pyplot and ticker from matplotlib

**Data Assumption:**

We are making the winning team column in categorical data by changing column values to “2” if the home team won, “0” if the away team won, “1” if it’s a tie match.

**Data Limitation:**

FIFA ranking was created in the 90’s, thus a huge portion of the dataset is lacking. So,we will use the results of historical matches since the beginning of the championship for all participating teams.

A few record values are not available and null, which will further be cleaned.

**Data Constraint:**

1) The name of the country at the time of the match is used for country names ,even if the names of the home team and the country don't match. For example, Ghana was playing at home in Accra, Gold Coast in the 1950s. The neutral column for those matches reads FALSE, indicating that they did not take place in a neutral site.

2) The current name of the team has been used for both home and away teams. For example, when a team known as Ireland played versus England in 1882, it is referred to as Northern Ireland in this dataset because the current Northern Ireland squad is the heir to the 1882 Ireland side. This is done to make it easier to keep track of a team's history and statistics

**Exploratory Data Analysis:**

Below are the screenshots of our EDA

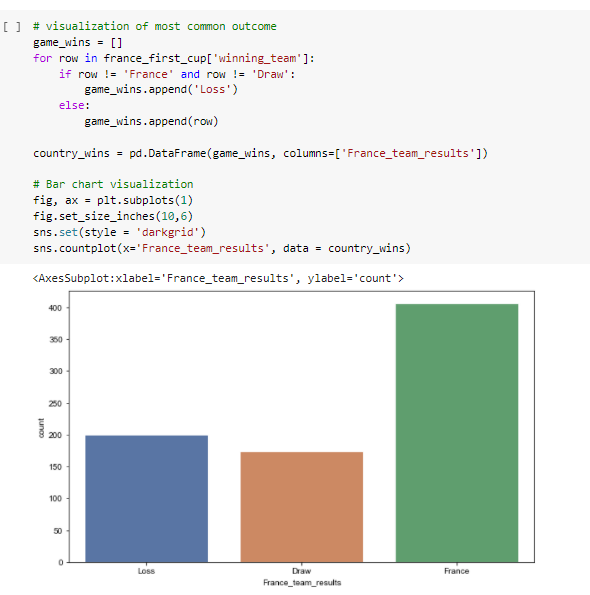
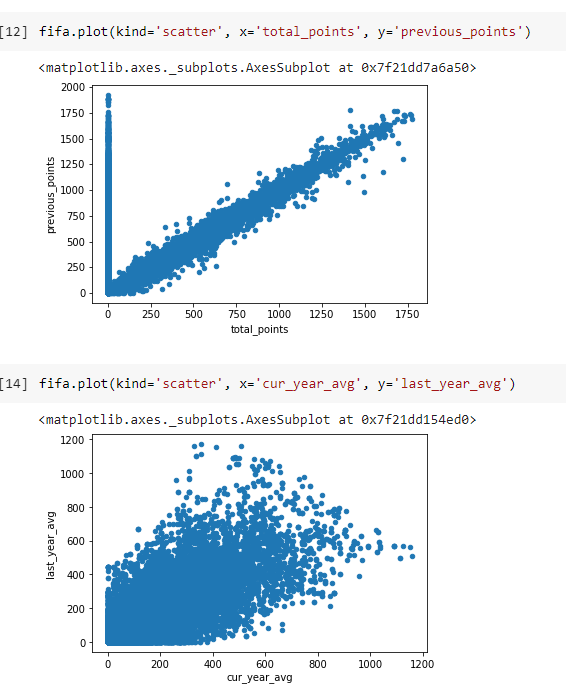
In our EDA, we have found a certain summary showing us the winners in each of the match tournaments. The output is as per the goal differences in the matches played.



Also, we have an EDA showing the output of the data by the countries who have just played by single countries.



Also, we have created some graphs like the scattered plot graphs and bar graph for the EDA.



**Modelling**

**Choice of models**

Since our data is mostly related to the world of sports, where the results can vary randomly, we can still think of some models that can help us to achieve the best of the output results. We will therefore test and compare the following classification models capable of generating probability predictions: –

* Logistic Regression
* Decision Tree
* Random Forest Classifier
* Convolutional Neural Network (CNN)

**Evaluation Criteria**

To evaluate a model performance, one would classify match results into home wins, away wins and draws and then look at the number of matches that the model has correctly identified, using a standard classification matrix. In this case, classification accuracy can be used for evaluation.

Models can be evaluated and then can be labeled "win", "lose" and "draw"

**Modeling Approaches:**

We are planning to go forward with the logistic regression modelling as it is the simplest model and can expect that that the model will give the best of the accuracy results.

Pros :

* Easier to implement, interpret, and very efficient to train.
* Fast at classifying unknown records
* Can easily extend to multiple classes(multinomial regression)
* Better accuracy than other classifying models

Cons :

* If the number of observations is lesser than the number of features, it may lead to overfitting.
* Logistic Regression requires average or no multicollinearity between independent variables.

**Model Architecture**

In the project, we are trying to predict the winner of the FIFA World Cup 2023. We are planning to use two different techniques for splitting the data into training dataset and testing dataset. Firstly, using the train test function of sklearn we will divide our train test split with a ratio of 70% for training and 30% testing. However, the disadvantage of this technique would be that our model doesn’t go through the entire dataset. Also, for better accuracy we would require a good amount of dataset for testing too. In machine learning we always would have a bias-variance trade off. To overcome this problem, we can plan to use K-fold cross validation. Due to this, our model will go through the entire dataset and we can get better results.

2) Therefore, we can use cross-validation to train our model multiple times using different training sets, and generate performance metrics on the test set for which we then take the average for all cross-validation training runs.

3) We use the best model, i.e, logistic Regression algorithm, to do prediction on N-fold cross validation(N folds for training, X fold for testing).

4) Models would then be evaluated on following criteria like Precision, recall , F1, which can be carried out for each label as "win", "lose" and "draw".

**References:**

<https://www.sciencedirect.com/science/article/pii/S2210832717301485>

<https://content.iospress.com/articles/journal-of-sports-analytics/jsa200463>

<https://www.imperial.ac.uk/media/imperial-college/faculty-of-engineering/computing/public/1718-ug-projects/Corentin-Herbinet-Using-Machine-Learning-techniques-to-predict-the-outcome-of-profressional-football-matches.pdf>

**Team Members :**

Smriti Raina – 100830255

Devangraj Vaidya – 100780615

Prasann Patel – 100833325

Samarth Patel – 100827936