Statement of Work - v2

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

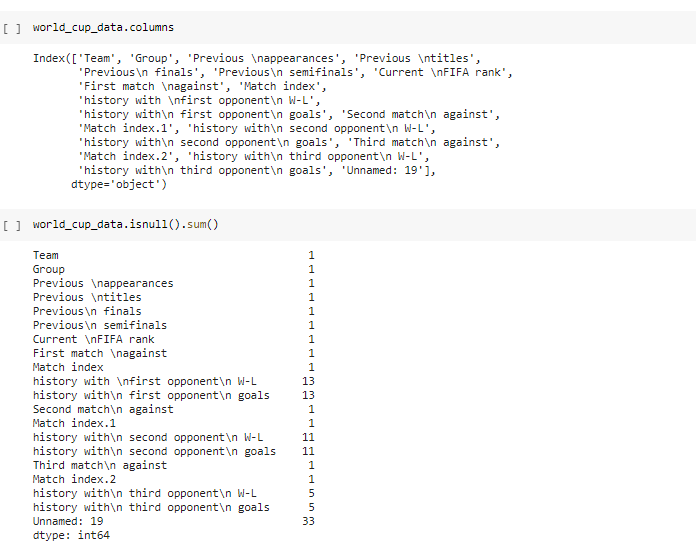
3. World cup 2018

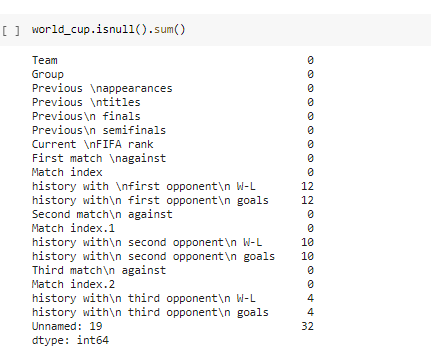
**Data Assumption:**

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 Manipulation/Cleaning/Preparation**

Initially, we are checking the list of columns from our dataset. After discovering that there are null values in our data, we identified them and then planned to reduce/remove the null values from the dataset, so that the data is clean, and has minimal chances of error.





**Exploratory Data Analysis:**

Below are the screenshots of our EDA

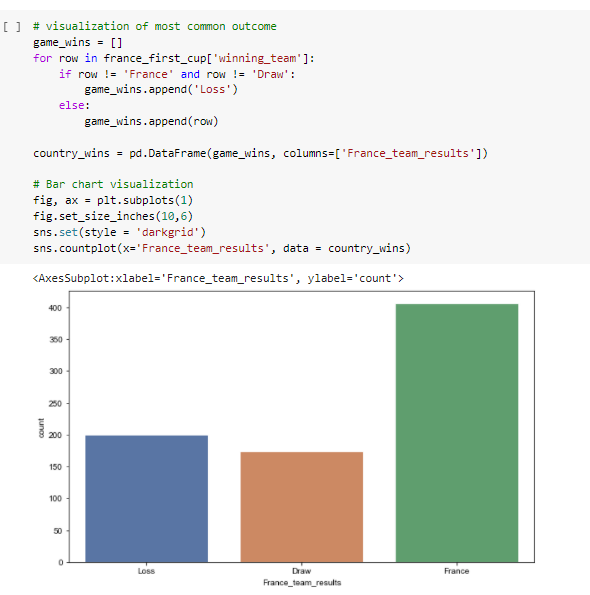
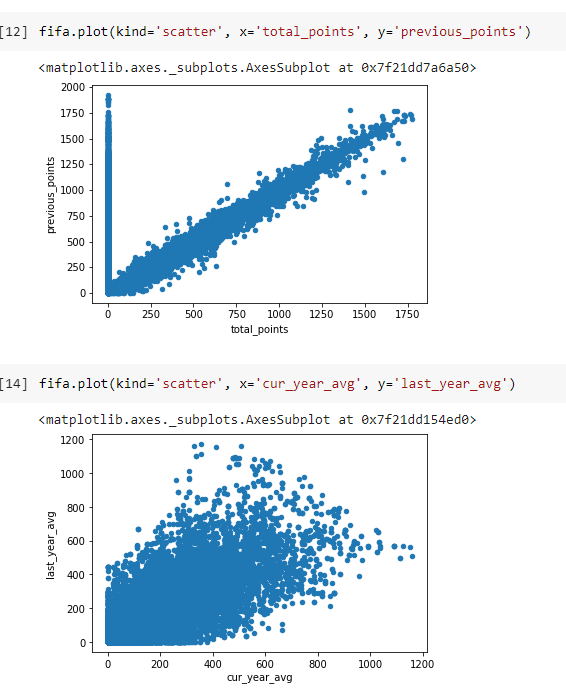
In our EDA, we have found a certain summary showing us the winners in each of the match tournaments. The output is as per the goal differences in the matches played.



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.



**Modeling Approaches:**

We are planning to go forward with the logistic regression modelling.

Pros :

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

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.

**References:**

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

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

**Team Members :**

Smriti Raina – 100830255

Devangraj Vaidya – 100780615

Prasann Patel – 100833325

Samarth Patel – 100827936