Mchezo Pesa

Introduction

Today we are going to analyze the dataset for Mchezo pesa and see if we can make predictions for them. We are provided with two datasets, one containing the rankings of the countries and another containing the results.

Business Understanding

Since Mchezo Pesa is a betting domain that charges 100 for every bet / prediction placed on the results. It is our duty as data scientists to use these machine learning models to come up with predictions to enable them to see how much profit they can make off these bets, and the probability of a person winning.

Problem Statement

Can we predict the FIFA results using our model ? How accurate is the rank in predicting the results.

Data Description

The datasets had the following columns :

- Rank
- Country Abbreviation
- Total Points
- Previous Points
- Rank Change
- Average Previous Years Points
- Average Previous Years Points Weighted (50%)
- Average 2 Years Ago Points
- Average 2 Years Ago Points Weighted (30%)
- Average 3 Years Ago Points
- Average 3 Years Ago Points Weighted (20%)
- Confederation
- Date - date of the match
- Home_team - the name of the home team
- Away_team - the name of the away team
- Home_score - full-time home team score including extra time, not including penalty-shootouts

- Away_score - full-time away team score including extra time, not including penalty-shootouts
- Tournament - the name of the tournament
- City - the name of the city/town/administrative unit where the match was played
- Country - the name of the country where the match was played
- Neutral - TRUE/FALSE column indicating whether the match was played at a neutral venue

Modelling procedure

Since we are looking at the FIFA World Cup predictions we made a sample with only the fifa records. We then added the rank column from the first dataframe into the sample.

We first created a new column that had win,lose and draw.We then converted that column into numerical where -1 was a loss , 0 was a draw and 1 was a win. We then trained the difference, rank and scores column. I then performed Logistic Regression on this model.

We then had an additional polynomial regression model where we tried to use the rank column to predict the scores.The model did not perform that well.

Modelling results

Our model was perfect. It did not contain any wrong predictions, there was no need of hyperparameter tuning since it did the correct thing. We had an RMSE value of 0.

However the polynomial regression did not predict that well.We can see this from the scatter plot.

Summary and conclusions

We can easily predict whether the game will be won provided we have the previous scores of the games. However the rank and previous scores have the highest weightings.