Moringa Data Science

Independent Project

Core Week 10: Supervised Learning - Regression Mini Project

Prediction Results for MchezoPesa Ltd

Data Report

Table of Contents

Item	Description	URL
Google Colab	-	Link

Introduction

The beginnings of contemporary football can be tracked back to 19th century Britain. Prior to this, folk football was played with numerous regulations but it was at the public schools that a standardized form of the game began to develop.

You have been recruited as a football analyst in a company - Mchezopesa Ltd and tasked to accomplish the task below.

A prediction result of a game between team 1 and team 2, based on who's home and who's away, and on whether or not the game is friendly (include rank in your training).

3. Defining the Question

How to predict the outcome of a soccer match based on the ranking of the home and away teams and the type of tournament

4.Metrics for Success

- ➔ Optimal and reliable prediction Model:
  - ◆ About 80% Accuracy score and above
  - ◆ With the least RMSE score achievable

5. The Context

The FIFA Men's World Ranking ranks teams of FIFA's member nations on the basis of their game results from full international matches, awarding points to teams depending on the outcome. Brazil is presently ranked highest, followed by eight other countries who have held the top spot since its introduction in December 1992. In response to criticism, the ranking system has been revised multiple times, most recently transitioning to the Elo rating system found in chess in August 2018.

A more detailed explanation and history of the rankings is available here: [Link]

7. Experimental design taken

- ➔ Perform your EDA
- ➔ Perform any necessary feature engineering
- ➔ Check of multicollinearity
- ➔ Building a model
  - ◆ Approach 1: Polynomial regression model
    - Model 1: Predict how many goals the home team scores
    - Model 2: Predict how many goals the away team scores
  - ◆ Approach 2: Logistic regression model
    - Figure out from the home team's perspective if the game is a Win, Lose or Draw (W, L, D)
    -
- ➔ Cross-validate the model
- ➔ Compute RMSE
- ➔ Create residual plots for the model
- ➔ Assess Heteroscedasticity using Bartlett's test
- ➔ Challenge the solution.
- ➔ Create a dashboard that communicates the findings.

8. Appropriateness of the available Data

- ➔ Two datasets are available:
- ➔ Ranking dataset: contains the team ranks from 1993 to 2018
- ➔ Results dataset: contains matches and the team scores since 1892 to 2019

- ➔ **Resources**
- ➔ Datasets: Link
- ➔ Tools used
  - ◆ Google Colab
  - ◆ Github
  - ◆ Moringa school canvas access for data and problem statement access

- ➔ **Assumptions**
  - ◆ Data provided is sufficient to fulfil the research objectives

Constraints

- Data is from varied times.

Cost

Analysis time.

Data Understanding

Data Structure and Exploration

The available data appears to be appropriate for answering the given question. The results.csv file contains information on the outcomes of past soccer matches, including the rankings of the home and away teams, while the fifa_ranking.csv file contains the rankings of various soccer teams. This information should be sufficient for building a regression model to predict the outcomes of soccer matches.

Data Preparation

We loaded the necessary libraries required to carry out our analysis then uploaded our data.

After checking the variables in each dataset we noticed that the maximum scores were very high although factual. We chose to remove scores greater than 5 as they were not realistic.

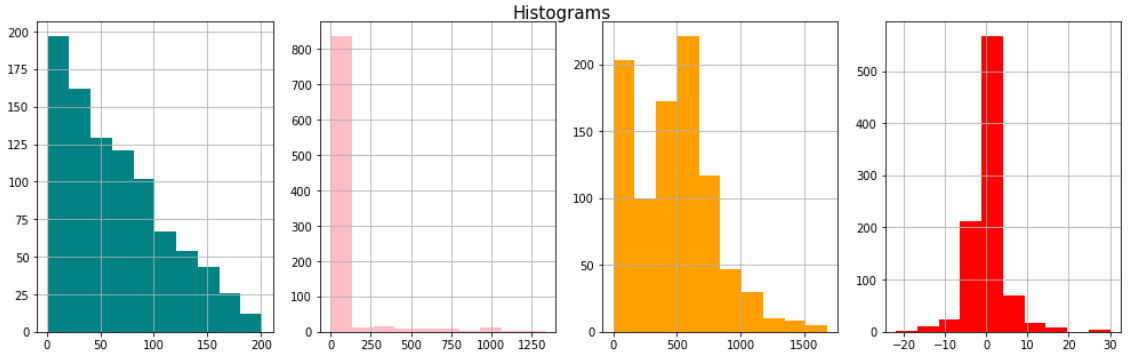
- Univariate Analysis

This was done in different ways.

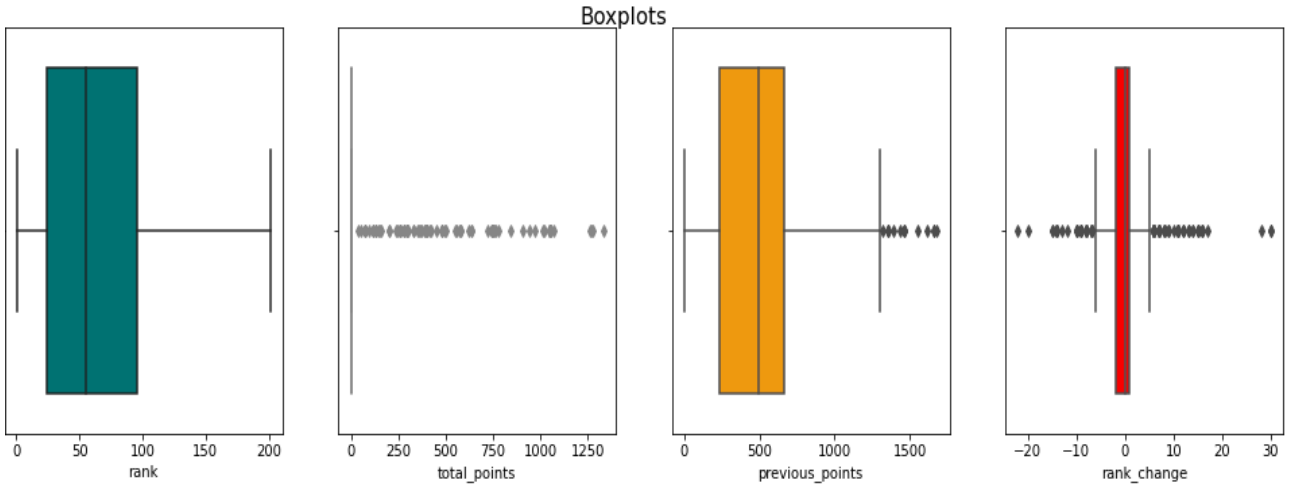
We started by looking at the descriptive statistics: count, mean, standard deviation, minimum and maximum values and the quantiles, range, interquartile ranges, skewness and kurtosis.

We also checked distributions of individual variables using histograms

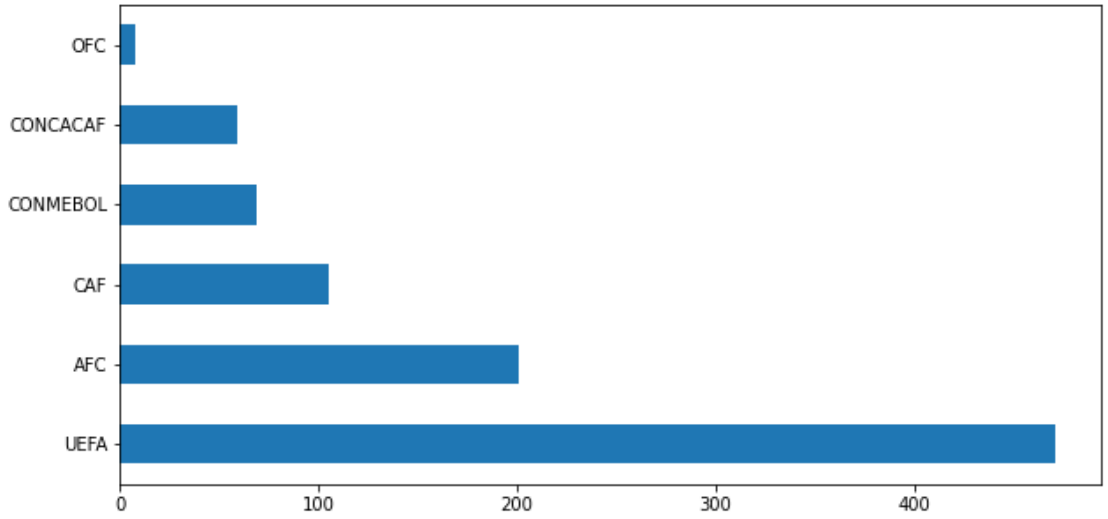
histogram showing numeric variable distribution



Box plots used to check the presence of outliers. We chose to keep these outliers as they represented the true state of things



- Horizontal bar charts showing the distribution of specific variables such as the tournaments, neutrality of grounds where games were played and the confederations involved in planning the games.

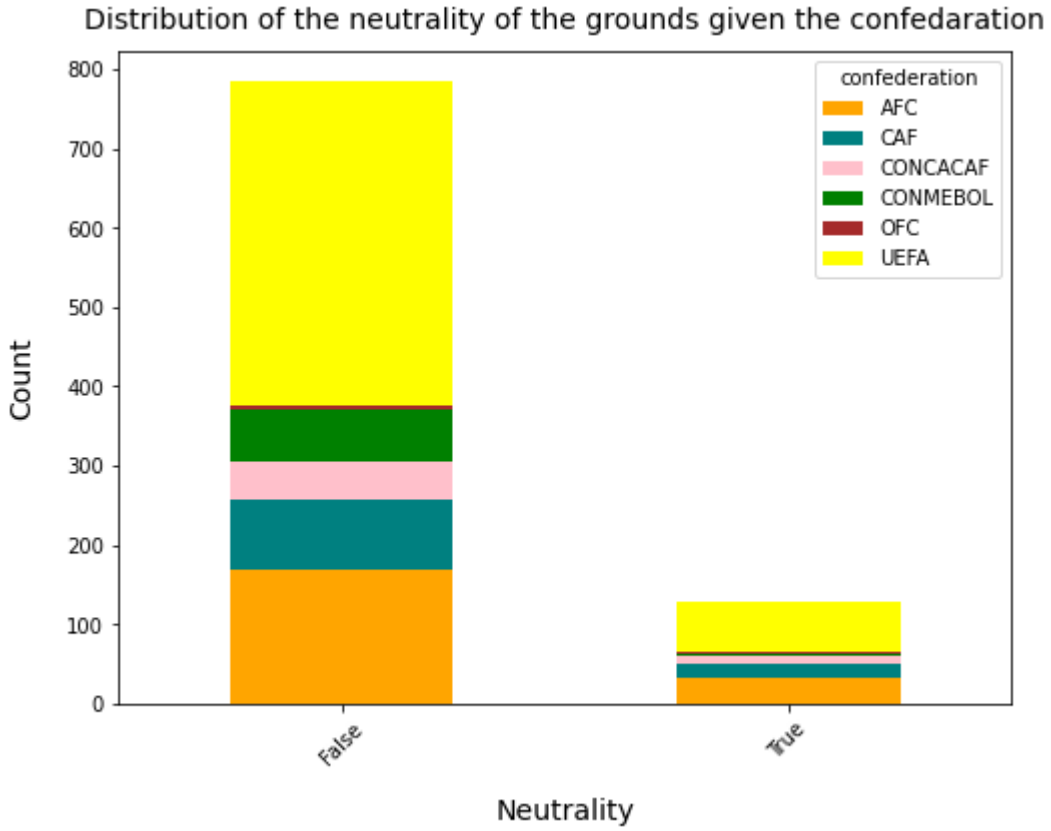


.

- Bivariate Analysis

In this section we explored the relationships between our variables in pairs.

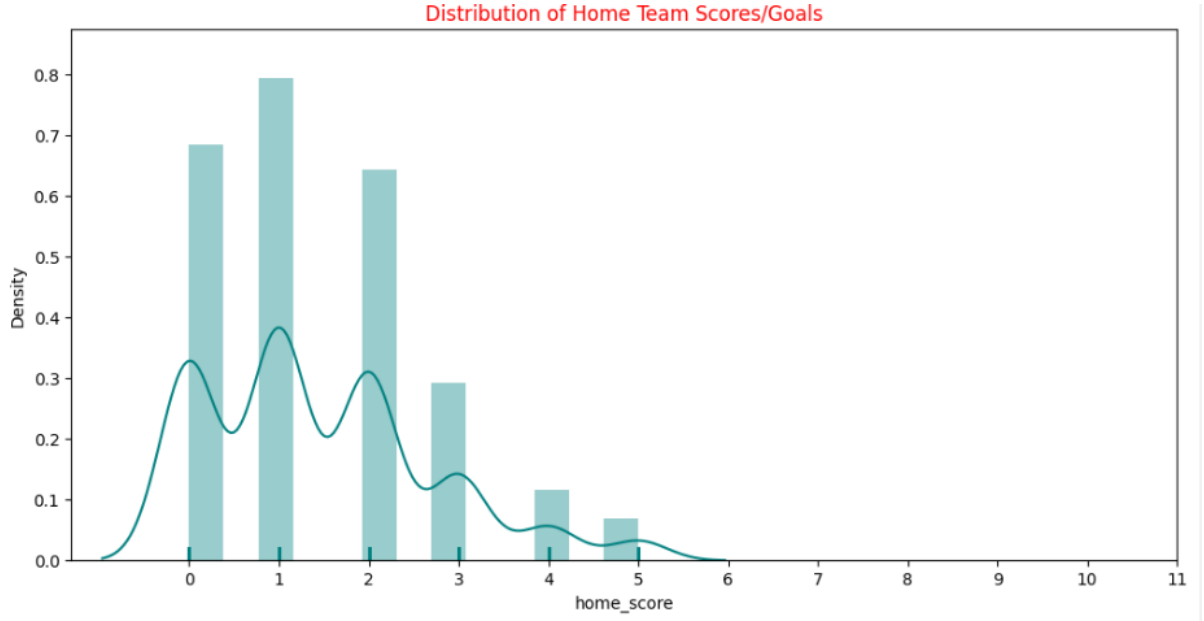
This included running correlation tests which we visualised using heat maps for faster understanding, generating visualisations such as stacked column charts and then encoding our categorical variables using LabelEncoder and running the same tests.

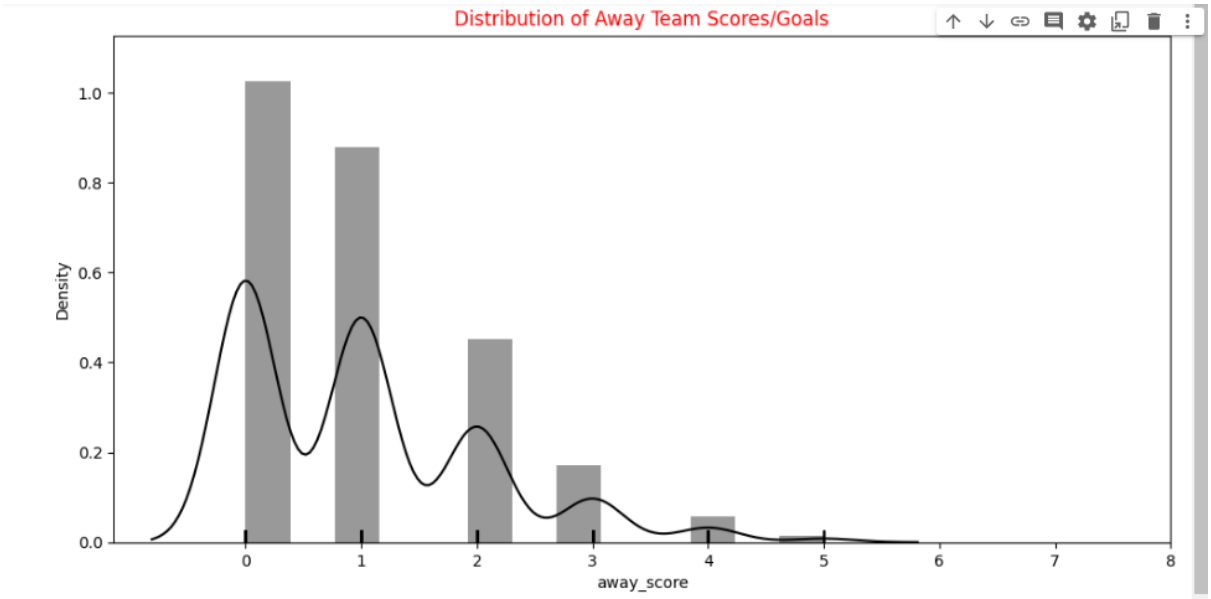


Data Preparation

a) Feature Engineering

Here we create new features that may be useful for the model, and the use of regression techniques to build and evaluate the model.





The histograms show that the home teams are highly likely to score at most one goal compared to the away team.

The charts also show that the goals scored in both home and away team is skewed to the right.

This means that there is a high probability of scoring more goals.

i.e. the outliers lie on the positive side.

Building a Model

We built the model using appropriate regression techniques, and cross-validated the model to ensure its accuracy. We then computed the RMSE to measure the model's performance

Conclusion

. 1.Predicting whether home team wins, loses or draws in a match

The best accuracy score using two hyperparameters is 57% Using XGboost algorithm the accuracy score was 75% Therefore, another algorithm can be considered to improve the accuracy score or more hyperparameters should be tuned

---

2. Predicting the rank of a team

None of the regularised regression models is a good fit to predict the rank; they all have RMSE scores greater than the mean of the actual. Since this is also more of a classification problem, using regressions is not the best approach though this project was restrictive.