SEGMENTING AND CLASSIFYING THE BEST STRIKERS REPORT

My Data Analytics Project

SEPTEMBER 22, 2025

OKE TOLULOPE

Contents

Introduction

In the world of football, strikers play a pivotal role in deciding the fate of matches and championships. Identifying the best strikers among a pool of talent involves a comprehensive analysis of various factors ranging from performance metrics to personal attributes. In this project, titled "Segmenting and Classifying the Best Strikers," I delve into a dataset containing information on 500 strikers, aiming to uncover patterns, insights, and classifications that distinguish top-performing strikers from the rest.

Project Description

The project involves utilizing data analytics techniques to explore and understand the characteristics and performance metrics of strikers. By employing descriptive statistics, data visualization, feature engineering, and machine learning algorithms, I aim to identify the key attributes that contribute to a striker's success on the field and classify them into different categories based on their performance.

Purpose

The primary purpose of this project is to provide a systematic framework for analyzing and categorizing strikers based on their performance metrics and personal attributes. By doing so, coaches, scouts, and football analysts can gain valuable insights into the characteristics of top-performing strikers and make informed decisions in team selection, recruitment, and strategic planning.

Dataset Description

The dataset comprises various variables related to 500 strikers, encompassing both demographic information and performance metrics. Key variables include nationality, footedness, marital status, goals scored, assists, shot accuracy, dribbling success, and many more, providing a comprehensive overview of each striker's profile and on-field performance.

- **Striker ID:** Unique identifiers assigned to each striker.
- **Nationality:** The country of origin for each striker.
- **Footedness:** Indicates whether the striker is right or left-footed.
- **Marital Status:** Indicates whether the striker is married (yes) or unmarried (no).
- **Goals Scored:** The total number of goals scored by the striker, a fundamental performance metric.
- **Assists:** The number of assists provided by the striker, indicating their ability to create goal-scoring opportunities for teammates.
- **Shots on Target:** The number of shots taken by the striker that hit the target, reflecting their ability to create scoring opportunities and test the goalkeeper.
- **Shot Accuracy:** The percentage of shots on target out of total shots taken, showing the striker's precision and effectiveness.
- **Conversion Rate:** The percentage of shots that result in goals, revealing the striker's efficiency in front of goal.

- **Dribbling Success:** A metric indicating the striker's ability to bypass defenders and create goal-scoring opportunities through individual skill.
- **Movement off the Ball:** Reflects how actively the striker moves to find space and create opportunities for themselves and teammates.
- **Hold-up Play:** Measures the striker's ability to retain possession and bring teammates into play with passes or layoffs.
- **Aerial Duels Won:** The number of aerial duels won by the striker, important for strikers strong in the air as it can create scoring chances.
- **Defensive Contribution:** Reflects the striker's defensive efforts such as tracking back, pressing opponents, and making interceptions.
- **Big Game Performance:** Indicates the striker's performance in important matches, which can elevate their reputation.
- **Consistency:** Reflects how regularly the striker performs at a high level over the course of a season or multiple seasons.
- **Versatility:** Measures the striker's ability to adapt to different tactical systems and roles within the team.
- **Penalty Success Rate:** The efficiency of the striker from the penalty spot, crucial in tight matches.
- **Impact on Team Performance:** Reflects how the team's results and overall attacking play are influenced by the striker's presence.
- **Off-field Conduct:** Measures the striker's professionalism, leadership, and behavior, which can impact their overall performance and value to the team.

Required Tools

- Python programming language
- Jupyter Notebook

Your Job - Questions to solve

1. Data Cleaning**:**

- Load the dataset it into Jupyter notebook.
- Load all the relevant and necessary packages for the required tasks.
- Check for missing values within any column and use `SimpleImputer` to impute the missing values. Use strategy 'median' for numeric and 'most frequent' for nominal columns.
- Check for the correct data types and assign integer data types for specific variables: `'Goals Scored', 'Assists', 'Shots on Target', 'Movement off the Ball', 'Hold-up Play', 'Aerial Duels Won', 'Defensive Contribution', 'Big Game Performance', 'Impact on Team Performance', 'Off-field Conduct'.`

2. Descriptive Analysis**:**

- Perform descriptive analysis on the dataset. Round the output values by 2 decimal points.

3. Data Visualization**:**

- Perform percentage analysis on the variable Footedness and create a pie chart on the output using **`matplotlib`**.
- Visualize the distribution of players' footedness across different nationalities in a countplot of **`seaborn`**.

4. Statistical Analysis**:**

- Determine which nationality strikers have the highest average number of goals scored.
- Calculate the average conversion rate for players based on their footedness.
- Find whether there is any significant difference in consistency rates among strikers from various nationalities. Before doing the appropriate test, must check for the assumptions.
- Check if there is any significant correlation between strikers' Hold-up play and consistency rate. Must check for the assumptions.
- Check if strikers' hold-up play significantly influences their consistency rate.

5. Feature Engineering**:**

- Create a new feature - Total contribution score by summing up specific columns: **`'Goals Scored', 'Assists', 'Shots on Target', 'Dribbling Success', 'Aerial Duels Won', 'Defensive Contribution', 'Big Game Performance', 'Consistency'.`**
- Encode the Footedness and marital status by **`LabelEncoder`**.
- Create dummy variables for Nationality and add them to the data.

6. Clustering Analysis**:**

- Perform **`KMeans`** clustering:
- Select features by dropping the **`Striker_ID`** from the updated data.
- Calculate the Within-Cluster-Sum-of-Squares (WCSS).
- Visualize the elbow chart to select the optimal number of clusters (The breakpoint of elbow chart must show 2).
- Build the **`KMeans`** cluster with the optimal number of clusters and add the labels into the data.
- Calculate the average total contribution score by the value of clusters.
- Assign the tag **`'Best strikers'`** for **`0`** and **`'Regular strikers'`** for **`1`** and add a new column **`'Strikers types'`** into the data. Drop the Clusters variable.
- Use feature mapping to map the new feature Strikers types: **`'Best strikers'`** for **`1`** and **`'Regular strikers'`** for **`0`**.

7. Machine Learning Model**:**

- Select the features into **`x`** and the target column Strikers types into **`y`**. Must delete unnecessary columns (i.e., **`'Strikers_ID'`**) while selecting the features.
- Perform feature scaling with **`StandardScaler`** and split the data into train and test sets where the test data size will be 20%.
- Build a logistic regression machine learning model to predict strikers type.
- Make predictions and evaluate by calculating the accuracy percentage.
- Create the confusion matrix and visualize it.
  Finally, answer the question asked in this assignment and you are done!

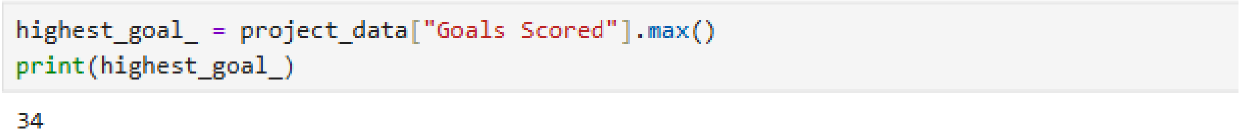
Conclusion

Through a comprehensive analysis of the dataset, we've gained valuable insights into the characteristics and performance metrics of strikers. By segmenting and classifying the strikers based on their attributes and performance, we've provided a framework for identifying top-performing strikers and predicting their performance type. This project serves as a valuable resource for football professionals and enthusiasts alike, aiding in talent identification, team selection, and strategic planning.

Questions for this project.

1. What is the maximum goal scored by an individual striker?

The max goal scored is 34.

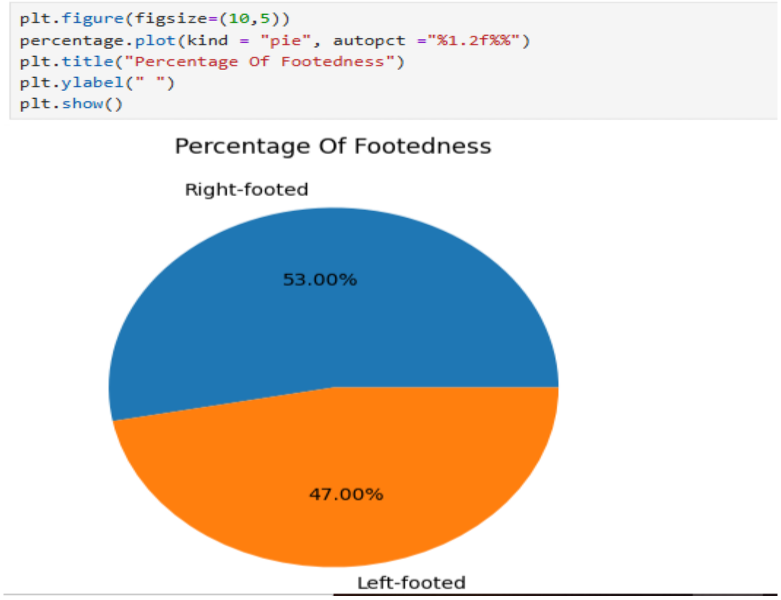


```
highest_goal_ = project_data["Goals Scored"].max()
print(highest_goal_)
```

```
34
```

2. What is the portion of Right-footed strikers within the dataset?

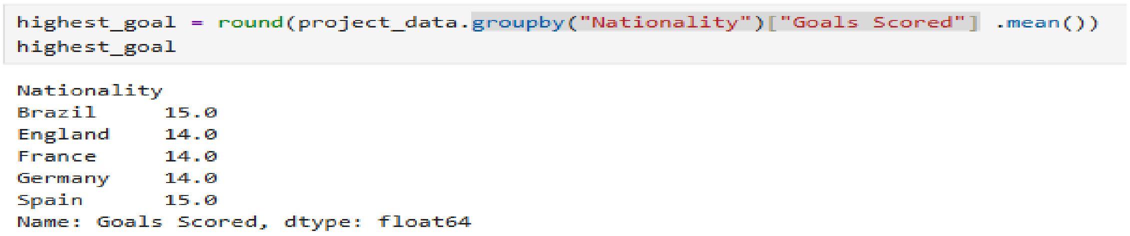
The portion of right-footed strikers is 53.00%.



```
plt.figure(figsize=(10,5))
percentage.plot(kind = "pie", autopct ="%1.2f%%")
plt.title("Percentage Of Footedness")
plt.ylabel(" ")
plt.show()
```

3. Which nationality strikers have the highest average number of goals scored?

Strikers from Spain and Brazil have the same and highest number of average goal scored.

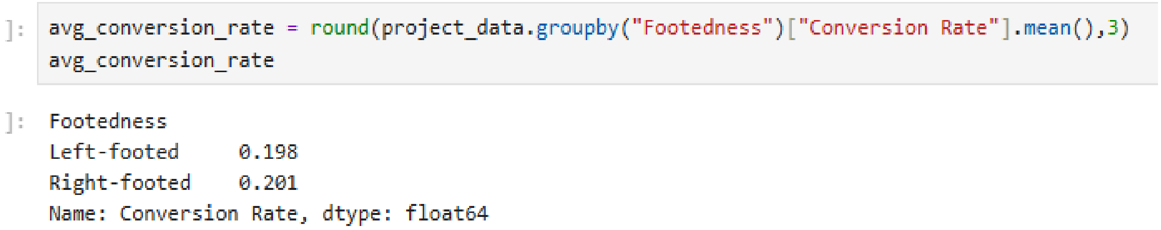


```
highest_goal = round(project_data.groupby("Nationality")["Goals Scored"] .mean())
highest_goal
```

```
Nationality
Brazil      15.0
England     14.0
France      14.0
Germany     14.0
Spain       15.0
Name: Goals Scored, dtype: float64
```

4. What is the average conversion rate for left-footed player?

The conversion rate for left-footed players is 0.198.

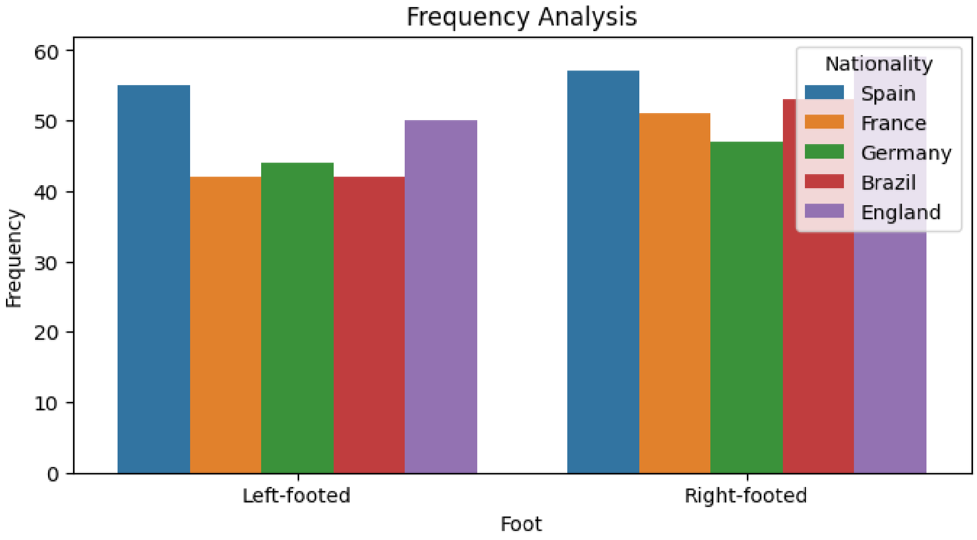


```
]:  avg_conversion_rate = round(project_data.groupby("Footedness")["Conversion Rate"].mean(),3)
    avg_conversion_rate
```

```
]:  Footedness
    Left-footed     0.198
    Right-footed    0.201
    Name: Conversion Rate, dtype: float64
```

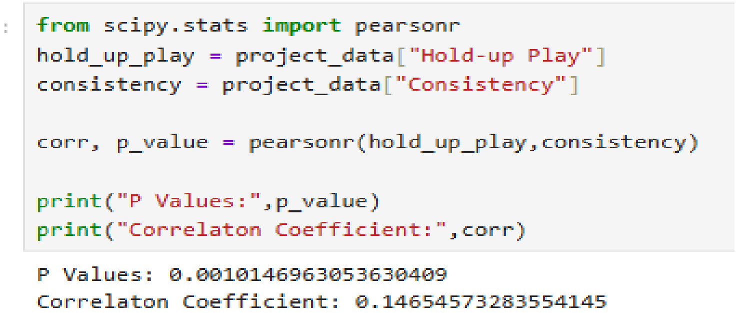
5. How many left footed players are from France?

There are 45 left-footed players in France.



6. What is the correlation co-efficient between hold up play and consistency score?

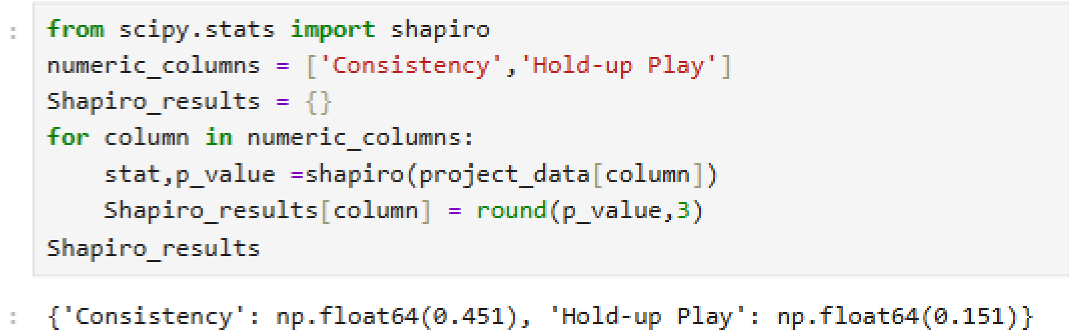
The correlation coefficient between hold up play and consistency score is 0.1465.



```
:   from scipy.stats import pearsonr
    hold_up_play = project_data["Hold-up Play"]
    consistency = project_data["Consistency"]

    corr, p_value = pearsonr(hold_up_play,consistency)

    print("P Values:",p_value)
    print("Correlaton Coefficient:",corr)

    P Values: 0.0010146963053630409
    Correlaton Coefficient: 0.14654573283554145
```

7. What is the p-value for the shapiro wilk test of consistency score? Is it normally distributed?

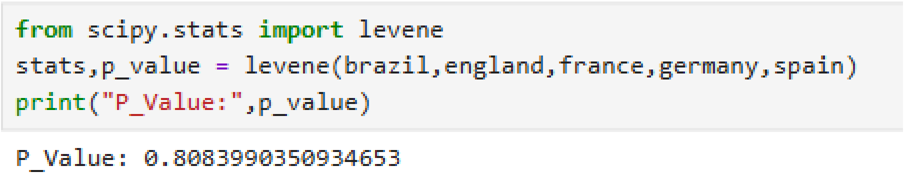
The p-value for the test is 0.451 which is higher than the significant value(0.05) & this indicates that the consistency rate is normally distributed.



```
:   from scipy.stats import shapiro
    numeric_columns = ['Consistency','Hold-up Play']
    Shapiro_results = {}
    for column in numeric_columns:
        stat,p_value =shapiro(project_data[column])
        Shapiro_results[column] = round(p_value,3)
    Shapiro_results

:   {'Consistency': np.float64(0.451), 'Hold-up Play': np.float64(0.151)}
```

8. What is the p-value for the levene's test of ANOVA analysis? Is the heteroscedasticity assumed?

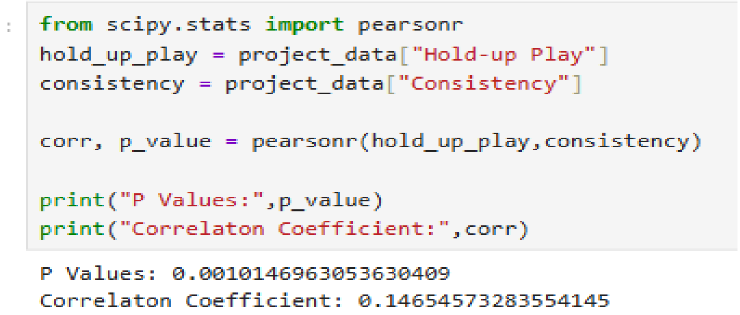
The p-value of the levene’s test is 0.808 and there is no statistical evidence of heteroscedasticity.



```
from scipy.stats import levene
stats,p_value = levene(brazil,england,france,germany,spain)
print("P_Value:",p_value)

P_Value: 0.8083990350934653
```

9. Is there any significant correlation between strikers' Hold-up play and consistency rate?

There is a weak correlation between the strikers’ hold up play and the consistency rate.

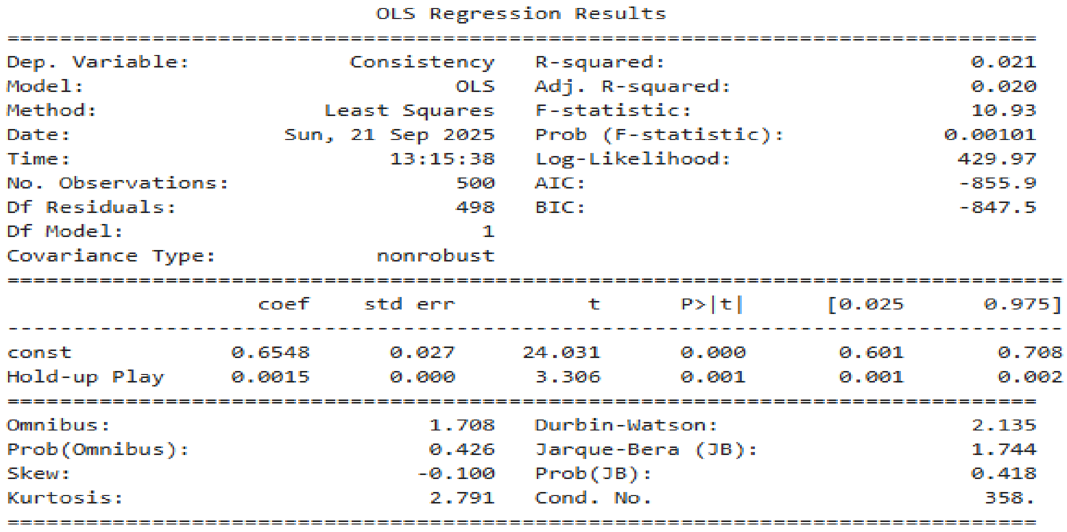


```
from scipy.stats import pearsonr
hold_up_play = project_data["Hold-up Play"]
consistency = project_data["Consistency"]

corr, p_value = pearsonr(hold_up_play,consistency)

print("P Values:",p_value)
print("Correlaton Coefficient:",corr)
```

```
P Values: 0.0010146963053630409
Correlaton Coefficient: 0.14654573283554145
```

10. Describe the beta value of Hold-up Play you have found in your regression analysis.

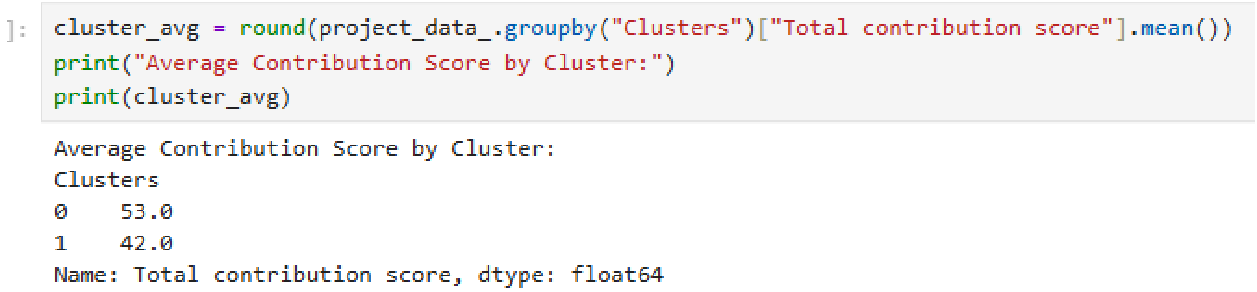
The beta value is 0.0015, which indicates that whenever the hold up play increases by 1 the consistency score increases by 0.0015. The x(hold up play) variable has a positive influence on the y(consistency score) variable.



```
                            OLS Regression Results
==============================================================================
Dep. Variable:            Consistency   R-squared:                       0.021
Model:                            OLS   Adj. R-squared:                  0.020
Method:                 Least Squares   F-statistic:                     10.93
Date:                Sun, 21 Sep 2025   Prob (F-statistic):            0.00101
Time:                        13:15:38   Log-Likelihood:                 429.97
No. Observations:                 500   AIC:                            -855.9
Df Residuals:                     498   BIC:                            -847.5
Df Model:                           1
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
const            0.6548      0.027     24.031      0.000       0.601       0.708
Hold-up Play     0.0015      0.000      3.306      0.001       0.001       0.002
================================================================================
Omnibus:                        1.708   Durbin-Watson:                   2.135
Prob(Omnibus):                  0.426   Jarque-Bera (JB):                1.744
Skew:                          -0.100   Prob(JB):                        0.418
Kurtosis:                       2.791   Cond. No.                         358.
==============================================================================
```

11. What is the average Total contribution score you get for the best strikers?

The average total contribution score for best strikers is 42.0.

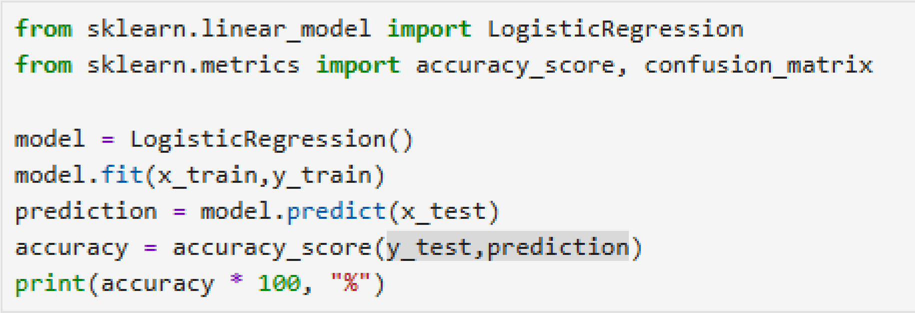


```
cluster_avg = round(project_data_.groupby("Clusters")["Total contribution score"].mean())
print("Average Contribution Score by Cluster:")
print(cluster_avg)
```

```
Average Contribution Score by Cluster:
Clusters
0    53.0
1    42.0
Name: Total contribution score, dtype: float64
```

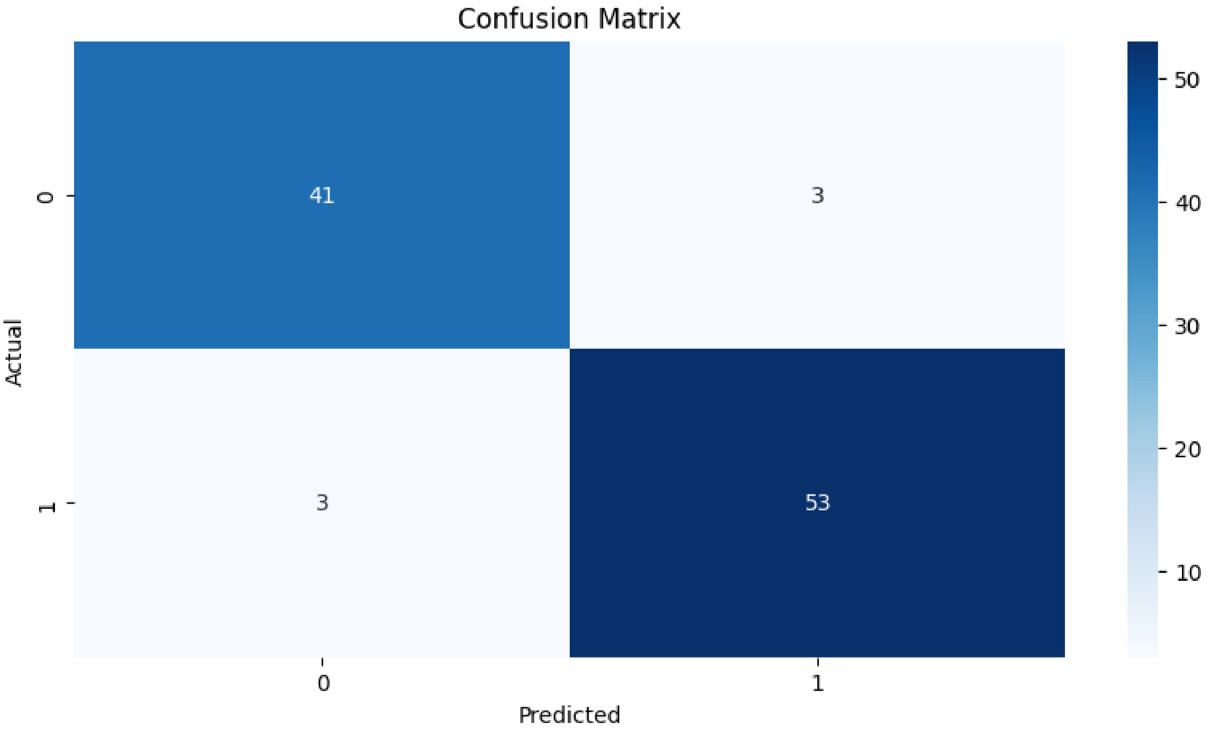
12. What is the accuracy score of your LGR model? How many regular strikers your model predicted correctly? How many best strikers your model predicted incorrectly?

The accuracy score of the Logistic Regression is 94%, the regular strikers correctly predicted are 41 and the best strikers incorrectly predicted are 3.



```python
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, confusion_matrix

model = LogisticRegression()
model.fit(x_train,y_train)
prediction = model.predict(x_test)
accuracy = accuracy_score(y_test,prediction)
print(accuracy * 100, "%")
```

```
94.0 %
```



Link to the Jupyter notebook:

https://drive.google.com/file/d/1BVUxn7InibJ2JwWpVtafxm_rHuJDsC8x/view?usp=drivesdk