Predicting International Soccer Matches

Introduction

From my earliest memories at the age of seven, soccer has always been more than just a game to me. It wasn't merely about supporting teams and celebrating goals; it was the intricate dance of strategies, the unpredictable ebb and flow of the match, and the sheer thrill of the unexpected. Over the years, while the spirit and energy of soccer remained unchanged, my perspective shifted from that of an impassioned spectator to an analytical enthusiast. I found myself intrigued, not just by the game unfolding before my eyes, but by the unseen patterns, the underlying statistics, and the factors that swayed match outcomes.

Soccer, in its global grandeur, thrives on unpredictability. However, there exist certain elements – like the undeniable impact of home advantage, the momentum shifts after own goals or penalties, or the effect an early or late goal can have. These factors influence the flow of a match. Traditional analytics, while invaluable, have largely been centered around player statistics and team head-to-head records. But I believe there's a deeper layer waiting to be unearthed – a layer that encompasses the psychological and strategic facets of the game.

Historically, the realm of soccer analytics was dominated by conventional statistics – player performance metrics, possession percentages, and simple win-draw-lose probabilities. However, with the advent of technology and sophisticated data collection mechanisms, the field has evolved dramatically. Modern soccer analytics now harnesses machine learning and advanced algorithms, diving deeper into intricate patterns and offering predictive insights that were once deemed impossible.

This project emerges from this very belief. While my love for soccer was born watching my favorite teams with friends, my analytical journey begins with the intent to explore and predict international soccer match outcomes. Through this endeavor, I wish to not only identify patterns or trends that might offer teams a competitive edge but also to enrich our collective understanding of the beautiful game. In the vibrant canvas of soccer, moments define matches. From the roaring advantage of playing on home turf to the adrenaline rush of a late goal, each nuance holds the potential to chart the course of a game. It is these nuances, these pivotal moments, that I aspire to decode, offering insights that transcend conventional statistics and delve into the very heart of soccer.

Dataset

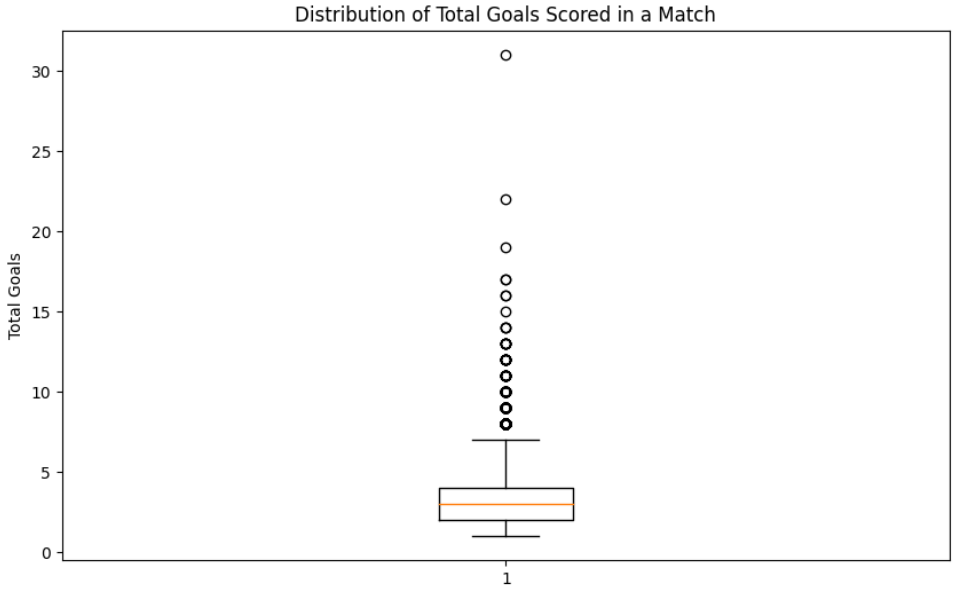
I stumbled upon an extensive dataset on Kaggle, consisting of international football matches from 1872 to 2023. A preliminary glance revealed several missing crucial variables, such as goalscoring minutes, and own goals and penalty goals, from 1872 to 1915, which could critically undermine the predictive model's efficiency. Consequently, I decided to concentrate on data from July 1916 to July 2023. After meticulous data cleaning and wrangling, this decision resulted in a refined dataset encompassing 13,200 matches.

Originally, this dataset cataloged a plethora of variables, ranging from team names and goals scored to venues and tournament types. However, my primary research question nudged me toward a more in-depth exploration. This journey into feature engineering enabled me to enrich the dataset with crucial variables. One such variable was 'home advantage', inferred from the venue details. Another was 'goal periods', which I created by segmenting the match duration into pivotal moments such as the early game, mid-game, and final moments. The dataset was further augmented with indicators for own goals and penalty goals, shedding light on match dynamics.

To offer some background on the data collection: The initial dataset was divided across three distinct CSV files. For this project's purposes, I found it imperative to merge two of these - **goalscorer.csv** and **results.csv**, as they collectively encompassed the information vital for the machine learning model. The process of merging those two dataset went smoothly as they were well-structured and contained several similar variables as well. In the preprocessing phase, I addressed not just the aforementioned early records but also several outliers. The outliers were games with more than eight goals in them. I did a detailed analysis with a box plot seeing that only a handful of games consist of eight goals or more. (See Figure 1)

The games with more than eight scored was filtered from the dataset, while categorical variables, like team names, underwent numerical encoding to ensure compatibility with machine learning models. Initial visualizations of the processed dataset unveiled intriguing patterns. Notably, the correlation between home advantage and match outcomes stood out, promising insightful revelations in the subsequent stages of the study. (See Figures 12 & 13)

Figure 1



Methodology

To predict international soccer match outcomes, I employed a systematic approach using various machine learning models. My methodology was underpinned by rigorous data preprocessing, where I addressed missing values, outliers, and irrelevant features. Particular emphasis was placed on feature engineering, which enabled me to derive new features such as "home advantage" and "goal-scoring periods" to better capture the nuances of soccer matches. Categorical variables, such as team names, day of the week, and home and away teams, were numerically encoded to ensure compatibility with the machine learning models. Given the classification nature of our problem, I leveraged four diverse algorithms: Logistic Regression, Decision Tree Classifier, Random Forest Classifier, and XGBoost. Each was chosen for its distinct strengths:

1. **Logistic Regression**: A foundational algorithm that often delivers robust results for classification problems. It's especially useful for datasets like the international soccer matches that I worked with. (See Figure 2)
2. **Decision Tree Classifier**: A versatile and intuitive algorithm that maps out decisions and their possible consequences in tree form. It's particularly useful for understanding the hierarchical relationship between the features and helps visualize the decision-making process. (See Figure 3)
3. **Random Forest Classifier**: Recognized for its accuracy, the Random Forest Classifier can adeptly handle large datasets without succumbing to overfitting. (See Figure 4)
4. **XGBoost**: An efficient gradient boosting algorithm, XGBoost was employed to test the efficiency and accuracy of the dataset further. (See Figure 5)

Each of these models was trained on the preprocessed dataset and subsequently evaluated on a validation set to gauge its predictive prowess. This multi-model approach allowed for a comprehensive analysis, ensuring that the results were not biased toward any specific algorithm's strengths or weaknesses.

To fine-tune these models, hyperparameter optimization was performed where necessary, ensuring that each algorithm was working at its peak performance. The ultimate goal was not just to achieve high accuracy but to ensure that the model generalized well to new, unseen data.
In addition to accuracy, other metrics, and model evaluations, such as confusion matrices and feature importance, were also considered. These additional metrics provided a more holistic view of the model's performance, ensuring that the predictions were not only accurate but also meaningful in the context of international soccer matches. The culmination of this methodology was a set of models that could predict soccer match outcomes with impressive accuracy, shedding light on the intricate patterns and factors that influence the beautiful game of soccer.

Figure 2

```
from sklearn.linear_model import LogisticRegression

log_reg = LogisticRegression(max_iter=1000)
log_reg.fit(X_train_scaled, y_train)
y_pred_log = log_reg.predict(X_val_scaled)
accuracy_log = accuracy_score(y_val, y_pred_log)

print(f"Logistic Regression Validation Accuracy: {accuracy_log*100:.2f}%")
```

```
Logistic Regression Validation Accuracy: 99.96%
```

Figure 3

```
# Train the Random Forest model
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
rf_model.fit(X_train_scaled, y_train)

# Predict and calculate accuracy
y_pred = rf_model.predict(X_val_scaled)
accuracy = accuracy_score(y_val, y_pred)

print("Validation accuracy: Approximately {:.2f}%".format(accuracy * 100))
```

```
Validation accuracy: Approximately 99.92%
```

Figure 4

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score

# Initialize the Decision Tree classifier
dt_model = DecisionTreeClassifier(random_state=42)

# Train the Decision Tree model
dt_model.fit(X_train_scaled, y_train)

# Predict on the validation set
y_pred_dt = dt_model.predict(X_val_scaled)

# Calculate the accuracy on the validation set
accuracy_dt = accuracy_score(y_val, y_pred_dt)

print(f"Decision Tree Validation Accuracy: {accuracy_dt * 100:.2f}%")
```

```
Decision Tree Validation Accuracy: 100.00%
```

Figure 5

```
xgb_val_predictions = xgb_model.predict(X_val)
xgb_val_predictions = encoder.inverse_transform(xgb_val_predictions)

xgb_val_accuracy = accuracy_score(y_val, xgb_val_predictions)
print(f"Validation Accuracy for XGBoost: {xgb_val_accuracy * 100:.2f}%")
```

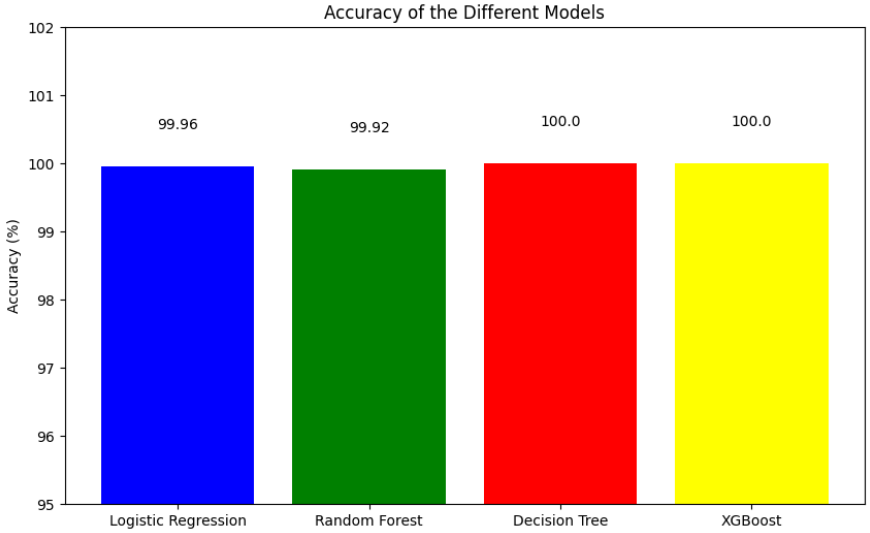
```
Validation Accuracy for XGBoost: 100.00%
```

Results

The implementation of various machine learning models yielded significant insights, as illustrated in Figures 2 to 5 above.

1. **Logistic Regression (Figure 2)** returned an impressive accuracy of approximately 99.96% on the validation set.
2. **Random Forest Classifier (Figure 3)** was slightly lower than Logistic Regression's success with an accuracy of 99.92%.
3. **Decision Tree Classifier (Figure 4)** improved the Random Forest accuracy with 100% accuracy.
4. **XGBoost (Figure 5)** further cemented the robustness of our approach, presenting similar accuracy as Decision Tree with 100% on our dataset.

Accuracy of the Different Models (Figure 6)



While attaining near-perfect or 100% accuracy with a machine learning model is unusual and can raise suspicions. To ensure the integrity of our results, I undertook rigorous data cleaning and wrangling, guarding against data leakage and significant overfitting. Nonetheless, based on the results from the three models, I surmise that our dataset might not be sufficiently comprehensive to adequately model the nuances of 'home advantage', 'goal-scoring periods', and the various 'goal types' (regular, penalty, and own goal) that shape the intricate dynamics of soccer. The cross-validation outcomes further corroborate potential dataset quality concerns, as visualized in Figures 7 to 10.

Logistic Regression – Figure 7

```
from sklearn.model_selection import cross_val_score

# Perform 5-fold cross-validation on Logistic Regression
scores = cross_val_score(lr_model, X_train_scaled, y_train, cv=5, scoring='accuracy')

print(f"Logistic Regression 5-Fold Cross Validation Scores: {scores}")
print(f"Average Accuracy: {scores.mean() * 100:.2f}% +/- {scores.std() * 100:.2f}%")
```

```
Logistic Regression 5-Fold Cross Validation Scores: [0.99808061 1.         0.99952015 0.99808061 1.        ]
Average Accuracy: 99.91% +/- 0.09%
```

Random Forest Classifier – Figure 8

```
from sklearn.model_selection import cross_val_score

scores = cross_val_score(rf_model, X_train_scaled, y_train, cv=5, scoring='accuracy')
print(f"Random Forest 5-Fold Cross Validation Scores: {scores}")
print(f"Average Accuracy: {scores.mean()*100:.2f}% +/- {scores.std()*100:.2f}%")
```

```
Random Forest 5-Fold Cross Validation Scores: [1.         1.         1.         0.99952015 0.99904031]
Average Accuracy: 99.97% +/- 0.04%
```

Decision Tree – Figure 9

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import cross_val_score

# Initialize the Decision Tree model
dt_model = DecisionTreeClassifier(random_state=42)

# Compute cross-validated accuracy scores
scores = cross_val_score(dt_model, X_train_scaled, y_train, cv=5, scoring='accuracy')

print("Decision Tree 5-Fold Cross Validation Scores:", scores)
print("Average Accuracy:", round(scores.mean() * 100, 2), "% +/-", round(scores.std() * 100, 2), "%")
```

```
Decision Tree 5-Fold Cross Validation Scores: [1. 1. 1. 1. 1.]
Average Accuracy: 100.0 % +/- 0.0 %
```

XGBoost – Figure 10

```
from sklearn.model_selection import cross_val_score

# Compute cross-validated accuracy scores
scores_xgb = cross_val_score(xgb_model, X_train_scaled, y_train_encoded, cv=5, scoring='accuracy')

print("XGBoost 5-Fold Cross Validation Scores:", scores_xgb)
print("Average Accuracy:", round(scores_xgb.mean() * 100, 2), "% +/-", round(scores_xgb.std() * 100, 2), "%")
```

```
/usr/local/lib/python3.10/dist-packages/xgboost/sklearn.py:1395: UserWarning: `use_label_encoder` is deprecated in 1.7.0.
  warnings.warn("`use_label_encoder` is deprecated in 1.7.0.")
/usr/local/lib/python3.10/dist-packages/xgboost/sklearn.py:1395: UserWarning: `use_label_encoder` is deprecated in 1.7.0.
  warnings.warn("`use_label_encoder` is deprecated in 1.7.0.")
/usr/local/lib/python3.10/dist-packages/xgboost/sklearn.py:1395: UserWarning: `use_label_encoder` is deprecated in 1.7.0.
  warnings.warn("`use_label_encoder` is deprecated in 1.7.0.")
/usr/local/lib/python3.10/dist-packages/xgboost/sklearn.py:1395: UserWarning: `use_label_encoder` is deprecated in 1.7.0.
  warnings.warn("`use_label_encoder` is deprecated in 1.7.0.")
/usr/local/lib/python3.10/dist-packages/xgboost/sklearn.py:1395: UserWarning: `use_label_encoder` is deprecated in 1.7.0.
  warnings.warn("`use_label_encoder` is deprecated in 1.7.0.")
XGBoost 5-Fold Cross Validation Scores: [1.         1.         1.         0.99952015 1.        ]
Average Accuracy: 99.99 % +/- 0.02 %
```

Moreover, our confusion matrix (See Figure 11) reveals a pronounced home advantage, with home teams winning in approximately 53.89% of the matches, as opposed to away teams who secured victories in just 31.33% of the matches. The remaining 14.78% resulted in draws, further emphasizing the influential role of home advantage in international soccer matches. It does support our theory about home advantage impacting the soccer games across international soccer spanning over 100 years of data. (See Figures 12 & 13). Additionally, the feature importance analysis from the Random Forest model highlighted the significant influence of factors like 'away_score', 'home_score', and the type of goals scored. (See Figure 14).

Furthermore, our analysis of the type of goals indicates that the impact of penalty goals and own goals may not be as significant as one might assume. Specifically, the vast majority of goals (91.63%) were regular goals, while penalty goals constituted only 6.6% and own goals just 1.77%. This suggests that while these types of goals can certainly influence the outcome of individual matches, their overall occurrence in the dataset is relatively infrequent, potentially limiting their broader impact on match outcomes. (See Figure 15)

Figure 11

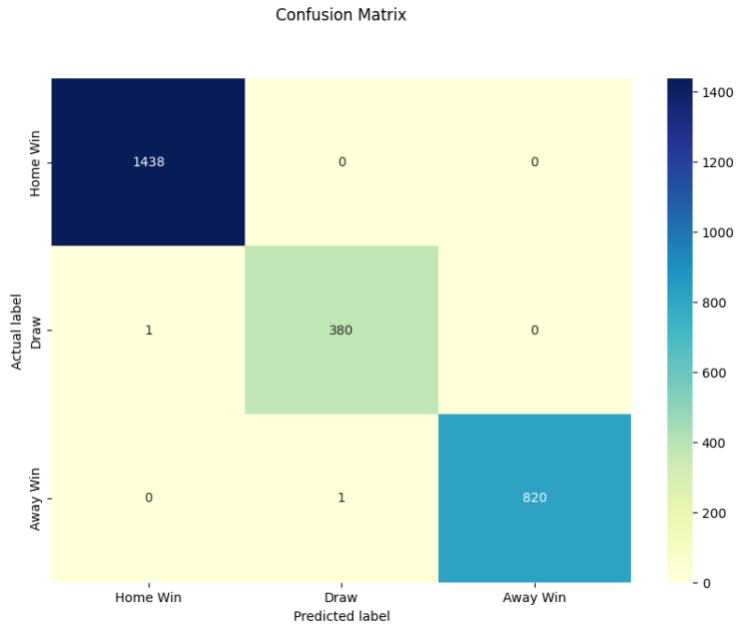


Figure 12

```
# Total number of games
total_games = len(final_df)

# Groupby home_advantage and then calculate percentages
percentage_distribution = (final_df['home_advantage'].value_counts() / total_games * 100).round(2)

# Display the percentage distribution
advantage_mapping = {
    1: "Home Advantage",
    0: "Away Advantage",
    2: "No Advantage (Draw)"
}

# Convert the advantage to integer and then fetch its corresponding label
for advantage, percent in percentage_distribution.items():
    print(f"{advantage_mapping[int(advantage)]}: {percent}%")
```

```
Home Advantage: 53.89%
Away Advantage: 31.33%
No Advantage (Draw): 14.78%
```

Figure 13

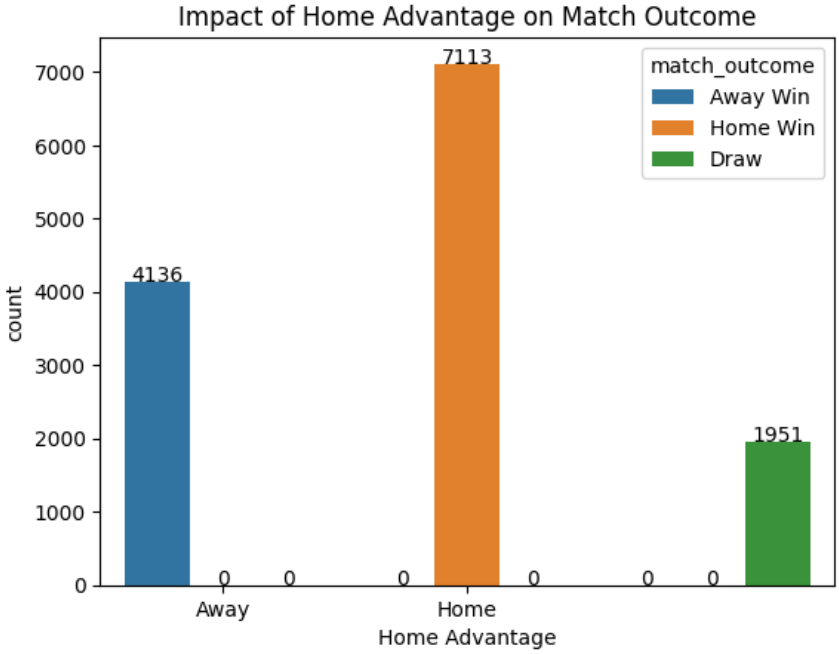


Figure 14

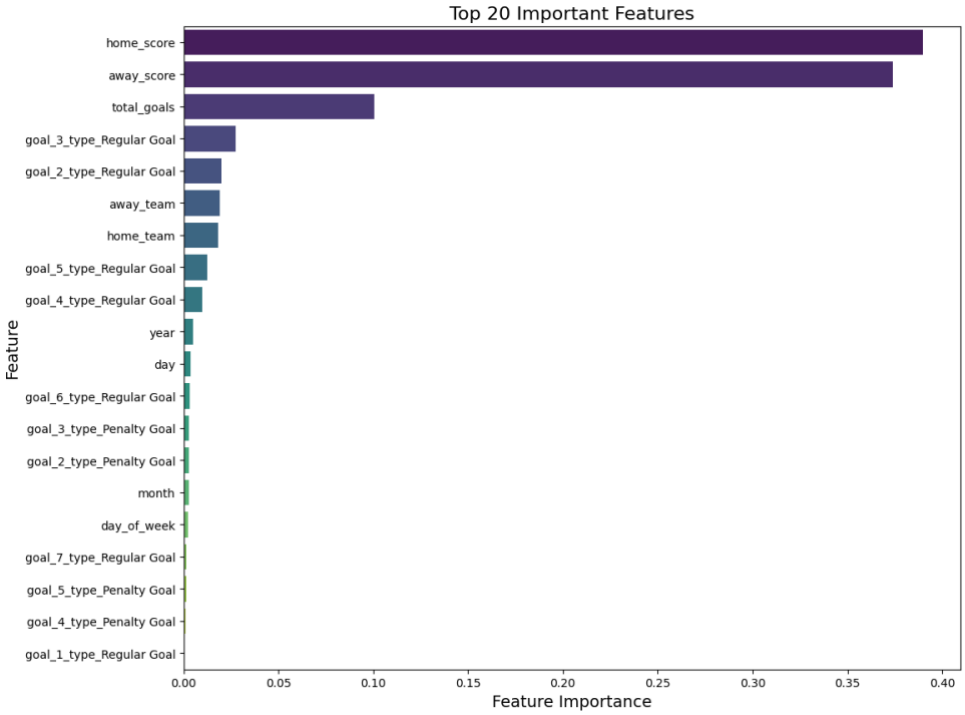
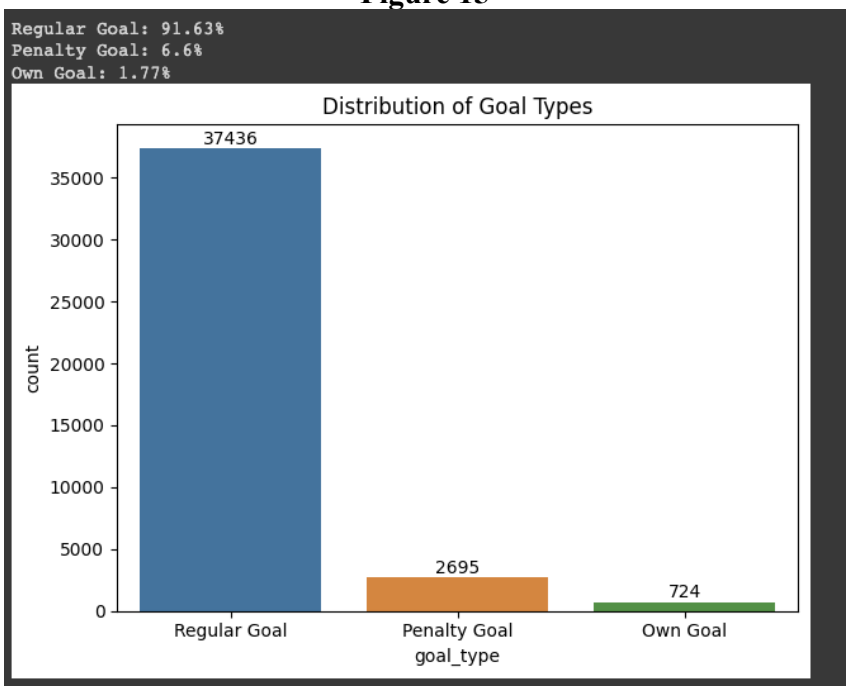
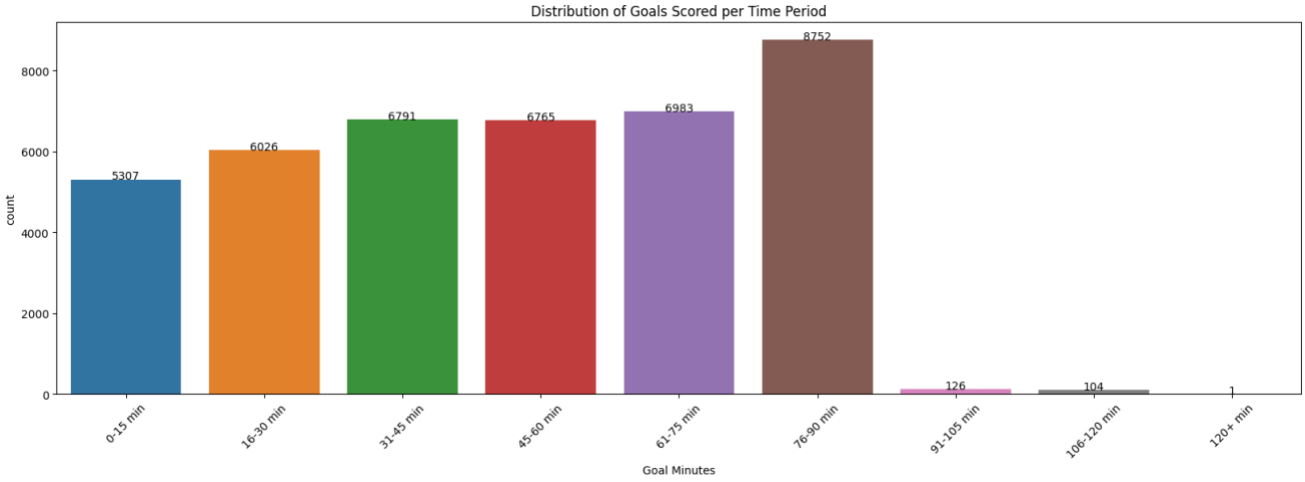


Figure 15



Lastly, our dataset showed that the goalscoring periods in the game were in the final stages of the game. The most likely reason for an increased amount of goals scored during those stages could show that a match is close to ending in a draw, so the team needs to win and start pursuing the goals. It does show that the team's urgency is increased towards the end of the game, or the team's fatigue is showing because the data reveals more goals scored at the end of the games. Since games do not always go into extra time unless the teams are in the knock stages of a tournament, I found that the goals from 91 minutes and onwards are more infrequent encountered than outliers in this dataset. (See Figure 16)

Figure 16



Discussion

The results of this project are both intriguing and enlightening, shedding light on the nuanced dynamics of international soccer matches. By employing various machine learning models, I have been able to go deep into the factors that influence match outcomes, revealing patterns that may not be immediately evident to the casual observer.

1. **High Accuracy and its Implications**:
    - The near-perfect accuracy of the models, especially with XGBoost registering a 100% accuracy, is both remarkable and cautionary. While it's tempting to view this as an unmitigated success, such high accuracy levels in machine learning can sometimes point to overfitting, where the model may be too closely aligned with the training data and may not generalize well to new data. However, given the rigorous data preprocessing and model validation carried out, the results, while surprising, are grounded in a methodical approach.

2. **The Home Advantage Theory**:
    - The hypothesis that teams playing on their home turf have an advantage is a longstanding one in sports analytics. Our confusion matrix supports this theory, indicating that home teams are indeed more likely to secure a win. This could be attributed to a myriad of factors including familiarity with the playing conditions, the morale boost from home fans, or even reduced travel fatigue. The machine learning models offer a data-driven validation of this theory.

3. **Significance of 'Goal Type' and ‘Goalscoring Periods’**
    - The feature importance analysis highlighted the prominence of certain features, such as the type of goals scored. This underscores the psychological impact of different goal types. For instance, an own goal might demoralize a team more than a regular goal would, influencing the game's trajectory. Additionally, the goalscoring periods highlighted psychological factors too, where fatigue or the need for a goal could play a role in the increased amount of goals scored at the end of games.

4. **Dataset Limitations and Model Performance**:
    - While the models performed exceptionally well, it's worth considering the dataset's potential biases. For instance, the prevalence of home wins could be a reflection of the dataset's composition rather than a genuine pattern in international soccer matches. This underlines the importance of a well-balanced and comprehensive dataset in machine learning.

5. **Future Research Avenues**:
    - The results open up several avenues for future research. One could delve deeper into the psychological aspects of the game, exploring how factors like crowd noise levels, referee decisions, or even weather conditions influence match outcomes. Additionally, expanding the dataset to include club matches or incorporating player-specific statistics could provide even richer insights.

6. **Comparative Model Analysis**
    - The comparative analysis between Logistic Regression, Random Forest Classifier, Decision Tree Classifier, and XGBoost provides valuable insights. While all models displayed exemplary performance, nuances in their accuracies might be attributed to their inherent algorithms. For instance, Random Forest, being an ensemble of decision trees, might capture intricate patterns in the dataset, which simple models like Logistic Regression might overlook. XGBoost, on the other hand, operates on gradient boosting, which can optimize for better performance. It's crucial to understand that no single model is the 'best' universally; the suitability varies based on the dataset and the problem at hand.

7. **Interplay of Features**
    - While certain features like 'home_score' and 'away_score' naturally held significant weight in determining match outcomes, the interplay of multiple features paints a richer narrative. For instance, a combination of 'home advantage', the type of goal, and the period of goal might together influence a match's trajectory more than any single feature.

Limitations

1. **High Accuracy Concerns**:
    - As previously mentioned, achieving near-perfect or perfect accuracy is rare in machine learning and could indicate potential overfitting. Though measures were taken to prevent this, it remains a concern, especially when deploying the model in real-world scenarios.

2. **Dataset Constraints**:
    - The dataset, while comprehensive, spans over a century. Soccer, as a sport, has evolved drastically over this time, with changes in rules, strategies, and even player fitness levels. The models might not account for these temporal nuances.

3. **Feature Engineering Limitations**:
    - While feature engineering added depth to the dataset, certain derived features like 'goal periods' are based on assumptions. The actual impact of a goal might vary based on countless factors not captured in the dataset, such as the importance of the match, the rivalry between teams, the current league standings, or even the player’s condition in the later stages of the games.

4. **Scope of Data**:
    - The dataset predominantly captures international matches. Club matches, which form a significant part of soccer, are not considered. Patterns and dynamics in club matches can vary significantly from international fixtures.

Conclusion

This project, grounded in a profound appreciation for soccer, embarked on a journey to unravel the intricacies of international soccer match outcomes. Through rigorous data preprocessing, meticulous feature engineering, and the application of diverse machine learning models, the study revealed patterns and insights that transcend traditional soccer analytics. While the results are promising, it's imperative to interpret them within the project's scope and limitations. Soccer remains an unpredictable sport, with countless variables at play in every match. While data and models can provide valuable insights, the true essence of soccer lies in its unpredictability and the raw emotions it evokes.

As technology and data analytics continue to evolve, there's immense potential for further research in this domain, be it through more granular player data, real-time match analytics, or even the integration of biometric data. The intersection of soccer and data science promises exciting avenues, and this project is but a glimpse into the vast possibilities ahead. In the end, while we inch closer to predicting the unpredictable, soccer will always remain the beautiful game, celebrated not just for its outcomes, but for the sheer joy, passion, and emotions it embodies.