Yash Maheshwari (ym463)
Vaibhav Suresh (vs759)

Theoretical Report

UEFA EURO 2024 Players' Performance Analysis

Dataset(s):
https://www.kaggle.com/datasets/damirdizdarevic/uefa-euro-2024-players
https://www.kaggle.com/datasets/bugralpp/euro-2024-group-stage-fbref-scrape-data?select=shots_all.csv

Defining the Project: This project seeks to analyze player performance data from the UEFA Euro 2024 tournament to understand the factors influencing team success. By examining individual player metrics such as goals scored, assists, passing accuracy, and defensive actions, the project will identify patterns and correlations that contribute to team performance. This relates closely to the concepts discussed in the analytics and data-driven decision-making lectures, applying statistical methods to create a better experience for sports fans.

Summary: The top soccer players in Europe will represent their nations in the UEFA Euro 2024 soccer competition. However, it can be confusing for a new viewer to identify which players, besides household names, to focus on during such a large-scale event. Our goal is to solve this issue through data analysis, creating a model that identifies top players at each position on the pitch, and providing new users with insights into whichever game they are watching. We propose to quantify player abilities, focusing on several key metrics at a time, making it easy for the casual viewer to identify the players that provide the most impact/value on the field.

Purpose & Significance:

This proposal outlines a data-driven analysis of player performance forecasted for UEFA EURO 2024, aiming to provide insights into strategic aspects of team dynamics and individual player contributions and how these factors impact individual players market value.

Understanding player performance in Euro 2024 is crucial for several reasons. It allows coaches and analysts to optimize team selection and strategy based on empirical data rather than intuition and gamefilm alone. Sports industry management procedures now in place ignore the complex player contributions that are made by players by depending on subjective evaluations along with basic statistics.

As these competitions gain appeal on a global scale, sports data becomes increasingly important to study and work with. We hope this data can be beneficial to introduce fans across the world to different players/clubs in the tournament. These statistics also play a critical role in the international sports betting industry, including setting betting limits, moneyline bets, prop

bets, and individual statistic-based betting. Along with that, this practice can easily be translated to measure player metrics in other professional sports/leagues.

Initial plan

1. **Data Collection:** Player performance data (goals, assists, etc.) will be collected from UEFA EURO 2024 official records on Kaggle, including key metrics like goals per minute, using assists and saves to quantify impact at different positions on the field (examples of KPIs per position), etc.
    a. Data Preparation:
        i. Gather relevant data on team performance, individual player statistics, and other potentially influential factors.
        ii. Normalize or standardize the data if necessary.

2. **Models/Techniques/Algorithms:** Statistical analysis methods like regression analysis, will be employed to identify key performance indicators (KPIs) and their impact on team success.
    a. Regression Analysis:
        i. Use multiple linear regression to identify relationships between various factors and team success.
        ii. Ex: Player_Impact = $\beta 0 + \beta 1$(points_scored) + $\beta 2$(shots_on_goal) + $\beta 3$(net_positive_passes) + ... + ε

3. **Hypothesis:** The hypothesis is that certain player metrics (including but not limited to the ones mentioned above) significantly correlate with team success metrics, e.g., goals scored, points earned, and current/future performance in tournaments. These correlations will be tested using appropriate statistical tests. We can use this to predict The market value of the players and show who they players to watch out for are.

4. **Evaluation:** The success of the method will be measured by its ability to predict match outcomes based on player data and validate these predictions against actual match results, which we will be able to witness in real time as the UEFA tournament continues through the summer.
    a. Cross-Validation:
        i. Implement cross-validation techniques with the actual data from the tournament to ensure the robustness of the algorithms across different subsets of the data.

Important Files & Data

critical values are highlighted

Yash Maheshwari (ym463)
Vaibhav Suresh (vs759)

- **euro2024_players.csv**: 623 non-null objects in the following categories (and datatypes)
  - Name - object
  - … [6 other columns]
  - Goals - int64
  - MarketValue - int64
  - Country - object
- **shots_all.csv**: varied by category, but highlighted columns have 630 non-null objects
  - player - object
  - team - object
  - xg_shot - float64
  - psxg_shot - float64
  - outcome - object
  - distance - float64
  - … [7 other columns]
- **summary.csv**: 749 non-null objects in the following columns
  - … [4 other columns]
  - goals - int64
  - assists - int64
  - pens_made - int64 } use these attr to compute
  - pens_att - int64 } pen% = pens_made/pens_att
  - …
  - shots - int64 } use these attr to compute
  - shots_on_target - int64 } shot_on_target% = shots_on_target/shots
  - cards_yellow - int64
  - cards_red - int64
  - touches - float64
  - … [8 other columns]
  - passes_completed - float64
  - passes - float64
  - passes_pct - float64
  - progressive_passes - float64
  - carries - float64
  - progressive_carries - float64
  - take_ons - float64 } use these attr to compute
  - take_ons_won - float64 } take_on_win% = take_ons_won/take_ons
  - name - object
  - match - object

Example of applying linear regression **after JOINing the tables above**

```
# market value per goal
players_df['value_per_touch'] = players_df['MarketValue'] / players_df['touches']

from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score

# Define the features (X) and target (y)
X = players_df[['value_per_touch', 'goals', 'passes_pct', 'shot_on_target%', ...]]
y = players_df['MarketValue']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Create and train the regression model
model = LinearRegression()
model.fit(X_train, y_train)

# Make predictions
y_pred = model.predict(X_test)

# Evaluate the model
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)

print("Mean Squared Error:", mse)
print("R-squared:", r2)
# Merge summary data with players data to get market values
merged_df = pd.merge(summary_df, players_df, left_on='shirtnumber', right_on='Name',
how='left')

# Calculate goals per minute
merged_df['minutes_per_goal'] = merged_df['minutes'] / merged_df['goals']

features = ['minutes_per_goal', 'assists', 'saves', 'shots', 'xg', 'npxg',
'key_passes', 'progressive_passes']
merged_df = merged_df.dropna(subset=features)

# Define the features (X) and target (y)
X = merged_df[features]
y = merged_df['MarketValue']
```

```
# Display the first few rows of the feature set
X.head(), y.head()
```

this is **completely theoretical, and there would have to be a large amount of data cleaning and stripping if we were to JOIN all three data sets together**

Cross-Validation (as mentioned above)

```
from sklearn.model_selection import cross_val_score

# Perform cross-validation
cv_scores = cross_val_score(model, X, y, cv=5)

print("Cross-Validation Scores:", cv_scores)
print("Mean CV Score:", cv_scores.mean())
```

would be possible after the predictive data has been stripped clean, normalized, and put through the linear regression **AND the data from the actual result of the tournament has been collected to cross-reference with**

Applications of the above method(s)

- The first method takes into consideration how many touches, goals, passes and shots are valued on target to predict the market value of a player. We take these variables and make a linear regression model and then check them using mean squared error to the actual market value of the a player to maybe see if there is a reason why a player is worth so much.
- We also use our own attribute which is minutes per goal. The less minutes between goals also adds value to the player as to how efficient they are. We add all these variables and create a feature.
- We also used a cross validation method to predict new data and to test the learning models efficiency.

Outputs

Player Name	Position	Country	Predicted Market Value
Rodri	Defensive Midfield	Spain	67,225,404.95
Pedri	Central Midfield	Spain	82,808,963.29
Vitinha	Central Midfield	Portugal	76,321,610.58

Here are 3 players with predicted market values using our model

Yash Maheshwari (ym463)
Vaibhav Suresh (vs759)

There actual market values are :

Player Name	Position	Country	Market Value
Rodri	Defensive Midfield	Spain	120,000,000
Pedri	Central Midfield	Spain	80,000,000
Vitinha	Central Midfield	Portugal	50,000,000

Conclusion

- Our predicted model is not completely accurate and there could be more factors to then there market value
- But our model predicts the ability of the players and gives them a higher market value for their skills. This help to know who the best players are in the league. The higher the market value the better the player according to our model.