Predictive Insights For Financial Recovery In Club Football

Manas Kalangan - 50608803, Sahil Kakad - 50607550

Abstract

Football Club Barcelona is facing severe financial difficulties, driven by underperformance on the field and declining evaluations in the transfer market. This project leverages advanced data analytics and machine learning techniques to uncover actionable insights that can support the club's financial recovery. By analyzing player valuations, transfer trends, and competition performance, this study highlights opportunities for smarter decision-making in recruitment, sales, and match strategies. The analysis also identifies undervalued players, evaluates key attributes driving transfer fees, and optimizes player lineups for greater financial and competitive returns. Utilizing models like Random Forest, XGBoost, and Gradient Boosting, the findings provide a data-driven framework for financial recovery and strategic improvement, enabling the club to align its decisions with revenue growth and performance excellence.

Introduction

Football Club Barcelona, one of the most iconic names in global football, is currently grappling with a financial crisis. The club's struggles stem from a combination of underwhelming on-field performance and diminishing evaluations in the transfer market. As revenue streams falter and costs mount, Barcelona faces the urgent challenge of realigning its strategies to ensure financial stability while maintaining competitive integrity.

This project explores the potential of data-driven decision-making to aid Barcelona's recovery. Leveraging historical and current data, we aim to identify key insights into player market valuations, performance metrics, and the financial dynamics of matches and competitions. By understanding these factors, the club can optimize its player recruitment, transfers, and match-day strategies.

Motivation

The project "Analyzing Player Valuations to Optimize Financial Recovery" highlights the transformative power of data-driven decision-making in football. With Football Club Barcelona facing financial challenges, this project seeks to leverage analytics and machine learning to address key issues, making significant contributions to the club's financial recovery and strategic optimization. The motivation behind this project lies in its potential to offer actionable insights that can influence not only FC Barcelona but also broader practices in football management.

Optimizing Transfer Market Decisions: By identifying undervalued players, the project provides a framework for smarter investments and sales in the transfer market. These insights can drive profitability and financial stability, ensuring sustainable operations for FC Barcelona.

Enhancing Recruitment and Development Strategies: Understanding which player attributes and performance metrics drive transfer fees allows clubs to refine their recruitment strategies. The findings can also guide player development, focusing on attributes that offer the greatest financial and competitive returns.

Maximizing Financial Returns from Competitions: The analysis of attendance and performance trends identifies which competitions are most profitable. Prioritizing high-revenue matches ensures efficient resource allocation and revenue generation, enabling clubs to balance finances effectively.

Strategically Optimizing Team Lineups: Insights into player combinations that maximize match success and attendance can significantly influence team selection. By aligning performance goals with financial returns, the project empowers clubs to create high-impact strategies.

Driving Evidence-Based Decision-Making: With data as its foundation, this project promotes evidence-based policymaking in football management. From transfer policies to match-day strategies, data-driven insights ensure decisions align with financial goals.

Addressing Financial Sustainability: The project directly tackles FC Barcelona's financial instability by offering scalable solutions that blend analytics and domain expertise. These solutions are vital for achieving long-term sustainability in an increasingly competitive football landscape.

In essence, this project is motivated by the aspiration to harness the power of analytics and machine learning to create a financially sustainable and strategically sound framework for football management. By addressing key challenges and unlocking actionable insights, this study aims to contribute to FC Barcelona's resurgence as both a competitive and financially robust football club.

Data Retrieval

To conduct this study, multiple datasets were sourced and analyzed to uncover actionable insights into player valuations, transfer trends, and team optimization strategies for Football Club Barcelona. These datasets were hosted on Kaggle and required manual download. The extracted files were stored in the same directory as the project notebook for seamless access and analysis:

Datasets Used

Player-Data
The dataset **players.csv** provided detailed information about individual players, including attributes like age, position, nationality, and historical performance metrics. This dataset served as the cornerstone for identifying undervalued players and analyzing player attributes related to market value and transfer trends.

Player-Valuations
The **player_valuations.csv** dataset offered insights into current and historical market valuations for players. Key columns such as market_value_in_eur and highest_market_value_in_eur were used to evaluate discrepancies in player valuations and detect undervalued assets.

Transfer-Data

The **transfers.csv** dataset detailed historical transfer activity, including fees, involved clubs, and player details. This dataset was crucial for correlating player attributes with transfer fees and understanding market dynamics.

Competition and Match Data

- **competitions.csv** and **games.csv** provided comprehensive records of competition types, match outcomes, and performance metrics. These datasets were integral to identifying competitions with the highest financial returns and analyzing team performance trends.
- **appearances.csv** and **game_lineups.csv** captured player participation in matches and lineup configurations, enabling analysis of optimal player combinations and their impact on match outcomes and attendance.

Match-Events

The **game_events.csv** dataset tracked in-game events such as goals, assists, and cards, offering granular insights into player performance and match dynamics.

Attendance-Data

Information on attendance figures and stadium capacities from **club_games.csv** supported the analysis of revenue patterns and the financial impact of high-attendance matches.

Integration of Datasets

Integration of Datasets

The integration of these datasets enabled a holistic analysis that combined player performance, financial metrics, and match outcomes. This approach facilitated:

- Identifying undervalued players by correlating player attributes with their market values.

- Understanding the impact of demographic factors such as age and position on transfer fees and performance metrics.

- Bridging attendance and revenue data to prioritize competitions and optimize match-day strategies.

Data Cleaning

To ensure consistency, accuracy, and reliability in the analysis, extensive data cleaning techniques were applied to the datasets. These steps were critical for preparing the data for exploratory analysis and predictive modelling. Below is a detailed breakdown of the data cleaning processes:

Removing Missing Values

Rows with missing values in critical columns were dropped to ensure the validity of the analysis. For example:

- Players with missing values in market_value_in_eur, highest_market_value_in_eur, and position were excluded.
- In the transfers dataset, rows with missing values in transfer_fee were removed.
- Attendance values in match data were imputed using the median to address sparsity in this column.

Fixing Data Formats

- **Date Columns**: Columns such as date_of_birth, transfer_date, and game_date were converted to datetime format using pd.to_datetime() for consistent date-based calculations.

- **Numerical Columns**: Columns like market_value_in_eur, highest_market_value_in_eur, and minutes_played were converted to numeric types using pd.to_numeric() to ensure compatibility with statistical analysis and modeling.

Handling Duplicates
Duplicate rows were identified and removed across datasets to ensure the uniqueness of each entry:

- Players were deduplicated using player_id.
- Player valuation entries were deduplicated based on player_id and date.
- Transfers were deduplicated using a combination of player_id and transfer_date.

Merging Relational Tables
To create a comprehensive dataset, relational tables were merged on shared keys:

- Player information was merged with player valuations on player_id.
- Match appearances were merged with game data using game_id.
- Match data was integrated with competition information to provide a holistic view of game outcomes and contextual factors.

Standardizing Categorical Variables
Categorical columns like position, sub_position, and hosting were standardized:

- Text-based fields were converted to lowercase and stripped of extra spaces for uniformity.
- Binary fields like hosting were encoded numerically for compatibility with machine learning models.

Imputation of Missing Values
Missing values in non-critical columns were imputed using statistical methods:

- Median values were used for numerical fields like attendance in game data.
- Contextually appropriate defaults were applied to categorical fields where feasible.

Normalization
Numerical columns with large variances in scale, such as market_value_in_eur and highest_market_value_in_eur, were normalized to prevent dominant features from skewing the analysis.

Ensuring Data Integrity
Throughout the cleaning process, data integrity was maintained by verifying:

- Column data types for compatibility with analysis.
- Logical consistency between columns (e.g., transfer_date after date_of_birth).
- Completeness of merged datasets to avoid introducing bias or missing key records.

These cleaning and preprocessing steps ensured that the datasets were ready for rigorous analysis and machine learning applications, enabling accurate and actionable insights.

Problem Statements

Question 1- Sahil

Which players are undervalued in the transfer market and could potentially yield higher returns in the future?
Why This Question:
After considering datasets like **players**,

players_valuation, and **transfers**, we identified key columns such as market_value_in_eur, highest_market_value_in_eur, and transfer_fee. By comparing current market values, we can determine which players are undervalued and predict their future market growth.

Why This Matters:
Identifying undervalued players allows clubs to make smarter, more profitable transfer decisions, which could improve the club's financial situation and future competitive performance.

- **Hypothesis 1**: **FC Barcelona's current market standing compared to other European clubs**
  This hypothesis aims to compare FC Barcelona's market value to that of other European clubs. We will analyze if Barcelona's market is underperforming due to recent factors like poor team performance or a decrease in overall market visibility.

- **Hypothesis 2**: **Age's impact on a player's market valuation**
  We aim to analyze how a player's age impacts their market value. A scatter plot comparing **age vs market value** will help to understand if younger players are undervalued and have the potential for growth.

Question 2- Manas

How do different player attributes and performance metrics impact transfer fees?
Why This Question:
By analyzing attributes like **goals**, **assists**, and **minutes_played** from the **players**, **transfers**, and **appearances** datasets, we seek to understand the correlation between player performance metrics and their transfer fees. This knowledge will help FC Barcelona optimize its recruitment and selling strategies based on these key factors.

Why This Matters:
Understanding which attributes influence transfer fees can help refine the club's recruitment strategy, as well as improve the financial outcomes of player sales.

- **Hypothesis 1**: **Performance values and real-world transfer fees correlation**
  We will examine the correlation between performance metrics (goals, assists, minutes played) and real-world transfer fees using a scatter plot to determine how well performance metrics predict actual transfer fees.

- **Hypothesis 2**: **The impact of goals and assists on transfer fees**
  This hypothesis will explore if **goals** and **assists** correlate with higher transfer fees. A scatter plot of **transfer fee vs goals/assists** will help identify if one metric has a stronger influence than the other.

- **Hypothesis 3**: **Analyzing market trends over time to understand the current standings**
  We will analyze market trends over a period of time to understand how player valuations and transfer fees have evolved, especially in relation to performance metrics like goals and assists.

Question 3 - Manas

Which competitions or matches are the most financially beneficial for the club?
Why This Question:
Given the club's financial goals, identifying competitions or matches that attract the most

attendance or generate higher revenue is crucial. By analyzing **competition type**, **match performance**, and **attendance data**, we aim to determine which competitions provide the highest financial returns for FC Barcelona.

Why This Matters:
Focusing on improving performance in the most financially rewarding competitions can enhance the club's revenue generation and overall financial strategy.

- **Hypothesis 1**: **Competitions with higher attendance lead to greater financial gains**
  This hypothesis will analyze the relationship between **attendance** and **financial metrics** (revenue), illustrating how matches with higher attendance correlate with increased financial success. A scatter plot will visualize this relationship.

- **Hypothesis 2**: **Teams that win more matches earn higher financial rewards**
  By analyzing the impact of **win rates** on financial returns, we aim to show that teams with more successful records tend to generate greater revenue through ticket sales, sponsorship, and other match-day revenues.

Question 4 - Sahil

Which player combinations and lineups perform the best in terms of match success and financial returns?
Why This Question:
By analyzing **player performance data** (goals, assists, minutes played), **lineup details** (positions, captains), and **match outcomes** (attendance, goals scored), we aim to determine which specific player combinations and lineups result in the best match performance and highest financial returns for FC Barcelona.

Why This Matters:
Optimizing team lineups based on both **performance** and **financial impact** can help the club maximize returns from match success, ensuring better financial performance in the future.

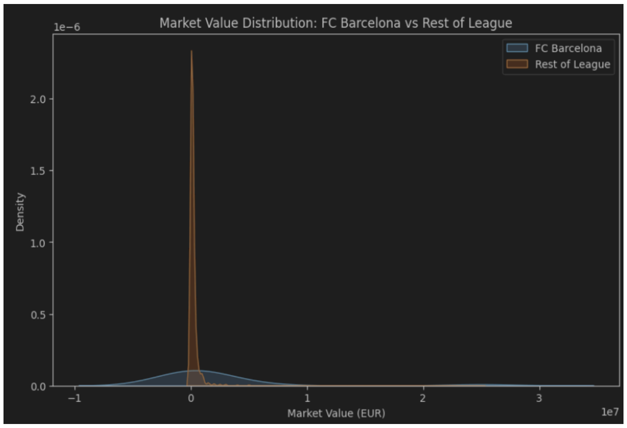
- **Hypothesis 1**: **High-attendance games are associated with specific player lineups**
  We hypothesize that certain player combinations, especially those with **star players** or effective formations, attract more spectators, leading to higher match-day attendance. A **bar plot** showing **average attendance** for different lineups will help us visualize this relationship.

Exploratory Data Analysis

EDA 1 - Sahil

Market Value Comparison: FC Barcelona vs the Rest of the League

Insights:
This analysis compares FC Barcelona’s market value to that of other top European clubs. The **market value distribution** plot clearly demonstrates that FC Barcelona holds a unique and dominant position in the European market. It shows a higher concentration of highly valued players compared to other clubs. This suggests that Barcelona tends to acquire and maintain high-value players, regardless of their performance or age.



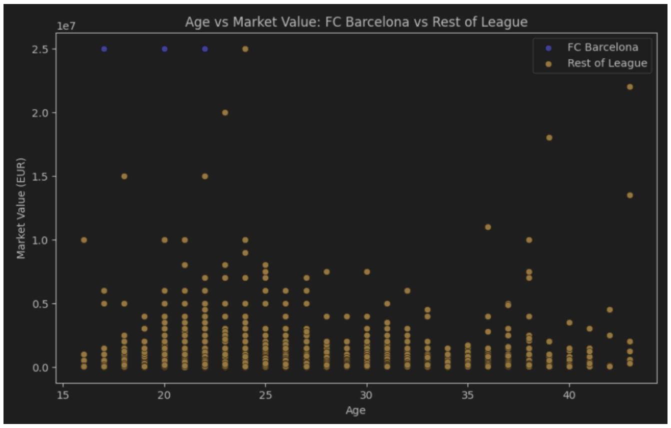
Inference:
Overall, the market value distribution plot supports the hypothesis that FC Barcelona's market value is concentrated in the higher range, especially in comparison to other European clubs. This reflects the club's strategy of targeting high-value players and maintaining a competitive advantage in the market.

EDA 2 - Sahil

Age vs. Market Value: FC Barcelona vs the Rest of the League

Insights:
The scatter plot between **age** and **market value** reveals some important trends:

- There is a **positive correlation** between age and market value, suggesting that older players typically have higher valuations.
- However, **FC Barcelona** shows a noticeable deviation from this general trend, with several **younger players** having exceptionally high market values. This indicates that FC Barcelona invests in young talent with high potential and a significant market appeal.



Inference:

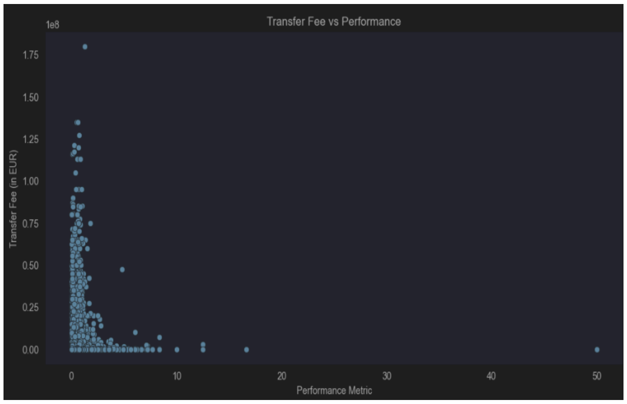
General Trend: The positive correlation between age and market value for both FC Barcelona and other European clubs indicates that older players typically command higher market values.

FC Barcelona's Strategy: The presence of **young high-value players** in FC Barcelona's squad supports the hypothesis that the club is focusing on acquiring and developing young, talented players. This aligns with their strategy of investing in promising talent rather than focusing solely on experienced players.

EDA 3 - Manas

Transfer Fee vs. Performance

Insights:
The scatter plot between **transfer fees** and **performance metrics** (such as goals, assists, etc.) does not show a strong correlation. The data points are scattered widely, indicating that transfer fees do not always correlate well with the player's subsequent performance. There are instances where players with high transfer fees have underperformed, and others who were transferred for lower fees have exceeded expectations.



Inference:

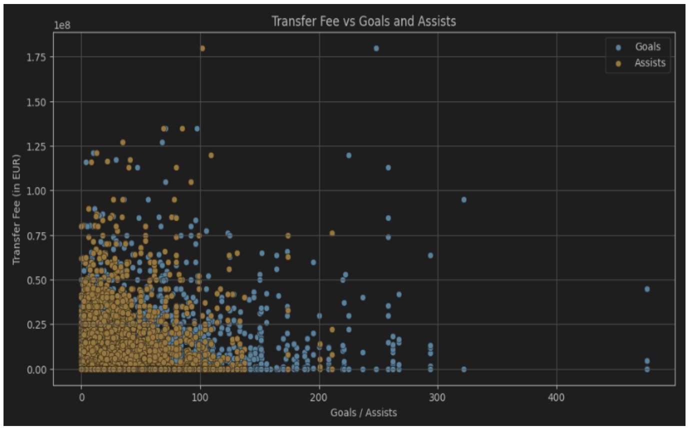
No Strong Correlation: The plot shows that **transfer fees** are not a reliable predictor of a player's **performance**. This suggests that while clubs may pay high transfer fees for certain players, other factors such as potential, team fit, and market dynamics play a significant role in their future performance.

Other Factors: Factors like **team dynamics**, **age**, and **playing style** appear to influence performance more than the initial transfer fee, challenging the assumption that high transfer fees equate to high performance.

EDA 4 - Manas

Transfer Fee vs Goals and Assists

Insights:
The scatter plot between **transfer fee** and **goals/assists** shows a **positive correlation**, indicating that players who contribute significantly to **goals** and **assists** tend to have higher transfer fees. However, there are some **outliers**, where players with high transfer fees have not performed well in terms of goals or assists, suggesting that **other factors** also play a role in determining transfer fees.



Inference:

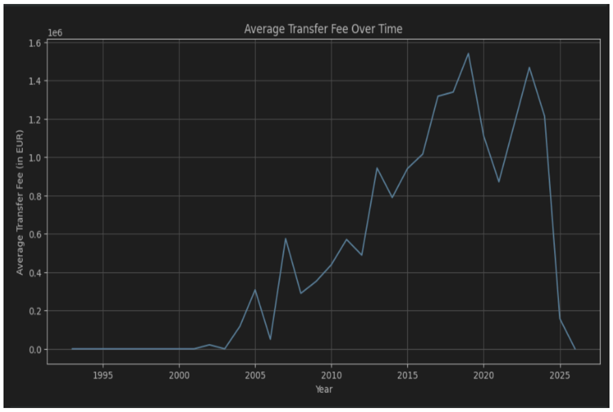
Goals vs. Assists: The analysis suggests that **goals** might be slightly more important than **assists** in determining a player's transfer fee. Players who score more goals are typically valued higher in the transfer market.

Other Influencing Factors: While goals and assists are significant contributors to transfer fees, **market demand**, **player potential**, and **age** also appear to be key factors that influence the final transfer fee.

EDA 5 - Sahil

Average Transfer Fee Over Time

Insights:
The line plot showing the **average transfer fee** over time reveals an **upward trend**, with a significant rise in the early 2000s and 2010s, followed by a slight decline in recent years. This upward trend can be attributed to factors such as the rise of the Premier League, increased club revenues, and higher player marketability. The recent decline may be linked to external factors like the **COVID-19 pandemic** and changing financial conditions in football.



Inference:

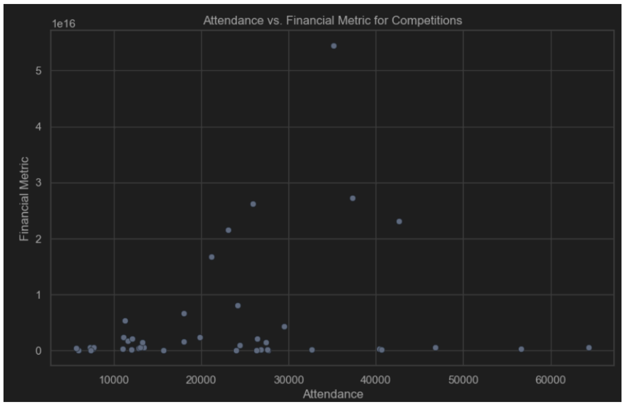
Upward Trend: The increase in **average transfer fees** reflects the growing financial power of football clubs and the increasing demand for high-value players.

Recent Decline: The slight decrease in average transfer fees in recent years could be due to economic disruptions such as the **COVID-19 pandemic** and evolving financial regulations. The data highlights how market conditions influence player valuations over time.

EDA 6 - Manas

Attendance vs. Financial Metric

Insights:
The scatter plot between **attendance** and **financial metrics** shows a **positive correlation**, suggesting that higher attendance generally results in greater financial returns for the club. However, the data points are spread out, indicating that **other factors**, such as ticket prices and sponsorship deals, also influence the financial success of a match.



Inference:

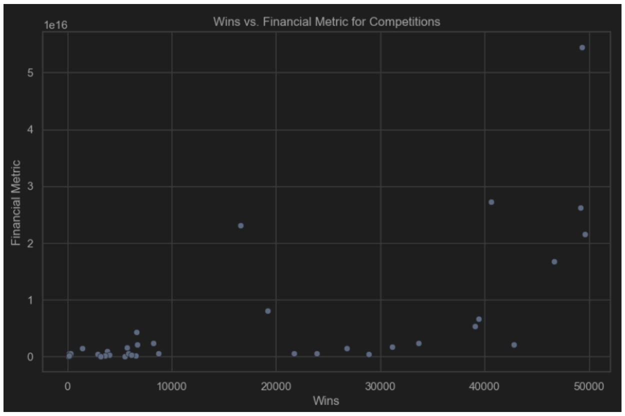
Attendance Impact: The plot supports the hypothesis that higher **attendance** correlates with **greater financial gains**. Matches with higher attendance contribute more to the club's financial success.

Other Factors: The spread of data points indicates that **attendance** is not the only factor influencing financial performance. Other variables such as **sponsorship revenue**, **ticket prices**, and **merchandise sales** play a significant role in the overall financial outcome of a match.

EDA 7 - Manas

Wins vs. Financial Metric

Insights:
The scatter plot between **wins** and **financial metrics** shows a **positive correlation**, suggesting that teams that win more matches tend to have higher financial returns. However, the data also highlights some outliers where teams with a high number of wins have lower financial returns, indicating that factors such as **marketability**, **fan base**, and **broadcast revenue** also play important roles.



Inference:

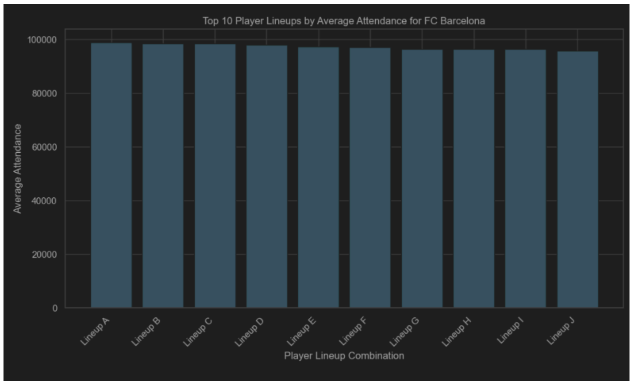
Wins Impact: The plot supports the hypothesis that **winning matches** contributes to higher **financial rewards**. Teams that win more tend to attract more sponsorship, media attention, and higher ticket sales.

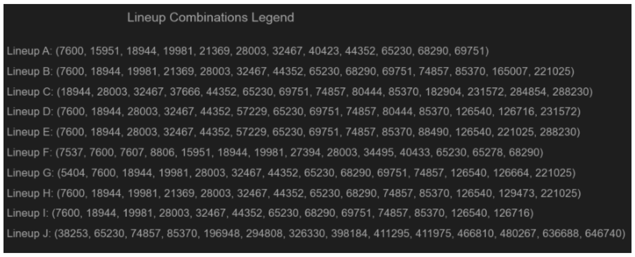
Other Factors: The spread of data points suggests that **wins** alone are not enough to guarantee financial success. The financial outcomes are also significantly influenced by factors like **team popularity**, **media deals**, and **brand strength**.

EDA 8 - Sahil

High-Attendance Games and Player Lineups

Insights:
This analysis examines the relationship between **player lineups** and **match attendance**. The findings support the hypothesis that specific **player lineups**, particularly those featuring **star players**, attract higher attendance. Certain lineups consistently draw larger crowds, likely due to factors such as **team performance**, **opponent strength**, and the presence of **high-profile players**.




Lineup Combinations Legend

Lineup A: (7600, 15951, 18944, 19981, 21369, 28003, 32467, 40423, 44352, 65230, 68290, 69751)
Lineup B: (7600, 18944, 19981, 21369, 28003, 32467, 44352, 65230, 68290, 69751, 74857, 85370, 165007, 221025)
Lineup C: (18944, 28003, 32467, 37666, 44352, 65230, 69751, 74857, 80444, 85370, 182904, 231572, 284854, 288230)
Lineup D: (7600, 18944, 28003, 32467, 44352, 57229, 65230, 69751, 74857, 80444, 85370, 126540, 126716, 231572)
Lineup E: (7600, 18944, 28003, 32467, 44352, 57229, 65230, 69751, 74857, 85370, 88490, 126540, 221025, 288230)
Lineup F: (7537, 7600, 7607, 8806, 15951, 18944, 19981, 27394, 28003, 34495, 40433, 65230, 65278, 68290)
Lineup G: (5404, 7600, 18944, 19981, 28003, 32467, 44352, 65230, 68290, 69751, 74857, 126540, 126664, 221025)
Lineup H: (7600, 18944, 19981, 21369, 28003, 32467, 44352, 65230, 68290, 74857, 85370, 126540, 129473, 221025)
Lineup I: (7600, 18944, 19981, 28003, 32467, 44352, 65230, 68290, 69751, 74857, 85370, 126540, 126716)
Lineup J: (38253, 65230, 74857, 85370, 196948, 294808, 326330, 398184, 411295, 411975, 466810, 480267, 636688, 646740)

Inference:

Player Lineups Impact: The visualization shows a clear association between certain player lineups and higher attendance, suggesting that **effective lineups** (e.g., featuring star players or fan favorites '*7600*') contribute significantly to **match-day revenue** through higher ticket sales and increased viewership.

Strategic Lineup Decisions: This reinforces the idea that clubs should optimize their lineups not just for performance but also for maximizing financial returns by considering the audience's preferences.

Machine Learning and Statistical Modeling Algorithms

Model 1 - Random Forest Classifier (Sahil)

Why Random Forest Classifier:
For the prediction problem of identifying undervalued players, a Random Forest

Classifier was chosen due to its flexibility and ability to handle both numerical and categorical features effectively. Random Forest is also robust against overfitting due to its ensemble nature, making it ideal for this task. Prior experiments with simpler models did not yield satisfactory results, prompting the choice of this more advanced model.

Model Architecture:

Preprocessing:

- Numerical features (market_value_in_eur, highest_market_value_in_eur, age, height_in_cm) were scaled using **StandardScaler**.
- Categorical features (position, sub_position) were one-hot encoded.

Classifier Details:

- The model uses **100 decision trees**.
- Random state was set to 42 to ensure reproducibility.

Training and Tuning:

The dataset was split into **80% training** and **20% testing** to evaluate the model's performance.

5-fold cross-validation was performed to validate the model and tune hyperparameters, balancing bias and variance.

Cross-validation results:

- Cross-Validation Scores: [0.9659, 0.9740, 0.9733, 0.9644, 0.8125]
- Mean Cross-Validation Score: 0.9381
- Standard Deviation of Scores: 0.062

Training-Details:
The model was trained on the cleaned and pre-processed dataset, ensuring that no missing values remained in essential features. During training:

- The Random Forest Classifier showed strong generalization across training data.

- Low variance across folds indicated model stability.

Evaluation Metrics:

Accuracy:

- Overall model accuracy was **0.98**, indicating the model performed exceptionally well in classifying undervalued players.

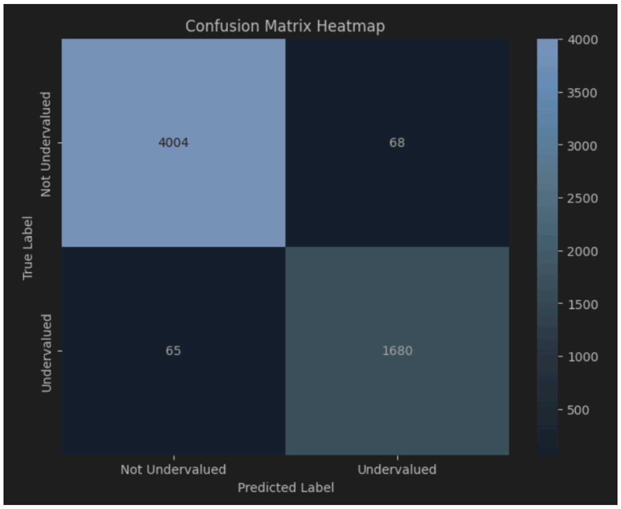
Classification Report:

- **Precision, recall, and f1-score** for both classes (0: Not Undervalued, 1: Undervalued) were high, showcasing a balanced model performance.
- Class 0 (Not Undervalued): f1-score = 0.98
- Class 1 (Undervalued): f1-score = 0.96

Confusion Matrix:

- The matrix revealed **4004 true negatives**, **1680 true positives**, and only **133 misclassifications** across the dataset.

Visualization:



Intelligence Gained:
The model effectively identified undervalued players, achieving high accuracy and balanced performance metrics. By analyzing features such as **age**, **market value**, and **position**, the model provided actionable insights:

- Players with a significant gap between their current and highest market values were flagged as undervalued.

- Younger players with high potential were identified, aligning with FC Barcelona’s strategy of building a dynamic squad.

Model 2 - XGBoost Regressor (Manas)

Why XGBoost Regressor:
For predicting football transfer fees based on player attributes and performance metrics, **XGBoost Regressor** was chosen due to its ability to handle complex, non-linear relationships between features and the target variable. XGBoost is also well-suited for tabular data and offers robust performance with relatively low risk of overfitting. Unlike simpler regression models, XGBoost can effectively model interactions between features, making it ideal for this task.

Model Architecture:

Preprocessing:

- **Feature Scaling:** Features such as age, market_value_in_eur_y, and performance were scaled using **StandardScaler**.
- **Log Transformation:** Skewed features like market_value_in_eur_y and transfer_fee were log-transformed to reduce the impact of extreme outliers.
- **Club Influence:** The average transfer fee per club was added as a feature to account for club-specific economic dynamics. Missing values were imputed with the overall average transfer fee.

Model Details:

- Objective: reg:squarederror for regression tasks.
- Hyperparameters:
  - **Number of Estimators:** 100
  - **Learning Rate:** 0.1
  - **Max Depth:** 5
  - Random State: 42 for reproducibility.

Training and Tuning:

- The dataset was split into **80% training** and **20% testing** sets to evaluate model performance.

- The model was trained on the preprocessed data, capturing the relationships between features like **goals**, **assists**, **performance**, and **transfer_fee**.

Evaluation Metrics:

Mean Absolute Error (MAE):

- Value: **0.91**
- This indicates that, on average, the model's predicted transfer fees are

approximately 0.91 units away from the actual fees.

Root Mean Squared Error (RMSE):

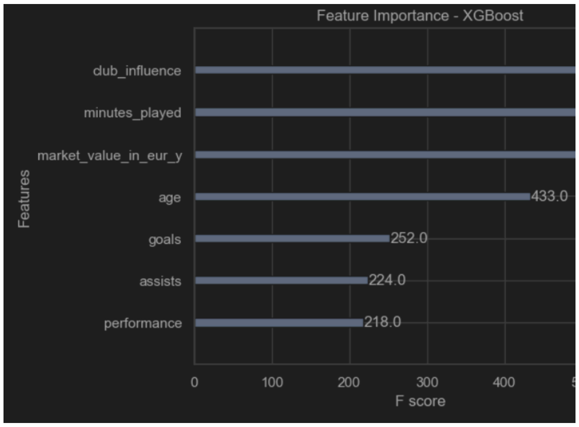
- Value: **1.16**
- RMSE provides a measure of the model's prediction error, with a lower value indicating better performance.

R-squared (R²) Score:

- Value: **0.52**
- The model explains approximately **52%** of the variance in transfer fees based on the selected features. While not perfect, this indicates a reasonable level of predictive power.

Visualization:

- **Feature Importance Plot:**
  The feature importance plot generated by XGBoost highlights the relative significance of each feature in predicting transfer fees.



Intelligence Gained:

- **Top Features:** club_influence, minutes_played, and market_value_in_eur_y are the most important features, contributing significantly to the model's predictions.
- **Actionable Insights:** The model provides valuable insights into how clubs can prioritize attributes such as goal contributions and player potential when evaluating transfer fees

Model 3 - Support Vector Regression (Manas)

Why Support Vector Regression (SVR):
SVR was chosen for modeling profitability in football competitions due to its ability to handle high-dimensional feature spaces and capture non-linear relationships through kernel functions. The **polynomial kernel** was used to model the complex interactions between features such as **goal difference**, **stadium seats**, and **match outcomes**. Unlike linear regression, SVR can generalize better for this type of data where relationships may not be strictly linear.

Model Architecture:

Preprocessing:

- Features such as stadium_seats, hosting, goal_difference, revenue, and profitability were scaled using **StandardScaler** to ensure uniform feature distributions for effective SVR training.
- The target variable, **profitability**, was calculated as a combination of **match revenue** and **win bonuses**, reflecting the financial impact of match outcomes.

Model Details:

- Kernel: **Polynomial**
- The polynomial kernel captures complex relationships between features, making SVR suitable for profitability prediction.

Training and Tuning:

- The dataset was split into **80% training** and **20% testing** to evaluate model performance.
- The SVR model was trained on the scaled training data.

Evaluation Metrics:

Mean Absolute Error (MAE):

- Value: **7557.50**
- This indicates that the model's predicted profitability is, on average, approximately 7557 units away from the actual profitability.

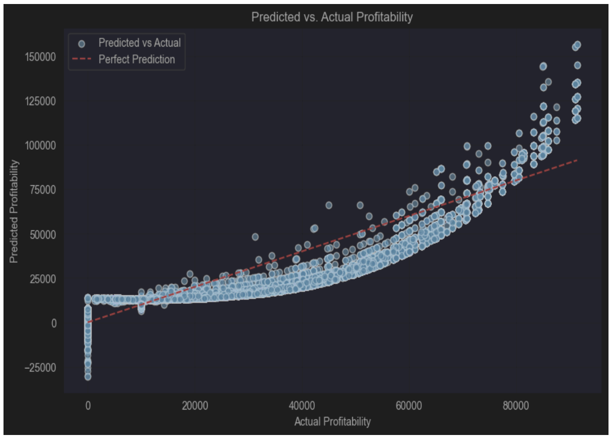
Root Mean Squared Error (RMSE):

- Value: **9941.00**
- This measure of error emphasizes larger discrepancies, with a lower value indicating better performance.

R-squared (R^2) Score:

- Value: **0.79**
- The model explains approximately **79%** of the variability in profitability, showcasing a strong predictive capability.

Visualization:



Intelligence Gained:

Significant Features:

- **Goal Difference:** Matches with higher goal differences contributed significantly to profitability, indicating that dominant performances attract more attention and revenue.
- **Hosting Advantage:** The encoding of hosting highlights the financial impact of home vs. away matches.

Profitability Drivers: The model identifies key drivers of profitability, such as match attendance (stadium_seats) and win bonuses, providing actionable insights for optimizing financial strategies.

Conclusion:
The SVR model achieved a robust R^2 score of **0.79**, demonstrating its effectiveness in predicting profitability based on match and club-level features. This model provides actionable insights into the financial dynamics of football matches, enabling clubs to optimize match-day strategies and resource allocation. Future work could involve exploring additional features such as **ticket pricing strategies**, **merchandising revenue**, and **sponsorship impacts** to further enhance the model's accuracy.

Model 4 - Gradient Boosting Classifier (Sahil)

Why Gradient Boosting Classifier:
Gradient Boosting was chosen for its ability to handle imbalanced datasets effectively and its performance in binary classification problems. The ensemble learning approach combines weak learners (decision trees) to create a strong model, making it ideal for predicting the likelihood of a player scoring a goal during a match.

Model Architecture:

Preprocessing:

- Missing values were dropped to ensure the data's integrity.
- **SMOTE (Synthetic Minority Oversampling Technique):** Used to handle class imbalance by oversampling the minority class.
- **Undersampling:** Applied to reduce the majority class samples, ensuring balanced class distributions for training.
- Class weights were calculated and integrated into the model to further address imbalance.

Model Details:

- Base Model: Gradient Boosting Classifier.
- Hyperparameters:
  - Learning rate and number of estimators were set to default (fine-tuning possible for optimization).

Feature Selection and Importance:
The following features were selected for their potential to impact the target variable (goal_scored):

- **Minutes Played:** Most significant contributor with an importance score of **30.29%**.
- **Home and Away Club Goals:** Combined importance of **54.72%**, indicating that team performance significantly influences a player's goal-scoring likelihood.
- **Assists:** Played a moderate role with an importance of **10.55%**, reflecting the value of team contributions.
- **Other Factors:** Features like attendance, yellow_cards, and red_cards had minimal impact, suggesting they are less relevant in this context.

- **Training and Tuning:**
  - The dataset was split into **80% training** and **20% testing** sets.
  - Class distributions were addressed using oversampling, undersampling, and class weight adjustments to ensure balanced training data.
  - The Gradient Boosting Classifier was trained on the preprocessed and balanced dataset.

Evaluation Metrics:

Model Accuracy:

- Value: **92%** (imbalanced data).
- Value: **62%** (balanced data).

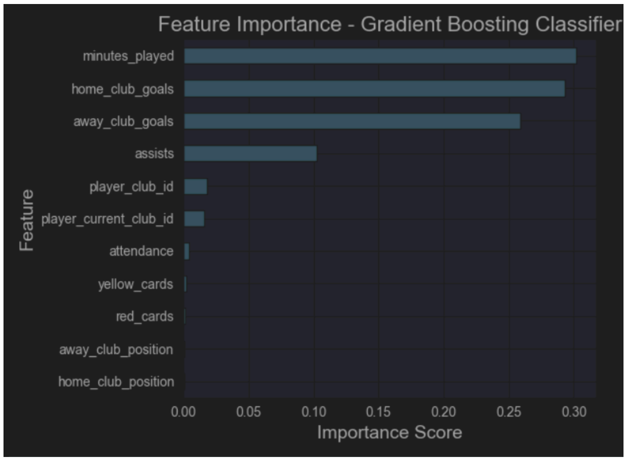
Indicates strong predictive performance, though balancing the classes reduces accuracy slightly due to increased sensitivity to the minority class.

ROC-AUC Score:

- Value: **0.73**.

Highlights the model's capability to distinguish between classes effectively, even with imbalanced data.

Visualization:



Intelligence Gained:

The analysis highlights that **minutes played**, **home club goals**, and **away club goals** are the most significant features influencing goal-scoring potential. Additionally, **assists** moderately affect the scoring probability, while other features like attendance, yellow cards, red cards, and club positions have negligible impact.

Conclusion:

Player performance and goal-scoring likelihood are primarily driven by time on the field and team performance metrics, emphasizing the importance of strategic playtime allocation and enhancing team dynamics to create scoring opportunities.