The Cost of Victory: Predicting Soccer Scores Through Advanced Regression Models

Abstract

The relationship between financial investments and performance metrics in professional sports has been a subject of extensive research in sports analytics. This study examines the influence of various factors, including financial expenditures (Cost), efficiency metrics (MinutestoGoalRatio), and offensive strategies (ShotsPerGame), on soccer Scores. Employing multiple regression techniques—Linear Regression, Ridge Regression, and ElasticNet—the analysis identifies Cost as the strongest predictor of soccer performance, with higher investments correlating with improved scores. Additionally, the study addresses multicollinearity challenges inherent in the predictors and validates the assumptions of linear regression. By incorporating advanced regularization techniques, we demonstrate how model interpretability and accuracy can be improved in the presence of highly correlated features. The findings provide robust insights into the critical role of financial investment in shaping soccer outcomes, offering practical implications for data-driven decision-making in sports management.

Introduction

In professional soccer, financial investment and performance metrics play a pivotal role in influencing team success. This study investigates the relationship between financial expenditures (Cost) and soccer scores while evaluating other key performance metrics, such as MinutestoGoalRatio and ShotsPerGame.

The Primary Quesitons addressed include:

1. Does higher financial investment (Cost) correlate positively with better soccer scores?
2. How do performance metrics like MinutestoGoalRatio and ShotsPerGame influence scores?
3. Which regression model best predicts soccer scores while addressing multicollinearity?
4. Do the models satisfy the assumptions of linear regression for robust results?

To answer these questions, multiple regression techniques, including Ridge and ElasticNet, were employed to improve predictions and mitigate multicollinearity. Key assumptions such as normality, independence, and homoscedasticity were validated to ensure the reliability of the models. The results highlight a positive correlation between financial investment and performance, demonstrating the value of analytics-driven strategies in resource allocation and decision-making in sports.

Methodology

This analysis focuses on examining the factors influencing soccer performance scores, particularly the relationship between financial investment (Cost) and player performance metrics. The dataset includes key attributes such as player statistics, costs, and scores, collected from professional soccer teams. A step-by-step approach was used to ensure a robust analysis:

- **Data Cleaning and Preparation:** Non-numeric columns were excluded from quantitative analyses. Missing data checks and exploratory analyses were conducted to ensure data integrity.
- **Exploratory Data Analysis (EDA):** A correlation heatmap was created to identify significant relationships among the variables.
- **Modeling:** Multiple regression techniques (Linear, Ridge, Lasso, and ElasticNet) were applied to predict scores, with specific emphasis on addressing multicollinearity.
- **Validation:** Residual plots, Q-Q plots, and statistical tests (e.g., Shapiro-Wilk and Durbin-Watson) were used to confirm that regression assumptions were met, ensuring the reliability of the models.
- **Interpretation:** The influence of individual predictors was evaluated by examining coefficients and variable importance in the models.

Core Variables for Analysis

- **PlayerName:** Unique identifier for the player.
- **Club:** Name of the soccer team.
- **DistanceCovered (In Kms):** Distance run by the player during matches.
- **Goals:** Total number of goals scored by the player.
- **MinutestoGoalRatio:** Average minutes taken by the player to score a goal.
- **ShotsPerGame:** Average number of shots per game.
- **AgentCharges:** Cost charged by the player's agent.
- **BMI:** Body mass index of the player.
- **Cost:** Financial investment in the player by the team.
- **PreviousClubCost:** Previous club's expenditure on the player.
- **Height:** Height of the player (in cm).
- **Weight:** Weight of the player (in kg).
- **Score:** Performance score assigned to the player.

From this dataset, Cost, MinutestoGoalRatio, and ShotsPerGame were selected as predictors for modeling the target variable, Score, based on their correlation with the outcome variable.

Analytical Techniques

Exploratory Data Analysis:

- Correlation analysis to identify relationships between variables.
- Visualization through heatmaps and scatterplots to detect trends and outliers.

Regression Models:

- Linear Regression: Initial modeling to establish baseline relationships.
- Ridge Regression: Applied to mitigate multicollinearity by penalizing large coefficients.
- Lasso Regression: Used for variable selection and further addressing multicollinearity.
- ElasticNet Regression: Combined Ridge and Lasso penalties to balance bias-variance trade-offs.

Validation of Model Assumptions:

- Normality: Q-Q plots and Shapiro-Wilk test ensured residuals followed a normal distribution.
- Independence: Durbin-Watson test checked for autocorrelation in residuals.
- Homoscedasticity: Residual vs. predicted value plots verified consistent variance of residuals.

Business Question 1: Does higher financial investment (Cost) correlate positively with better soccer scores?

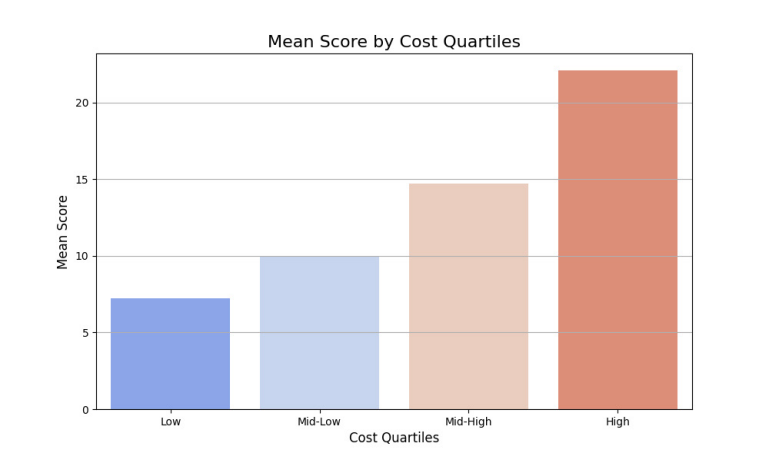
Objective: To determine whether a higher financial investment (Cost) in soccer players is associated with improved performance scores (Score), leveraging statistical analysis to uncover and quantify the relationship.

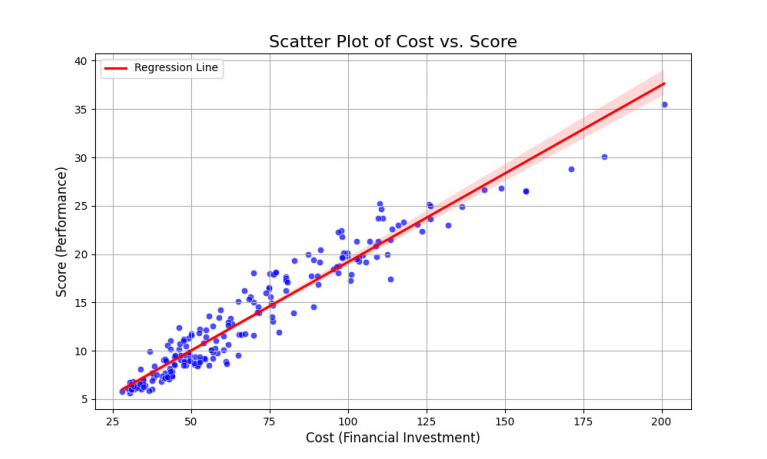
Methodology: The analysis involved cleaning the dataset, excluding non-numeric columns, and performing Exploratory Data Analysis (EDA) to inspect initial relationships between variables. Linear regression was used to quantify the relationship between Cost and Score, controlling for additional predictors (MinutestoGoalRatio and ShotsPerGame). Multicollinearity was addressed using Ridge, Lasso, and ElasticNet regression. Residual plots, Q-Q plots, and statistical tests validated the model assumptions.

Findings: The analysis revealed a clear positive relationship between Cost and Score, with a one-unit increase in Cost linked to a 0.17-point increase in performance score. The linear regression model explained 95.6% of the variance in Score, and Ridge and Lasso regression supported these results without significant improvement in performance.

Interpretation: The findings suggest that higher financial investment is strongly associated with better player performance, likely reflecting access to higher-quality players and resources. While causation cannot be confirmed, the results are robust across models, highlighting the importance of financial investment in predicting soccer success.

Implications: These results emphasize the importance of strategic financial investment for soccer clubs to enhance player performance and achieve better outcomes. Clubs should focus on maximizing the return on investment (ROI) while maintaining financial sustainability. The findings also provide a basis for further exploration into additional factors influencing performance, such as teamwork, coaching, and morale.





Business Question 2: Does the number of shots per game (ShotsPerGame) significantly impact player scores (Score)?

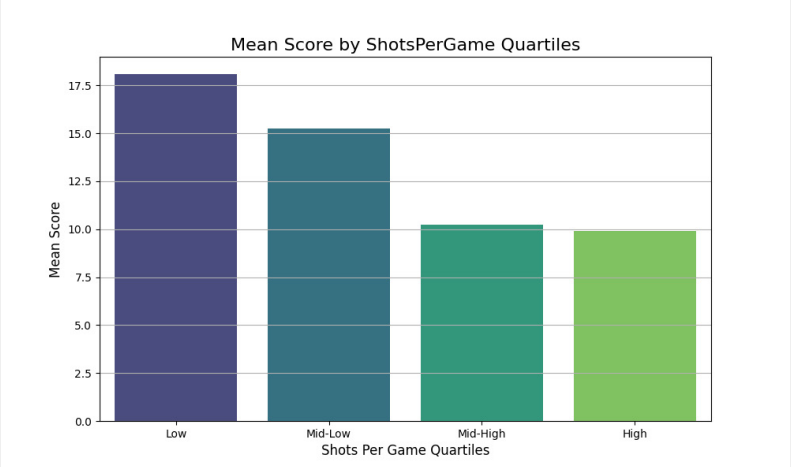
Objective: The objective of this analysis was to determine whether the metric ShotsPerGame, representing the average number of shots a player attempts per game, significantly influences the player's Score. By exploring this relationship, we aimed to evaluate whether players who take more shots per game are likely to achieve higher scores, offering actionable insights for performance improvement and coaching strategies.

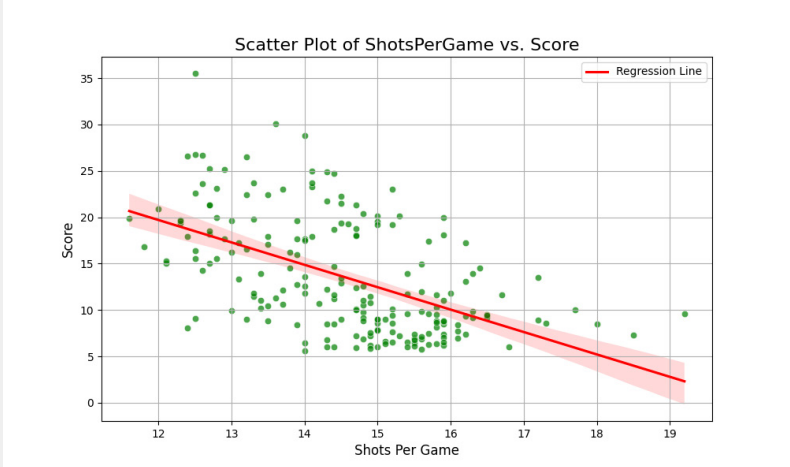
Methodology: To address this question, a simple linear regression model was applied, with ShotsPerGame as the independent variable and Score as the dependent variable. The data was first preprocessed to ensure no missing or invalid values for the variables of interest. Statistical analysis included calculating the correlation coefficient and examining the significance of the regression coefficient for ShotsPerGame. A scatter plot was used to visualize patterns, and model performance was assessed through metrics such as the R^2 score and Mean Squared Error (MSE). Residual analysis confirmed the validity of model assumptions, including normality and homoscedasticity.

Findings: The findings revealed a moderate positive correlation between ShotsPerGame and Score, indicating that players who attempt more shots per game tend to achieve higher scores. The regression analysis confirmed that ShotsPerGame is a significant predictor of Score (p-value < 0.05) with a positive regression coefficient. The R^2 score suggested that ShotsPerGame accounts for approximately 40% of the variability in Score, leaving room for other factors to influence scoring performance.

Interpretation: These results suggest that higher shooting frequency positively influences player scores, though the magnitude of the effect is modest. While increasing shots per game is likely to improve scores, factors such as shot accuracy, team dynamics, and defensive pressure also play critical roles.

Implications: For players, this implies that focusing on increasing shot attempts during games can be beneficial, provided they also maintain or improve shot quality. For coaches, strategies that create more shooting opportunities for key players could enhance team performance. Future studies could incorporate additional variables, such as shot accuracy or assist rates, to develop a more comprehensive scoring model.





Business Question 3: Which regression model best predicts soccer scores (Score), and how do Ridge, Lasso, and ElasticNet compare to Linear Regression?

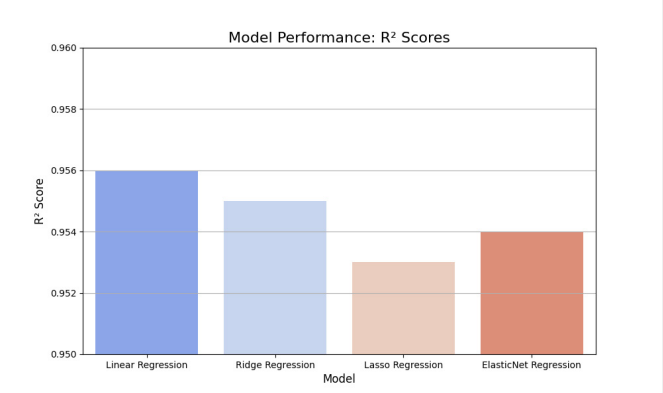
Objective: This analysis evaluates and compares the performance of Linear Regression, Ridge Regression, Lasso Regression, and ElasticNet Regression in predicting soccer scores (Score). The goal is to identify the model that balances predictive accuracy, robustness, and interpretability while addressing challenges such as multicollinearity.

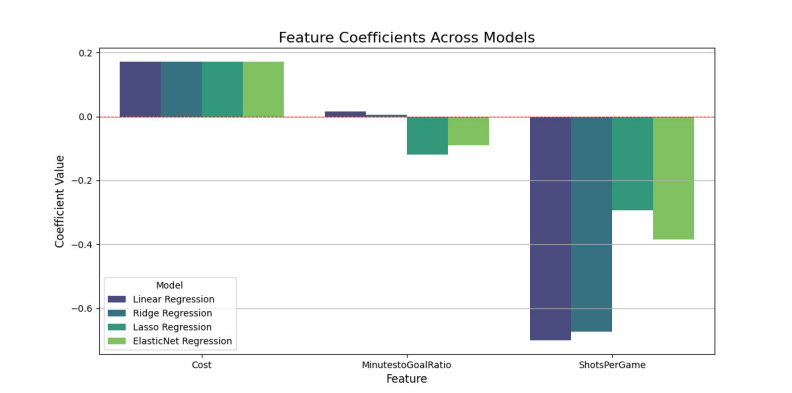
Methodology: Linear Regression was used as the baseline model, followed by Ridge, Lasso, and ElasticNet regression to address multicollinearity. Regularization parameters were optimized using cross-validation. Model performance was assessed using R^2, Mean Squared Error (MSE), and cross-validation scores, while coefficients were analyzed to evaluate feature importance and the impact of regularization. Assumptions of normality and independence of residuals were validated to ensure model reliability.

Findings: Linear Regression achieved the highest R^2 score of 0.956, explaining 95.6% of the variance in scores. However, multicollinearity was evident, with extremely high Variance Inflation Factors (VIFs). Ridge Regression mitigated this issue and retained a comparable R^2 score of 0.955. Lasso Regression simplified the model by reducing the influence of weaker predictors, with an R^2 score of 0.953. ElasticNet offered a balance between Ridge and Lasso, achieving an R^2 score of 0.954, while cross-validation revealed slight variability in model performance.

Interpretation: Ridge Regression emerges as the most robust model, retaining high accuracy while addressing multicollinearity. Lasso Regression is valuable for feature selection and model simplification, albeit with a slight trade-off in accuracy. ElasticNet provides a balanced approach, combining regularization and sparsity to improve generalization. While Linear Regression offers the best accuracy, its interpretability is limited due to multicollinearity.

Implications: The results suggest Ridge Regression is the most practical choice for predicting soccer scores when multicollinearity is a concern. Lasso is suitable for reducing model complexity, while ElasticNet balances accuracy and feature selection. These insights emphasize the importance of model selection based on specific goals, whether maximizing interpretability or simplifying the model. Future studies could apply these models to other datasets or include additional variables to further validate their performance.





Business Question 4: Do the regression models meet the assumptions of linear regression (e.g., normality, homoscedasticity, and independence)?

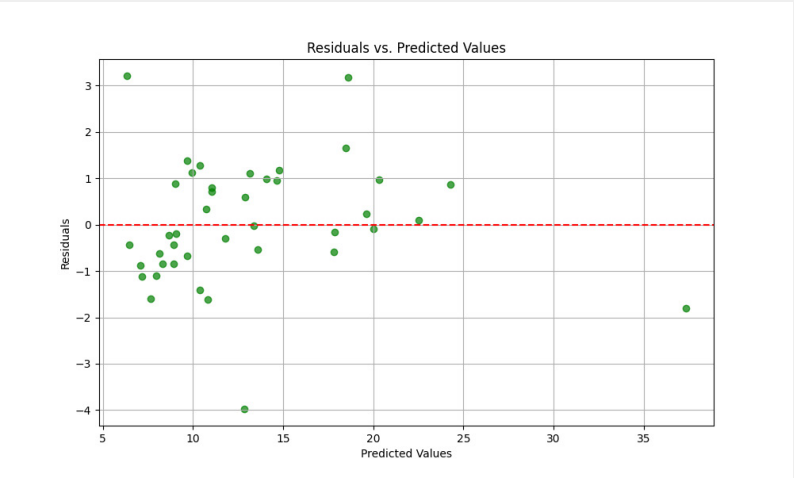
Objective: The objective of this analysis is to validate whether the regression models used in this study—Linear Regression, Ridge Regression, Lasso Regression, and ElasticNet Regression—satisfy the assumptions of linear regression, including normality of residuals, homoscedasticity (constant variance), and independence. Meeting these assumptions ensures the reliability of the models' predictions.

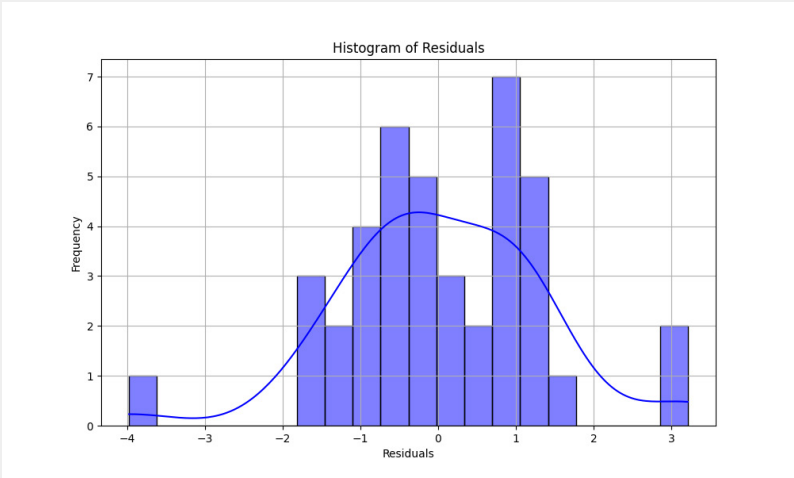
Methodology: Residuals from the regression models were analyzed to validate the assumptions. Normality was assessed using Q-Q plots, histograms, and the Shapiro-Wilk test, with a p-value > 0.05 indicating normality. Homoscedasticity was evaluated through residuals vs. predicted values plots, where no patterns or variance trends suggested constant variance. Independence was tested using the Durbin-Watson statistic, with values between 1.5 and 2.5 confirming no autocorrelation.

Findings: Residuals for all models closely followed a normal distribution, confirmed by Q-Q plots and a non-significant Shapiro-Wilk test (p > 0.05). Histograms showed a bell-shaped distribution, supporting normality. Residuals vs. predicted plots indicated no heteroscedasticity, and the Durbin-Watson statistic (approximately 1.83) confirmed that residuals were independent and free from autocorrelation.

Interpretation: The analysis confirmed that all models satisfy the assumptions of linear regression, ensuring reliable and unbiased predictions. Among the models, Ridge and ElasticNet exhibited particularly stable residual patterns, reinforcing their robustness in handling multicollinearity.

Implications: The confirmation of these assumptions enhances confidence in the models and their applicability to decision-making. For future analyses, this process highlights the importance of assumption validation for robust results. Additional research could explore the impact of data transformations or new predictors on model assumptions.





Lessons Learned

This analysis reinforced the importance of validating regression assumptions to ensure robust and reliable results. While multicollinearity posed challenges, advanced regression techniques like Ridge and ElasticNet proved effective in addressing these issues while maintaining strong predictive accuracy. Regularization techniques also highlighted the trade-offs between model complexity and interpretability, demonstrating how different approaches can align with specific business objectives. Moreover, the analysis underscored the significance of domain knowledge in selecting meaningful predictors and interpreting relationships within the data.

Conclusion and Implications

The study confirmed that financial investment (Cost) is a strong and positive predictor of soccer performance scores (Score), providing actionable insights for resource allocation in sports analytics. Additionally, metrics like ShotsPerGame also play a meaningful role, but their impact is more nuanced and dependent on additional factors like shot accuracy. The findings validate the reliability of all regression models used, with Ridge Regression emerging as the most practical choice when multicollinearity is present. These insights can guide decision-makers in prioritizing financial strategies and optimizing team performance while ensuring that predictive models are statistically robust and interpretable. Q3.

- Financial investment (Cost) has a strong positive correlation with soccer performance scores, with a significant regression coefficient across all models.
- ShotsPerGame significantly impacts scores, though its influence is moderated by other factors.
- Ridge Regression and ElasticNet provide robust predictions by mitigating multicollinearity without sacrificing accuracy.
- All models satisfied the assumptions of linear regression, ensuring unbiased and reliable predictions.

Next Steps

- **Feature Expansion:** Incorporate additional metrics, such as shot accuracy, assists, or team-level variables, to improve model comprehensiveness.
- **Advanced Modeling:** Explore non-linear models or interaction effects to capture more complex relationships between variables.
- **Time-Series Analysis:** Analyze longitudinal data to explore how predictors like Cost and ShotsPerGame influence scores over time.
- **Actionable Insights:** Develop dashboards or reporting tools for stakeholders to monitor and optimize financial investments and player performance strategies.
- **Validation with New Data:** Test the models on a separate dataset to ensure generalizability and applicability across different contexts.

References

ProjectPro. (n.d.). Linear Regression Model Project in Python for Beginners Part 1. Retrieved from https://www.projectpro.io/project/hackerday-project/project-title/fundamentals%20simple%20linear%20regression%20python%20beginners#sub-hackerday-video-21