Final Report: Transfer Expenditure and Football Success

# Introduction

Football is not only a game of passion and skill, but also a complex financial enterprise. In this project, I explored the link between financial decisions and team performance in the English Premier League. Specifically, I aimed to uncover whether spending big money on player transfers truly correlates with short-term success, measured in league points and standings.  
By combining my interests in football and data science, I developed a structured workflow to investigate this relationship. My analysis leveraged 10 years of data, covering transfer expenditures, earnings, squad changes, and seasonal performances. Later I also addes football club’s wage data as well to enrich my data. From this part one can also see the relation between spendings and success relation hopefully.

This report presents my step-by-step journey, starting from data collection and cleaning, through exploratory data analysis (EDA) and hypothesis testing, to the final machine learning (ML) modeling efforts that aim to predict next season’s team points.

# Data Collection and Cleaning

I collected the primary dataset from Transfermarkt, which provided comprehensive details about each team's transfer expenditures, income from player sales, and market values. Additionally, I gathered performance data (points, standings, trophies) from the Premier League and UEFA websites.

The raw dataset included financial metrics (Expenditure, Income, Net Balance), performance metrics (Wins, Draws, Goals Scored, Goals Conceded, League Points), and squad composition details (Number of Arrivals, Departures, Market Value).  
To ensure consistency and reduce multicollinearity, I removed redundant features such as Balance and '+/-', as they were directly calculated from other columns. Later with the enrichment the wage data for players were also available to use in order to gain more insight about these football clubs financial and success rate measures.

# EDA and Hypothesis Testing

After cleaning the dataset, I performed extensive exploratory data analysis (EDA) to identify relationships and potential patterns. This included generating scatter plots, correlation matrices, and boxplots to visualize connections between variables. Here I gained a relatively huge insight about my data and what my data really means came clear for reader hopefully. Therefore I made some key observations which led to these ideas:  
Key observations included a moderate positive correlation between expenditure and league points, no strong impact from net profit/loss, and mixed evidence about the number of squad arrivals on performance. And this resulted in having some hypothesis.

Based on these insights, I formulated several hypotheses to test potential causal relationships. Here are the hypotheses:

* **High Transfer Expenditure Leads to Better League Performance**
* **Net Transfer Profit (Balance) Is Inversely Related to Season Points**
* **Higher transfer expenditure in one season is associated with a positive change in league points in the following season.**
* **Clubs with higher transfer income are able to maintain or even improve their league points in the same season.**
* **There is a Relationship between the total yearly wage spending of Premier League teams and the points they earned.**

These hypotheses were tested using correlation analysis and basic significance tests to determine if the observed relationships held true statistically.

Hypothesizes were also pointed out with graphs and data analyses methods as an example:

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Yapay zeka tarafından oluşturulan içerik yanlış olabilir.

# Using EDA Insights to Develop a Machine Learning Idea

The insights from EDA and hypothesis testing guided me to develop a machine learning idea: predicting a football team’s next season points based on its current financial and performance data. This question was personally engaging, blending data science with a real-world topic I enjoy.

The final dataset, refined through EDA, served as the input for this modeling effort.

# Machine Learning Models Overview and Conclusion

I tested multiple models from lecture codes to identify the best predictive approach:  
- Linear Regression: Provided a strong baseline, explaining 76% of the variance.  
- K-Nearest Neighbors (KNN, k=15): Captured local patterns and performed reasonably well.  
- Decision Tree Regressor: Overfit the data and performed worse than simpler models.  
- Random Forest Regressor: Balanced bias and variance, similar to Linear Regression.  
- XGBoost Regressor: Did not outperform simpler models, possibly due to dataset size and overfitting.  
A comparison of R² scores is presented in the table below:

|  |  |
| --- | --- |
| Model | R² Score |
| Linear Regression | 0.7658 |
| KNN (k=15) | 0.7398 |
| Decision Tree | 0.4049 |
| Random Forest | 0.7233 |
| XGBoost | 0.4740 |

Overall, simpler models with good feature selection and ensembling (like Random Forest) provided the most balanced and reliable predictions. This project underscored how careful data cleaning and EDA play a crucial role in building accurate models, and how even complex sports data can be understood through statistical and ML tools.