Data Science

Project Proposal – NBA Game outcome Prediction

Introduction

In the high-stakes world of professional sports, data analytics has emerged as a game-changer, transforming how decisions are made on and off the court. This project harnesses the predictive prowess of machine learning to forecast NBA game outcomes, offering insights that could revolutionize team strategies and betting markets alike.

Research Questions

1. Can historical NBA game statistics be utilized to predict the outcome of a game?
2. What statistical indicators are most influential in determining the winning team?
3. Can the prediction model adapt dynamically to in-season changes such as player performance trends?
4. How does the home-court advantage factor into the predictive model?

Data to be used

Selenium and Python will be employed for the purpose of scraping NBA game statistics from the website "Basketball Reference" so that required data is captured for its analysis. Approach: Data Preprocessing:

This will describe the detail steps of the data cleaning and preparation needed to make the data readied up for analysis.

Data Preprocessing:

Initial steps will include the removal of outliers, handling missing values, and normalizing data for uniformity. Feature engineering will be a key focus, where we'll derive new variables that could significantly impact the model's performance, such as a player's recent performance trend or team fatigue levels.

Predictive Modeling: This will be done focusing on using a number of predictive models.

Logistic regression: To act as the baseline accuracy of prediction.

SVM: Serves as a robust alternative capable of capturing nonlinear patterns through the use of kernel functions, making it highly versatile for predicting outcomes based on intricate statistical data.

XG Boost: Remains as a top choice for its performance in handling tabular data, its efficiency, and its ability to deal with various types of features and complex relationships within the data.

Model evaluation—discuss the metrics accuracy, precision, recall, and the A AUC-ROC curve with respect to comprehensive model evaluation. Implementation and Future Work—Model practical applications and the way forward in advancing the study. Formatting and Submission

Proposals will be written in the form of a quality professional Jupyter Notebook, with clear and informative diagrams, section headers, and absolutely impeccable proofreading.

References:

Data collection –

https://www.basketball-reference.com/leagues/NBA_2023_games-november.html

https://www.basketball-reference.com/boxscores/202210180BOS.html

Inspired from - https://www.youtube.com/watch?v=MpLHMKTolVw

Questions:

- **Have you been able to download the data and test for size?**
  - We are working on the web scraping part, as per our research each NBA season consists of 2460 matches, if we collect 5 seasons that will be 12,300 matches (rows).
- · **What are the list of features you will use?**
  - We are focusing on 3 different categorical features
    - Line scores
    - Basic Stats
    - Advance Stats
- · **Is there a time period you will be focusing on?**
  - We are planning to consider the last 5 seasons.
- · **How will you deal with player specific effects?**
  - We are not able to see that data so, we thought of not considering this data.
- · **Sponsorship and other factors that make a team stronger than others?**
  - Definitely but unfortunately we can’t access that data too.
- · **Are there any budget related datasets about teams?**
  - No, there is no budget related data.

- · **How are you distributing work across the team?**

RACI

Task	Jetendra Mulinti	Goutham Vemula	Prajeeth Nakka
Data Collection and Preprocessing	A (Accountable)	R (Responsible)	C (Consulted)
Feature Engineering and Selection	A (Accountable)	C (Consulted)	R (Responsible)
Model Development	R (Responsible)	R (Responsible)	C (Consulted)
Model Evaluation and Selection	C (Consulted)	C (Consulted)	R (Responsible)
Documentation and Presentation	A (Accountable)	R (Responsible)	R (Responsible)

Additional Questions (Based on Feedback):

1. **What does the data look like? You should have outputted the data header in the document.**

   Sample_data.csv

2. **How will you deal with player specific effects? Sponsorship and other factors that make a team stronger than others?**
   - We are considering the best player per team score (max_score) also, along with total score.
3. **Are there any budget related datasets about teams?**
   - No we are considering budget related data.
4. **Each student MUST BUILD ATLEAST ONE MODEL.**
   - **Logistic Regression:** Prajeeth
   - **SVM:** Jetenra
   - **XG Boost:** Goutham

Team –

Jetendra Mulinti

Goutham Vemula

Prajeeth Nakka