Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Applied Probability I (STU22004) : Report

By

Rachel Ranjith, (23363463) and Luisa Amkhadova, (23374505)

Trinity Business School

TRINITY COLLEGE

UNIVERSITY OF DUBLIN

December 2024

Question 1: Simulating Temperature Behavior

a) Problem Statement

Consider a scenario where the temperature $X(t)$ varies randomly over a continuous time interval t, where t is in the range from 0 to 1. We begin with the assumption that $X(0) = 0$, which means that the temperature at time 0 is 0. Now, if we choose a small time increment represented by Δt, we can make the assumption that the change in temperature from time t to $t + \Delta t$, denoted as $X(t + \Delta t) - X(t)$, follows a normal distribution. This normal distribution is characterized by a mean of 0 and a variance of Δt.

I. Let P be the random variable denoting the proportion of time in $[0, 1]$ such that the temperature is positive. Estimate the distribution of P by Monte Carlo simulation and experiment with various values of Δt (e.g. $\Delta t = 0.01, 0.001, 0.0001$ …)

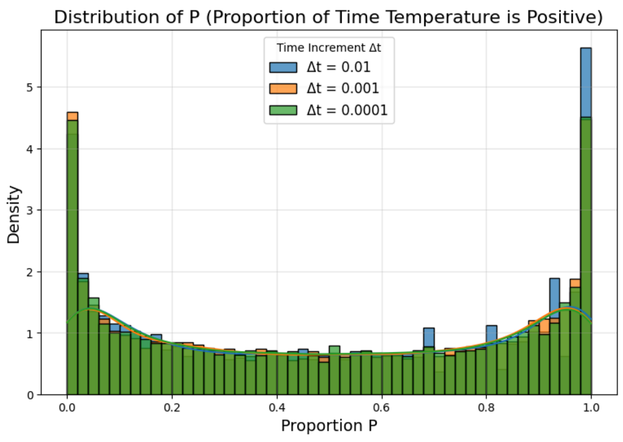
II. Let T_{max} be the random variable denoting the time in $[0, 1]$ such that the temperature is at its maximum. Estimate the distribution of T_{max} by Monte Carlo simulation and experimenting with various values of Δt (e.g. $\Delta t = 0.01, 0.001, 0.0001$…)

b) Part I

i. Simulation Methodology

1. Divide the interval $[0, 1]$ into $N = \frac{1}{\Delta t}$ small intervals, where Δt is the step size. The times are: $t_0 = 0,\ t_1 = \Delta t,\ t_2 = 2\Delta t,\ \ldots,\ t_N = 1.$
2. Starting with $X(0) = 0$, for $i = 1, 2, \ldots, N$, to simulate $X(t_i)$ using: $X(t_i) = X(t_{i-1}) + \sqrt{\Delta t} \bullet (Z_i)$
3. Count the number of intervals where $X(t) > 0$ and estimate P as: $P = \frac{Number\ of\ positive\ intervals}{N}$
4. Run the simulation for any amount of results and collect the values of P to estimate its distribution. Repeat the process for different values of Δt (e.g. $\Delta t = 0.01, 0.001\ldots$)

ii. Simulation



iii. Results

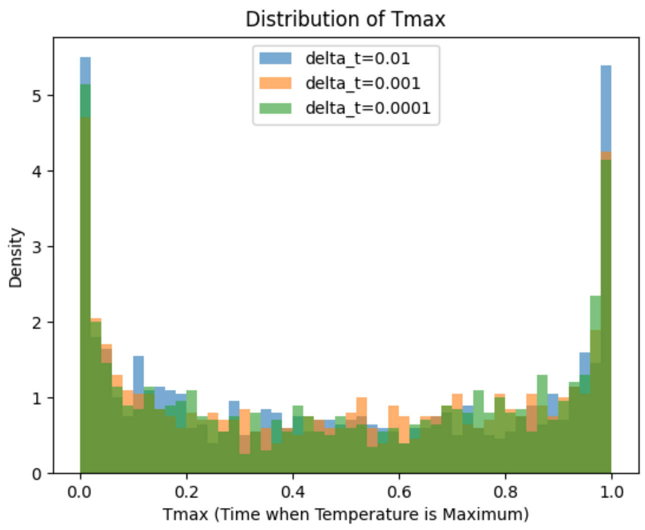
- **Graph Observations:**
  - The distributions with peaks around $P = 0$ and $P = 1$ suggest that in many simulations, the temperature remains mostly positive or negative.
  - For intermediate values of P, the density is relatively low, suggesting less frequent occurrences of a balanced positive/negative temperature.
- **Impact of** Δt:
  - Larger Δt (e.g., $\Delta t = 0.01$) results in a smoother but less precise estimation of P, as seen in the broader density curves.
  - Smaller Δt (e.g., $\Delta t = 0.0001$) improves the precision, resulting in sharper peaks at extreme values of PPP, though at a computational cost.

c) Part II

i. Simulation Methodology

1. Divide the interval $[0, 1]$ into $N = \frac{1}{\Delta t}$ steps, where Δt is the time increment.
2. Initialize $X(0) = 0$, then for each time step, update the temperature using $X(t_i) = X(t_{i-1}) + \sqrt{\Delta t} \bullet Z_i$, where $Z_i \sim N(0, 1)$ is a standard normal random variable.
3. For each simulation, identify T_{max}, the time at which $X(t)$ reaches its maximum value.
4. Perform the simulation for multiple trials (e.g., 10,000) to collect values of T_{max}.
5. Plot the distribution of T_{max} for different values of Δt (e.g. $\Delta t = 0.01, 0.001, 0.0001$) to observe the impact on precision and accuracy.

ii. Simulation



iii. Results

- **Distribution of T_{max}**:
    - The distribution of T_{max} is symmetric around $t = 0.5$, which is expected since maximum temperatures are most likely to occur near the midpoint.
- **Effect of Δt on the Distribution**:
    - As Δt decreases (e.g., from 0.01 to 0.0001), the distribution of T_{max} becomes smoother and more concentrated around the center ($t = 0.5$).
    - Larger Δt values result in wider distributions with more variability in the time of maximum temperature.

Question 2: Premier League Forecasting Model

a) Problem Definition

Create a probabilistic model and perform Monte Carlo simulations to forecast the final points for Premier League teams in the 2024-2025 season. This model will include some unknown parameters that you will determine based on the data you gather.

b) Model Details

Parameters

1. **Offensive Strength**: $O_i = \frac{\Sigma W_y \times Goals\ Scored\ in\ Year\ y}{\Sigma W_y \times Matches\ Played\ in\ Year\ y}$, where W_y is the weight for year y
2. **Defensive Strength**: $D_j = 5 - \frac{\Sigma W_y \times Goals\ Conceded\ in\ Year\ y}{\Sigma W_y \times Matches\ Played\ in\ Year\ y}$, where W_y is the weight for year y
3. **Recent Form**: Weighted average of performance in the last n matches, calculated as: $F_i = \frac{\sum_{k=1}^{n} W_k \cdot R_k}{\sum_{k=1}^{n} W_k}$, where W_k are the weights (recent matches weighted higher) and R_k is the result for match k..
4. **Goal Conversion Rate**: Captures how efficient a team is at converting shots on target into goals: $C_i = \frac{Goals\ Scored\ by\ Team\ i}{Shots\ on\ Target\ by\ Team\ i}$

Proposed Model

For each match between teams i and j:

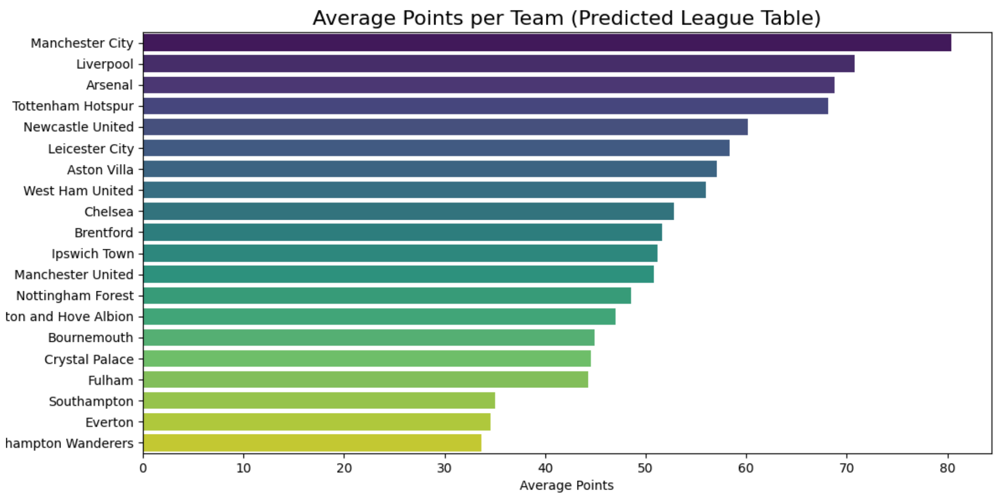
1. **Incorporate Recent Form**:
   - Adjust offensive strengths using recent form: $O_i^{'} = O_i \times (1 + w \cdot F_i)$
   - Adjust defensive strengths using recent form: $D_j^{'} = D_j \times (1 - w \cdot F_i)$
2. **Expected Goals Scored**:
   - Team i: $\lambda_{i,j} = O_i^{'} \times D_j^{'}$
   - Team j: $\lambda_{j,i} = O_j^{'} \times D_i^{'}$
3. **Incorporate Goal Conversion Rate**:
   - Refine λ using C_i: $\lambda_{i,j} = \lambda_{i,j} \times C_i$
4. **Simulate Match Results**:
   - Use a **Poisson distribution** to simulate goals: $G_i \sim Poisson(\lambda_{i,j})$, $Gj \sim Poisson(\lambda_{j,i})$

c) Simulation Methodology

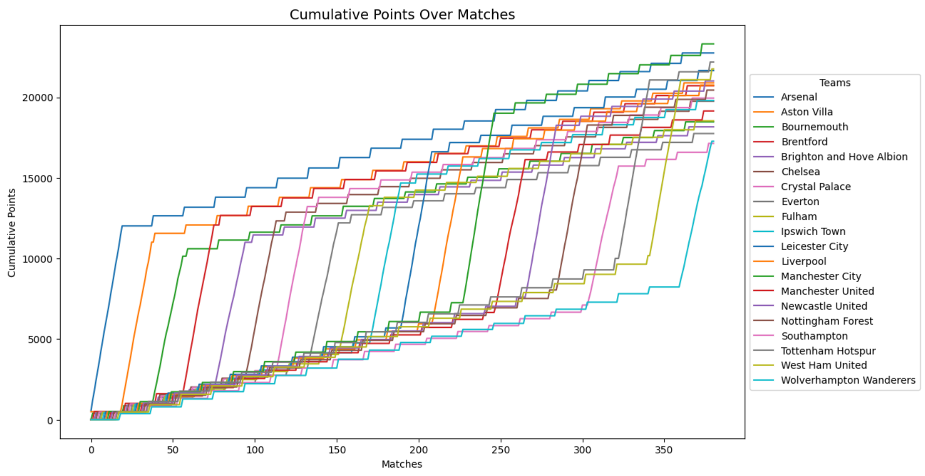
1. **Data Collection**
    a. Historical performance data (e.g., goals scored, goals conceded, shots on target) for teams was scraped from Premier League stats pages using Selenium and BeautifulSoup. The data spans four seasons (2020/21 to 2023/24), with more recent seasons weighted higher to reflect current relevance.
    b. Current season data, including team standings and recent form (match outcomes), was scraped from the Premier League table page. Match results were converted into weighted scores based on recency to calculate a Form score.
2. **Data Cleaning and Processing**
    a. The scraped data was cleaned to standardize team names and fill missing values.
    b. Historical data for offensive strength, defensive strength, and goal conversion rates was calculated by combining season statistics with their respective weights.
    c. The historical performance metrics were merged with current form data into a unified dataset.
3. **Simulation**
    a. A Monte Carlo simulation of the league was performed for 10,000 iterations.
    b. Each iteration simulated all possible matches between teams. Expected goals for both teams in a match were calculated using offensive and defensive strengths, goal conversion rates, and form adjustments. Goals scored were simulated using a Poisson distribution.
    c. Points were assigned based on match outcomes, and total points for each team were recorded.
4. **Results Aggregation**
    a. The average points and win probabilities for each team were calculated from the simulation results.
    b. A summary table was created to display team rankings based on predicted average points and win probabilities.
5. **Visualization**
    a. Scatter plots were generated to visualize offensive vs. defensive strengths across teams.
    b. A cumulative points chart was created to show team performance trends over simulated matches.
    c. Bar charts and tables summarized predicted final standings and performance metrics.

e) Simulation

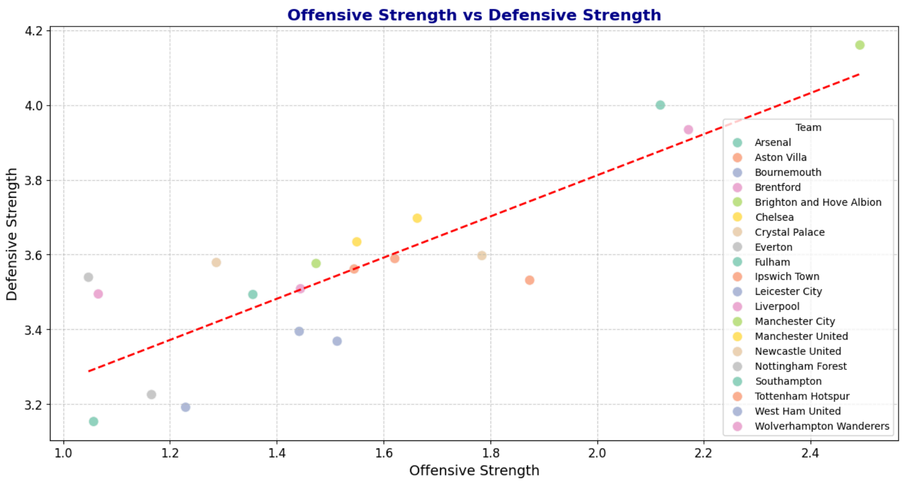
```
                         Team  Avg Points  Win Probability
0             Manchester City     80.3680           0.6634
1                   Liverpool     70.7876           0.1526
2                     Arsenal     68.7517           0.1082
3           Tottenham Hotspur     68.1765           0.0908
4            Newcastle United     60.2023           0.0138
5              Leicester City     58.3842           0.0099
6                 Aston Villa     57.0842           0.0056
7             West Ham United     55.9514           0.0046
8                     Chelsea     52.7701           0.0015
9                   Brentford     51.5926           0.0012
10               Ipswich Town     51.1529           0.0005
11          Manchester United     50.8263           0.0005
12          Nottingham Forest     48.5803           0.0002
13   Brighton and Hove Albion     46.9941           0.0004
14                Bournemouth     44.8986           0.0000
15             Crystal Palace     44.5412           0.0001
16                     Fulham     44.2994           0.0000
17                Southampton     35.0190           0.0000
18                    Everton     34.5353           0.0000
19    Wolverhampton Wanderers     33.6439           0.0000
```



Average Points per Team (Predicted League Table)
Manchester City
Liverpool
Arsenal
Tottenham Hotspur
Newcastle United
Leicester City
Aston Villa
West Ham United
Chelsea
Brentford
Ipswich Town
Manchester United
Nottingham Forest
ton and Hove Albion
Bournemouth
Crystal Palace
Fulham
Southampton
Everton
hampton Wanderers
0
10
20
30
40
50
60
70
80
Average Points



Cumulative Points Over Matches
Cumulative Points
Matches
Teams
Arsenal
Aston Villa
Bournemouth
Brentford
Brighton and Hove Albion
Chelsea
Crystal Palace
Everton
Fulham
Ipswich Town
Leicester City
Liverpool
Manchester City
Manchester United
Newcastle United
Nottingham Forest
Southampton
Tottenham Hotspur
West Ham United
Wolverhampton Wanderers



Offensive Strength vs Defensive Strength
Defensive Strength
Offensive Strength
Team
Arsenal
Aston Villa
Bournemouth
Brentford
Brighton and Hove Albion
Chelsea
Crystal Palace
Everton
Fulham
Ipswich Town
Leicester City
Liverpool
Manchester City
Manchester United
Newcastle United
Nottingham Forest
Southampton
Tottenham Hotspur
West Ham United
Wolverhampton Wanderers

f) Assumptions and Limitations

Key Assumptions

- **Weights:** More recent seasons and matches are weighted higher as they better reflect current team performance.
- **Goals Distribution:** Goals follow a Poisson distribution, a common assumption in sports analytics.
- **Form Influence:** Form impacts performance linearly and uniformly across teams.
- **Independence:** Match outcomes are independent of other matches.

Limitations

- **Data Integrity:** Relies on the accuracy and completeness of web-scraped data.
- **Simplified Form Effects:** The linear influence of form may not capture complex psychological or situational factors.
- **Static Parameters:** Team strengths and conversion rates are assumed constant during a season.
- **Filled-in Data:** Teams not previously a part of the Premier League are supplemented with Offensive Strength and Defensive Strength average to the current table.

g) Conclusion

The Monte Carlo simulation provides a probabilistic forecast of the final standings, reflecting team performance and variability across the season. While the model simplifies certain aspects (e.g., ignoring injuries, tactical changes), it offers valuable insights into likely outcomes and team strengths, with our model predicting **Manchester City** to win the 2024/2025 Premier League.

References

Premier League. (2024). *Tables*. *Premier League*. Available at: https://www.premierleague.com (Accessed: 23 November 2024).

Said, Adil. “Predicting Premier League Match Wins Using Bayesian Modelling.” *Medium*, Medium, 3 Apr. 2024, Available at : https://medium.com/@adilsaid64/predicting-premier-league-match-wins-using-bayesian-modelling-32eec733472e.

W3Schools. “Pandas Tutorial.” Www.w3schools.com, www.w3schools.com/python/pandas/default.asp.

Appendix: Github Repository

The link below leads to the GitHub that contains all code used to simulate the experiments. It also contains the graphs and tables created as a result of the experiments.

https://github.com/Rachel-R16/Applied-Probability-Final-Project