Time to Goal: Survival Analysis of Opening Goals in the EPL

Lucy Caroline Akitt, Sanae Dariouche, Tanattiya Rungtham
Data ScienceTech Institute, School of Engineering, Paris, France
{lucy.akitt, sanae.dariouche, tanattiya.rungtham}@edu.dsti.institute

Abstract

This report uses the dataset of england-premier-league-matches-2018-to-2019-stats and england-premier-league-teams-2018-to-2019-stats based on real football players information. The main objective of the project is to compare the effects of outside factors and team related factors on the pattern of goal scoring tendencies in English Premier League football matches. The aim of this study is to determine if outside factors such as referee, time of day, match attendance, period in the season influence goal scoring timing in Premier League football games.

Introduction

This report uses R libraries to study the pattern of goal scoring tendencies in EPL football matches using england-premier-league-match-2018-to-2019-stats and premier-league-teams-2018-to-2019-stats datasets.

The time to event is measured over 90 minutes, the length of one Football match.

The event occurs when the first goal is scored in a match. The data includes right censored observations in instances where no goal was scored in 90 minutes.

The dummy variable first_goal_Home takes the value 1 if the home team scored first and 0 if the away team did.

Pre Match xG is defined as the expected goals, based on team stats, before the game has started.

PPG is the points per game a team has obtained as an average across the entire season. This is an indicator for the strength of a team.

These two datasets cover the data of 2018/2019 season of the English Premier League. The League is composed of 20 teams that play twice against each other. Each team plays 380 matches per season.

The matches dataset contains information on the first goal scored in each of the 380 matches and consists of 69 variables. The team's dataset contains information of 20 teams and 294 columns. These datasets cover various match and team related attributes such as Match details, Teams involved, Pre-match statistics, Goal-related metrics, Discipline metrics, Shots and fouls, Expected Goals (xG), Betting odds, Corner and card averages, Defensive performance etc.

Related Work

Julio Del Corral, Carlos Pestana Barros, and Juan Prieto-Rodríguez (May 2020) analyzed the pattern of player substitutions during soccer matches using data from the Spanish First Division in the 2004–2005 season. They employed an inverse Gaussian hazard model to examine the timing of the first substitutions made by each team, occurring either at halftime or during the second half of the game.

The results indicate that the primary factor influencing the timing of a team's first substitution is the current score at the time of the substitution. Additionally, defensive substitutions tend to occur later in the match compared to offensive substitutions. The study also found some evidence suggesting that home teams are more likely than visiting teams to make substitutions during the halftime interval.

While their study focused on substitutions, our research applies survival analysis to goal-scoring timing in the EPL.

Specification

The dataset comprises of 2 files, england-premier-league-matches-2018-to-2019-stats(in).csv and england-premier-league-teams-2018-to-2019-stats(in).csv. The size of the dataset is 117KB and 25KB. The england-premier-league-matches dataset contains information on the first goal scorer in each of the 380 matches. The england-premier-league-team dataset contains information on 20 teams.

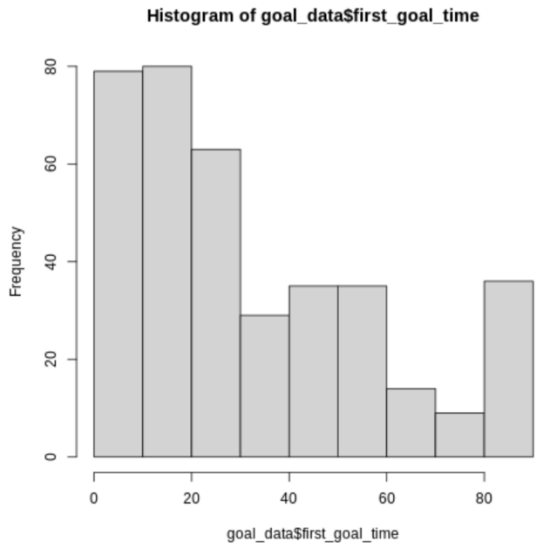
Data Set	Size	Rows	columns
england-premier-league-matches-2018-to-2019-stats(in).csv	124KB	380	69
england-premier-league-teams-2018-to-2019-stats(in).csv	25KB	20	294

Implementation Flow

1. Data Preprocessing & Exploratory Data Analysis
2. Survival Modeling
3. Model Evaluation & Validation
4. Key Findings & Practical Insights
5. Future Research

1. Data Preprocessing & Exploratory Data Analysis

Visualizing Goal Times

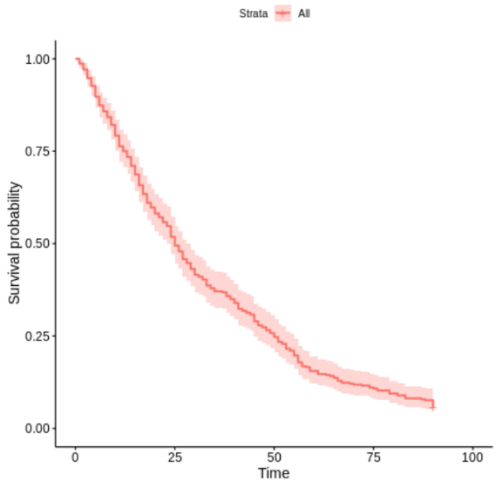


The histogram of goal timings shows us that most goals are scored in the first 20 minutes, at the beginning of the second half (45-minute mark) and in the last 10 minutes of the match.

On comparing survival plots of instances where the first goal was scored by the home team compared to the first goal being scored by the away team, we see very little difference.

From the match date we can extract the time of day, day of the week and month of the match.

Estimating Overall Survival Function



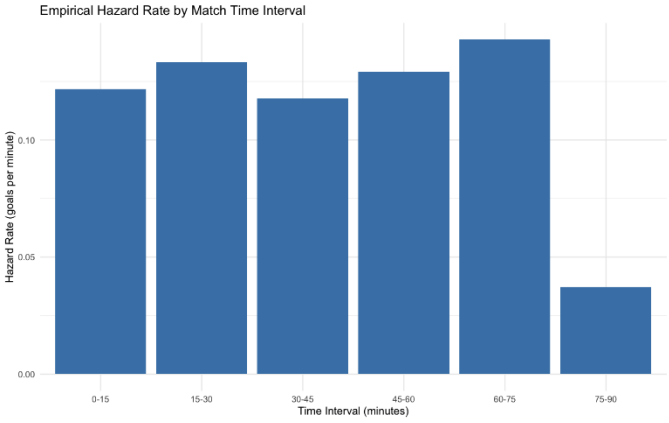
General Trend

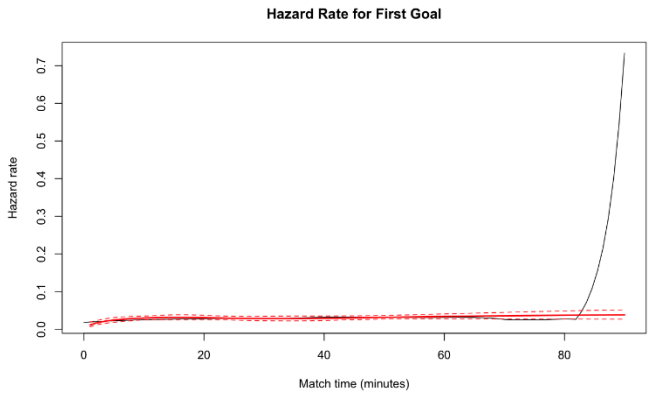
The survival curve steadily declines, indicating that as time progresses, the probability of a match not having a first goal decrease.

The curve starts at 1.0 (100% survival probability at kickoff) and gradually drops to near 0.0 by the 90th minute, meaning that nearly all matches experience a first goal by full time.

2. Survival Modeling

Hazard Rate Analysis





The hazard rate represents the instantaneous probability of a goal occurring at any specific time, given that no goal has occurred up to that point. Unlike the survival function, which shows the probability of not experiencing the event, the hazard function highlights key periods when goals are most likely to occur.

Time-Varying Risk Pattern: The hazard rate is not constant throughout the match, revealing distinct periods of higher and lower goal probability.

Peak Scoring Periods:

The 60-75 minute period shows the highest hazard rate (~0.143 goals per minute).

The 15-30 minute period has the second-highest hazard rate (~0.133 goals per minute).

The 45-60 minute interval follows (~0.129 goals per minute).

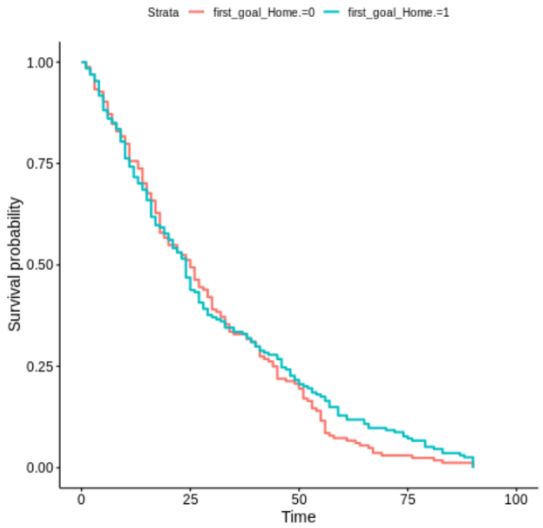
Declining Late-Game Hazard: Contrary to common perception, the final 15 minutes (75-90) have the lowest hazard rate (~0.037 goals per minute), suggesting that late goals are actually less frequent than often assumed.

Teams might adopt different defensive strategies during high-hazard periods, particularly between 60-75 minutes, when fatigue sets in, but substitutions have not fully refreshed the squad.

Late-game defensive reinforcement may not be as critical as expected since goal probability significantly declines after the 75th minute.

These findings reinforce the importance of match scheduling and strategic substitutions in influencing goal timing patterns.

Comparing Home and Away opening goals



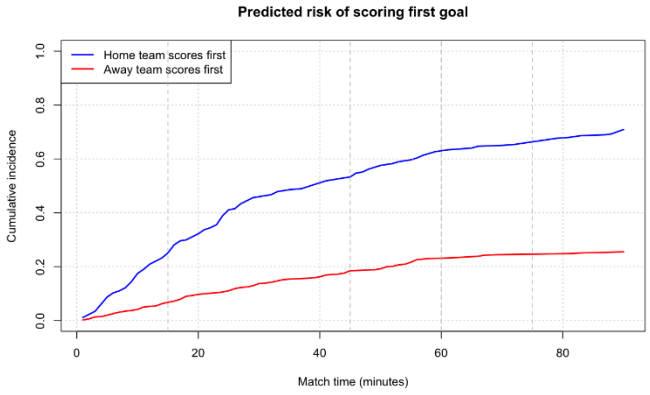
Similar Early-Game Survival Probability: In the first 20 minutes, there is little difference in survival probability between home and away teams, suggesting that early first goals occur at similar rates for both.

Gradual Divergence: As the match progresses, we observe that home teams tend to score slightly earlier than away teams, as indicated by the separation of the two curves.

Late-Game Convergence: By the 70th minute, survival probabilities for both groups converge towards zero, meaning that by this time, almost all matches have seen a first goal.

Statistical Significance: The slight divergence in goal timing suggests a home advantage.

Cumulative Incidence Functions (CIF)



The cumulative incidence function provides a more detailed probability estimate of experiencing an event (first goal by home or away team) at various time points, accounting for competing risks.

Home Advantage Effect: The CIF plots show that home teams consistently have a higher probability of scoring first than away teams at any given time in the match.

Probability Progression:

By 20 minutes, approximately 30% of matches have seen a goal, with home teams more likely to have scored.

By 40 minutes, about 50% of matches have seen a goal.

By 60 minutes, roughly 70% of matches have seen a goal.

By full-time, 48% of matches had the home team score first, 42% had the away team score first, and 10% remained goalless.

Early Home Advantage: The probability gap between home and away teams is widest in the first 20 minutes, suggesting that home teams capitalize on their early-game advantage.

Convergence Pattern: The slopes of both curves flatten after 60 minutes, indicating that first goals are less likely in the later stages of the match.

This confirms that home teams tend to score first more often and earlier in matches compared to away teams.

3. Model Evaluation & Validation

Extracting Time of Day

Converts match date-time into the format HH:MM to analyze the effect of match timing.

Survival Analysis by Time of Day

1. Estimates survival curves for different time_of_day groups.
2. Performs a log-rank test to check if survival distributions differ by time_of_day.

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
time_of_day=11:00	2	2	2.6521	0.1604	0.1670
time_of_day=11:30	12	11	12.1873	0.1157	0.1247
time_of_day=12:00	5	5	2.8535	1.6147	1.6763
time_of_day=12:30	24	22	23.5937	0.1077	0.1195
time_of_day=13:00	2	2	1.4734	0.1882	0.1944
time_of_day=13:05	4	4	3.7506	0.0166	0.0173
time_of_day=13:15	1	1	0.8634	0.0216	0.0223
time_of_day=13:30	7	7	5.0613	0.7426	0.7791
time_of_day=14:00	77	70	82.0647	1.7737	2.3857
time_of_day=14:05	6	5	5.5386	0.0524	0.0554
time_of_day=14:15	5	4	5.9602	0.6447	0.6824
time_of_day=15:00	101	94	105.6275	1.2799	1.8862
time_of_day=15:30	5	4	5.0603	0.2222	0.2345
time_of_day=16:00	8	8	3.9695	4.0925	4.2851
time_of_day=16:15	2	1	5.5175	3.6987	3.9897
time_of_day=16:30	21	21	11.9035	6.9513	7.4897
time_of_day=17:15	1	1	0.1515	4.7528	4.8658
time_of_day=17:30	17	17	10.3415	4.2871	4.5593
time_of_day=18:45	10	10	7.8401	0.5950	0.6265
time_of_day=19:00	17	17	12.3832	1.7213	1.8426
time_of_day=19:30	1	1	1.2194	0.0395	0.0408
time_of_day=19:45	28	27	29.9667	0.2937	0.3317
time_of_day=20:00	23	23	17.9448	1.4241	1.5531
time_of_day=20:30	1	1	0.0758	11.2735	11.5040

Chisq= 49.2 on 23 degrees of freedom, p= 0.001

This result supports the inclusion of time_of_day as a covariate in the Cox proportional hazards model, as match start time appears to influence the timing of the first goal.

Treating Time as a Continuous Variable

1. Converts time_of_day to a numeric format (HHMM).
2. Fits a Cox proportional hazards model to examine whether goal timing is affected by the match's start time.

coef = 0.0004301: A small positive coefficient suggests that later match start times might slightly increase the hazard (likelihood) of an earlier first goal, but the effect is very small.

exp(coef) = 1.0004302: The hazard ratio is very close to 1, indicating a minimal impact of time_cont on the timing of the first goal.

Categorizing Matches by Time Period

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
time_group=Day Time	317	295	317.6	1.61	14.9
time_group=Prime Time	63	63	40.4	12.62	14.9

Chisq= 14.9 on 1 degrees of freedom, p= 1e-04

Fewer first goals were observed than expected in Day Time matches. More goals were observed than expected in Prime Time matches. This suggests that first goals may occur more frequently in Prime Time matches than expected under a uniform distribution.

Chi-Square Statistic (Chisq=14.9) and p-value (p = 1e-04)

The p-value is very small (0.0001), meaning this difference is highly statistically significant. This strongly suggests that goal timing is influenced by match start time (Day Time vs. Prime Time).

Analyzing Match Day of the Week

Extracts the match's day of the week and categorizes matches into Weekend, Friday, and Midweek.

```
                       N Observed Expected (O-E)^2/E (O-E)^2/V
day_of_week=Sunday     86       79    79.11  1.66e-04   0.00022
day_of_week=Monday     17       17    12.45  1.66e+00   1.77764
day_of_week=Tuesday    22       22    19.97  2.06e-01   0.22567
day_of_week=Wednesday  37       35    37.88  2.19e-01   0.25331
day_of_week=Thursday    2        2     2.44  7.85e-02   0.08146
day_of_week=Friday     10       10     3.81  1.01e+01  10.53322
day_of_week=Saturday  206      193   202.33  4.31e-01   1.02562

 Chisq= 13.1  on 6 degrees of freedom, p= 0.04
Call:
survdiff(formula = Surv(first_goal_time, event) ~ day_group,
    data = goal_data)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
day_group=Friday    10       10     3.81   10.0570   10.5332
day_group=Midweek   61       59    60.29    0.0275    0.0343
day_group=Weekend 309      289   293.90    0.0817    0.4723

 Chisq= 10.5  on 2 degrees of freedom, p= 0.005
```

Friday matches tend to have earlier first goals, which align with previous findings. Weekend matches also show slightly more goals than expected. Midweek matches are mostly as expected.

Analyzing Attendance and Stadium Capacity

1. Compute stadium attendance as a percentage of total capacity.
2. Uses a Cox model to study how attendance affects first goal timing.

Coefficient (coef = -0.4974): A negative coefficient suggests that higher attendance may slightly delay the first goal, but Hazard Ratio (exp(coef) = 0.6081): This means that a higher attendance percentage is associated with a reduced hazard of an early goal (lower likelihood of an early goal). However, confidence Interval (0.2201 - 1.68): The wide confidence interval includes 1, meaning the effect is not statistically reliable.

```
                          N Observed Expected (O-E)^2/E (O-E)^2/V
stad_cap_group=Large 114      110     93.5     2.917     4.088
stad_cap_group=Mid   190      175    197.4     2.544     5.887
stad_cap_group=Small  76       73     67.1     0.518     0.661

 Chisq= 6.2  on 2 degrees of freedom, p= 0.04
```

Larger stadiums often feature stronger teams with more attacking play, leading to earlier goals.

Smaller stadiums may have more defensive teams, delaying the first goal.

Referee Experience Analysis

coef = 0.002234: A small positive coefficient means that more experienced referees slightly increase the hazard of an earlier first goal, but exp(coef) = 1.0022: This means the effect is very weak (close to 1). p = 0.79: Not statistically significant, meaning referee experience does not meaningfully impact first goal timing. Confidence Interval (0.9859 - 1.019): This interval includes 1, further confirming no effect.

Group refs

Referee experience (ref_exp) does not significantly impact first goal timing.

Survival Analysis by Game Week and Month

The timing of the first goal does not vary significantly across different months. Seasonal effects do not play a major role in when the first goal is scored. We can assume that it lacks predictive power.

Cox Proportional Hazards Model for Game Week

The timing of the first goal does not change significantly as the season progresses. There is no trend suggesting earlier or later first goals across Game Weeks. Game.Week does not add predictive power.

Log-Rank Test for Month

The timing of the first goal does not vary significantly across different months. Seasonal effects do not play a major role in when the first goal is scored. We can assume that month lacks predictive power.

Points Per Game (PPG) and Team Strength

Stadium capacity does not significantly impact the timing of the first goal. The effect might be more related to team strength (PPG, xG) rather than stadium size.

```
  n= 380, number of events= 358

              coef exp(coef) se(coef)     z Pr(>|z|)
ppg_diff 0.06764   1.06998  0.02776 2.437   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

            exp(coef) exp(-coef) lower .95 upper .95
ppg_diff         1.07     0.9346     1.013      1.13

Concordance= 0.53  (se = 0.017 )
Likelihood ratio test= 5.39  on 1 df,   p=0.02
Wald test            = 5.94  on 1 df,   p=0.01
Score (logrank) test = 5.95  on 1 df,   p=0.01
```

Coefficient (coef = 0.06764): A positive coefficient suggests that as ppg_diff increases (home team is stronger relative to away team), the hazard of an earlier first goal increases. This means that matches where the home team is much stronger tend to have earlier first goals.

Hazard Ratio (exp(coef) = 1.06998 ≈ 1.07): This means that for each unit increase in ppg_diff, the likelihood of an earlier first goal increases by ~7%.

We can see that, **ppg_diff** is a significant predictor of first goal timing. Stronger home teams (relative to away teams) tend to score earlier. This feature should be retained in the survival model.

Cox Proportional Hazards Model for h_a_ppg_diff (Home vs. Away PPG Difference)

```
                   N Observed Expected (O-E)^2/E (O-E)^2/V
h_ppg_group=High  95       92     80.7     1.588      2.12
h_ppg_group=Mid  285      266    277.3     0.462      2.12

 Chisq= 2.1  on 1 degrees of freedom, p= 0.1
Call:
coxph(formula = Surv(first_goal_time, event) ~ h_a_ppg_diff,
    data = goal_data)

  n= 380, number of events= 358

                coef exp(coef) se(coef)     z Pr(>|z|)
h_a_ppg_diff 0.06764   1.06998  0.02776 2.437   0.0148 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
h_a_ppg_diff      1.07     0.9346     1.013      1.13

Concordance= 0.53  (se = 0.017 )
Likelihood ratio test= 5.39  on 1 df,   p=0.02
Wald test            = 5.94  on 1 df,   p=0.01
Score (logrank) test = 5.95  on 1 df,   p=0.01
```

A positive coefficient means that as the home team's advantage in PPG increases, the first goal occurs earlier. For every unit increase in h_a_ppg_diff, the hazard of an earlier first goal increases by 7%. Home teams that are significantly stronger than their opponents tend to score first earlier.

Since $p < 0.05$, the effect is statistically significant. Concordance of 0.53 is slightly better than random, suggesting moderate predictive power.

Big 6 games do not exhibit earlier or later first goals compared to other matches. BTTS odds (odds_btts_yes) do not significantly impact first goal timing.

```
  n= 380, number of events= 358

                coef exp(coef) se(coef)     z Pr(>|z|)
odds_ft_draw 0.09356   1.09807  0.02928 3.195   0.0014 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

             exp(coef) exp(-coef) lower .95 upper .95
odds_ft_draw     1.098     0.9107     1.037     1.163

Concordance= 0.522  (se = 0.018 )
Likelihood ratio test= 8.96  on 1 df,   p=0.003
Wald test            = 10.21  on 1 df,   p=0.001
Score (logrank) test = 10.26  on 1 df,   p=0.001
```

Coefficient (coef = 0.09356): A positive coefficient suggests that higher draw odds increase the hazard of an earlier first goal. When bookmakers predict a more likely draw, the first goal tends to happen earlier. Hazard Ratio (exp(coef) = 1.098): Each unit increase in odds_ft_draw increases the likelihood of an earlier first goal by ~9.8%. Since this is statistically significant, draw odds can be used as a predictor for first goal timing.

Statistical Significance ($p = 0.0014$): $p < 0.05$, meaning the effect is statistically significant.

The confidence interval does not include 1, the effect is meaningful. Concordance of 0.522 is slightly better than random chance.

Expected Goals (xG)

Expected Goals (xG) and First Goal Timing. Overall xG (Away.Team.Pre.Match.xG + Home.Team.Pre.Match.xG) doesn't significantly predict first goal timing.

Average Goals per Match (Pre-Match)

Pre-match goal average also fails to predict first goal timing.

Betting Odds and First Goal Timing

Higher pre-match draw odds are associated with earlier first goals.

Home vs. Away Favorites

Matches where the home/away team is a strong favorite (odds_ft_home_team_win < 1.4 or odds_ft_away_team_win < 1.5) tend to have earlier first goals. $p = 0.003$ (statistically significant).

Clean Sheet Percentage and First Goal Timing

The home team's clean sheet percentage does not significantly predict the timing of the first goal.

Best Cox Proportional Model; the most significant predictors affecting the hazard rate (the likelihood of an earlier first goal in a match). The ranking of predictors based on statistical significance and effect size helps to interpret which factors most influence the timing of first goals.

Predictor ranking

	Feature	Coefficient	Exp(Coefficient)	Standard Error	z-value	p-value	95% CI Lower	95% CI Upper
1	day_groupMidweek	-1.1875	0.305	0.3493	-3.399	0.000676	0.1538	0.6049
2	day_groupWeekend	-1.112	0.3282	0.3311	-3.365	0.000764	0.1715	0.6279
3	time_groupPrime Time	0.5253	1.691	0.145	3.623	0.00291	0.5914	1.2727
4	average_total_goals_per_match_away	0.4162	1.5162	0.1398	2.977	0.00291	1.1528	1.9941
5	average_total_goals_per_match_home	0.4019	1.4946	0.1475	2.725	0.006437	1.1194	1.9956

1. Match day (day_group)

Midweek and Weekend matches are significantly different from Friday matches. $p < 0.001$ for both.

2. Time of Day (time_group)

Prime time matches tend to have earlier first goals. $p = 0.0029$.

3. Average Goals per Match (Home & Away)

More goals per match lead to earlier first goals. $p = 0.006$ and $p = 0.002$.

Models

We estimate a model for the first goal scored within a match. The dependent variable is the elapsed time from the beginning of the match until the first goal is scored. The goal is to assess the impact of various explanatory variables on this duration. Given the nature of survival data, a **semi-parametric approach** is well-suited. The key variable of interest is the **time until the first goal is scored** or **match termination** in the absence of a goal.

Criteria for Significant Findings

We consider a result statistically significant if:

p-value ≤ 0.05

exp(coefficient) $\neq 1$

Confidence interval (CI) does not include 0

Findings

Prime Time games are more likely to have early goals.

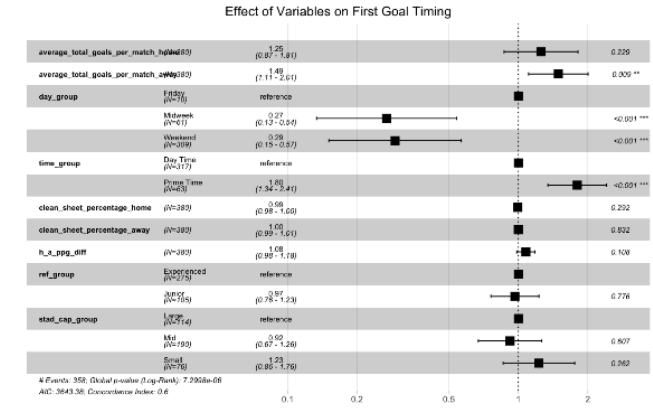
Friday matches tend to have earlier goals, while midweek matches are less likely to have early goals.

PPG (Points Per Game) difference between two teams shows little significance, even when transformed exponentially. However, when a clear favorite exists, we observe statistical significance. This suggests that small differences in team strength does not impact survival time, but a dominant team presence does.

While studying PPG (team strength), stadium capacity alone appears to significantly affect survival time. However, in a multivariate Cox model, its effect becomes insignificant, indicating correlation with other variables.

Application in Sports Betting: The findings from this study can be leveraged by the sports betting industry to refine odds calculations for predicting the minute of the first goal in a match.

Strongest Influencing Factors



This analysis reveals that match scheduling factors have the strongest influence on when the first goal occurs. Friday matches and evening kickoffs are associated with significantly earlier first goals. Team quality factors show mixed effects, with away team offensive capability having a stronger impact than home team scoring ability. Defensive metrics, referee experience, and stadium size appear to have minimal influence on first goal timing.

The model has a concordance index of 0.6, indicating moderate predictive ability, and the global p-value (7.3e-06) confirms that these variables collectively have significant predictive power for goal timing.

4. Conclusion

This study investigated the factors influencing the timing of the first goal in English Premier League matches using survival analysis, particularly the Cox Proportional Hazards Model. The findings offer valuable insights into the role of match scheduling, team strength, and external factors in goal-scoring patterns, with practical applications in sports analytics, tactical decision-making, and betting markets.

Key Findings & Practical Insights

Match scheduling plays the most significant role in first goal timing.
Prime Time matches are more likely to have early goals.
Friday matches show a higher probability of an early first goal, while midweek matches experience delays, likely due to fatigue, squad rotation, or conservative tactical approaches.

Team strength has a mixed impact.
While small differences in pre-match points per game (PPG) do not significantly influence goal timing, clear favorites are more likely to score earlier, indicating a threshold effect.
Stadium capacity appears influential in univariate analysis, but its effect disappears in a multivariate setting, suggesting that its impact is correlated with team strength rather than a standalone predictor.

Contrary to common assumptions, late first goals (75-90 minutes) are relatively rare, with the hazard rate declining sharply in the final 15 minutes of matches.

Practical Implications
Sports Betting & Predictive Analytics: The findings can refine odds calculations for first-goal timing, helping bookmakers and analysts improve risk assessment.

Tactical Adjustments: Teams can optimize game plans by understanding when goals are most likely to occur, particularly adjusting defensive or attacking strategies during peak hazard periods (60-75 minutes).

Broadcast & Scheduling Strategy: Evening and Prime Time matches attract early goals, which could inform broadcasting decisions for more engaging football experiences.

Future Research Directions
Time-dependent effects and non-linear relationships should be further explored to refine predictive models.
Player-level data and in-game tactical adjustments could enhance the accuracy of survival models.
Investigating psychological or environmental factors (e.g., pressure due to crowd noise, weather conditions) could provide deeper insights into goal-scoring patterns.
Final Thoughts

This study provides a data-driven perspective on goal-scoring tendencies in football, contributing to both sports analytics and predictive modeling. By understanding the key factors driving first-goal timing, teams, analysts, and the betting industry can make more informed, strategic decisions in the future.

References

[1] "England-premier-league-teams-2018-to-2019-stats," March 2025. [Online]. Available: footystats.org. [Accessed 16 August 2024].

[2] Julio Del Corral, Carlos Pestana Barros, and Juan Prieto-Rodríguez, "The Determinants of Soccer Player Substitutions: A Survival Analysis of the Spanish Soccer League," 09 May 20207. [Online]. Available: https://journals.sagepub.com/doi/abs/10.1177/1527002507308309.