

# Predicting Goal Outcomes in Ice Hockey Using Spatial-Contextual Data and Ensemble Machine Learning Models

Group name: Pandaria

Group member: Haoran Hua, Zekai Li, Bin Han, Kairui Li

**Abstract:** This study applies machine learning to predict goal-scoring events in ice hockey using the Linhac24-25 dataset. We frame it as a binary classification problem and use Random Forest and XGBoost to model spatial, contextual, and temporal features. XGBoost slightly outperforms Random Forest, particularly in handling class imbalance and rare event detection. Key predictors include puck location, manpower situation, and expected goals. Visualizations support model interpretability, and the findings align with domain knowledge, suggesting practical value for tactical and real-time applications.

**Keywords:** Ice hockey analytics, goal prediction, machine learning, Random Forest, XGBoost, spatial features, event classification, ensemble learning

## 1 Introduction

The capacity to predict whether a specific game event will lead to a goal represents a significant advancement in the field of sports analytics, particularly within the fast-paced and dynamic context of ice hockey. Accurate event-level prediction has the potential to inform coaching strategies, refine player evaluation processes, and enhance audience engagement through deeper insights.

This study aims to classify individual events as either goal or non-goal outcomes by analyzing detailed records from the Linhac24-25 dataset, which comprises over 500,000 logged hockey events. To capture the complex relationships inherent in the data, we employ machine learning techniques, focusing specifically on Random Forest and XGBoost classifiers due to their proven robustness and suitability for structured datasets. The primary objective is to identify the most influential features—such as position, player roles, and contextual game conditions—that contribute to the likelihood of a goal, and to evaluate the effectiveness of these models in predicting such rare but strategically significant events.

## 2 Background

Ice hockey is a dynamic and fast-paced sport that involves two teams of six players each (including a goaltender) competing to score goals by shooting a puck into the opponent's net. The game is divided into three periods, each lasting 20 minutes, with additional overtime or shootouts in the event of a tie. Given the fluidity of the game, numerous micro-events such as passes, blocks, receptions, checks, and shots occur within seconds, creating a complex and rich stream of data.

In the dataset used for this study, each row represents an individual in-game event and contains a diverse set of features. Spatial coordinates (`xadjcoord`, `yadjcoord`) and temporal features such as `compiledgametime` reflect when the event occurred. Contextual features like `manpowersituation`, `scoreddifferential`, and `ishomegame` capture the broader game state. Player-related attributes such as `playerprimaryposition` and `playerid` provide information about the roles and identities of the participants. Lastly, categorical features like `type`, `outcome`, and `teaminpossession` define the nature of the event and whether it was successful.

Understanding these components is crucial. For instance, the `manpowersituation` reflects whether the team was on a power play, short-handed, or at even strength—situations that drastically change scoring likelihood. Similarly, `scoreddifferential` affects strategic decision-making, as teams may play more conservatively or aggressively depending on the score.

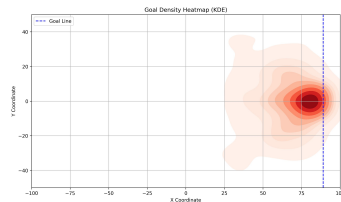
## 3 Algorithms

To address the prediction task, our project adopted two tree-based ensemble models: Random Forest (RF) and XGBoost (Extreme Gradient Boosting), both well-suited for structured data and capable of modeling complex nonlinear relationships. Random Forest constructs multiple decision trees and aggregates their predictions through majority voting. Its key strengths include handling missing and categorical data, reducing variance to mitigate overfitting, and providing feature importance for interpretability. We implemented two RF variants: RF1, a baseline model using selected features and one-hot encoding; and RF2, an improved version applying label encoding to all categorical variables and using `class_weight='balanced'` to address class imbalance.

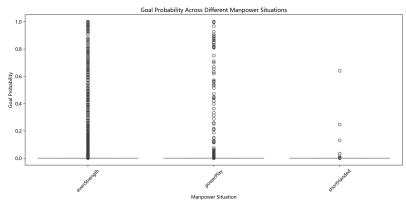
XGBoost is a fast and scalable gradient boosting algorithm widely used in both academic and applied contexts. It incorporates regularization to prevent overfitting, supports parallel training, and handles missing values automatically. We developed two XGBoost models: XGB1, a basic version with minimal preprocessing; and XGB2, an optimized model with one-hot encoding, class balancing, and visualization components to improve performance and interpretability.

## 4 Conclusion and Discussion

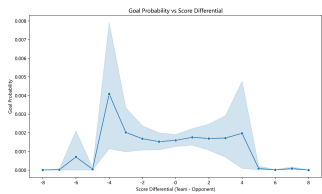
Our project reveals several insights that are relevant to in-game decision-making. Specifically, the majority of goals originate from the low-slot area directly in front of the net, reinforcing the strategic importance of controlling this zone offensively and defending it aggressively as shown in the goal density heatmap (**Fig. 1**). Shots taken during power play scenarios show significantly higher predicted scoring probabilities, indicating that maximizing shot volume and shot quality during man-advantage situations should be a tactical priority — a pattern clearly illustrated in **Fig. 2**. Furthermore, teams that are trailing by several goals exhibit a brief increase in scoring likelihood — likely due to more aggressive offensive play — while extreme leads correlate with a decline in scoring probability, possibly reflecting strategic conservatism or line rotations. This trend is visualized in **Fig. 3**, which shows the fluctuation of goal probability across different score differentials.



**Fig. 1. Goal Density Heatmap (KDE)**



**Fig. 2. Goal Probability Across Manpower Situations**



**Fig. 3. Goal Probability vs. Score Differential**

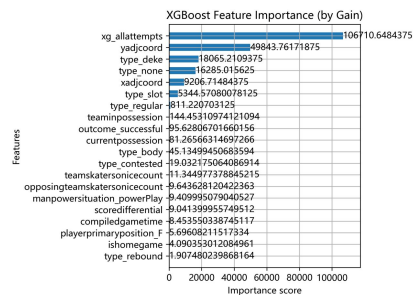
Building on these patterns, our machine learning models yielded strong performance in predictive accuracy. Both Random Forest and XGBoost classifiers demonstrated high overall precision and recall, with XGBoost outperforming Random Forest in terms of F1-score (0.44 vs. 0.2364) and recall (0.45 vs. 0.1585) for the rare positive class. Balancing class distributions through `class_weight='balanced'` in Random Forest and loss adjustments in XGBoost proved essential for reliable results.

Confusion matrix analysis confirmed this performance gap. XGBoost correctly identified 70 true positive goal events, compared to 26 by Random Forest, albeit with more false positives (91 vs. 30). This trade-off is acceptable in tactical settings where anticipating threats is more important than avoiding overprediction.

Feature importance analysis aligned closely with domain knowledge. Spatial coordinates (`xadjcoord`, `yadjcoord`) consistently ranked highest across both models, highlighting the influence of shot location. The `xg_allattempts` variable—used as a proxy for expected goals—was the top feature in XGBoost, as shown in the feature importance plot (**Fig. 4**) validating its theoretical value.

Visualizations strengthened these conclusions. Hexbin and KDE heatmaps showed concentrated goal densities near the crease. Scatter plots of predicted probabilities aligned well with actual shot patterns. Boxplots confirmed that power plays offer elevated scoring chances, and line plots revealed how goal probability fluctuates based on score differential—offering a view into tactical behavior shifts under different game states.

Finally, model interpretability was enhanced through feature rankings. Random Forest prioritized spatial and temporal features like `compiledgametime`, while XGBoost also highlighted contextually rich features such as shot type and puck control. These findings confirm that both models successfully learned context-aware and spatially grounded scoring patterns.



**Fig. 4. XGBoost Feature Importance Plot**

## 5 Summary

This study applies machine learning techniques to predict goal outcomes in ice hockey using over 500,000 events from the Linhac24-25 dataset. Random Forest and XGBoost classifiers were used to model event-level data, incorporating spatial, temporal, and contextual features. XGBoost achieved superior performance, particularly in recall and F1-score, and identified key predictors such as puck location and expected goals (xg\_allattempts). Visualizations confirmed that most goals occur near the crease and are influenced by factors like manpower situation and score differential. The results provide both accurate prediction and practical insights for tactical decision-making in hockey.

## 6 Future Work

Several avenues remain for extending this research. One promising direction involves incorporating temporal modeling, such as Recurrent Neural Networks (RNNs) or Long Short-Term Memory (LSTM) networks, to capture the sequential flow of events. This would allow the model to evaluate entire offensive sequences rather than isolated actions, potentially improving prediction of goal outcomes in dynamic game contexts.

Additionally, expanding the dataset with richer and more granular inputs—such as player movement tracking, puck speed, fatigue levels, or pressure zones—could significantly improve the model’s situational awareness. Combining structured event data with video-based features via deep learning may also yield deeper strategic insights.

Improving model interpretability remains important for practical application. Methods such as SHAP or LIME could help explain individual predictions, increasing trust and usability. Developing team- or player-specific models could further tailor outputs to specific tactical styles or roster compositions.

Lastly, evaluating generalization across different seasons or leagues and deploying the system as an API or interactive tool would enhance its robustness and utility, making it accessible to broader stakeholders such as coaching staff, analysts, and media professionals.

Our project can be found on: [https://gitlab.liu.se/tdde64\\_pandaria/tdde64-pandaria](https://gitlab.liu.se/tdde64_pandaria/tdde64-pandaria)