From Data to Goals: Predicting Hockey Scoring with Machine Learning

Travis Bates
Capstone 2 Presentation

Problem Statement

What opportunities exist for T.B Sports to increase average client contract by 25% for the 2025-2026 season using a machine learning to predict a player's season statistics?

Background

- ❑ TB sports is a hypothetical new and upcoming sports management agency representing professional athletes across multiple sports disciplines.
- ❑ The hockey division has seen a decline in clients and did not have an increase in average client contract from the previous year.
- ❑ Management want to turn to predictive analytics to help negotiate future contracts, endorsements deals, and guide future recruitment strategies.

Business Objective

Provide data-driven insights into player scoring efficiency

Help TB Sports identify undervalued talent and breakout stars based on goal probability, not raw totals.

Turning insights into better player contracts.

Data → Model → Insight → Contract

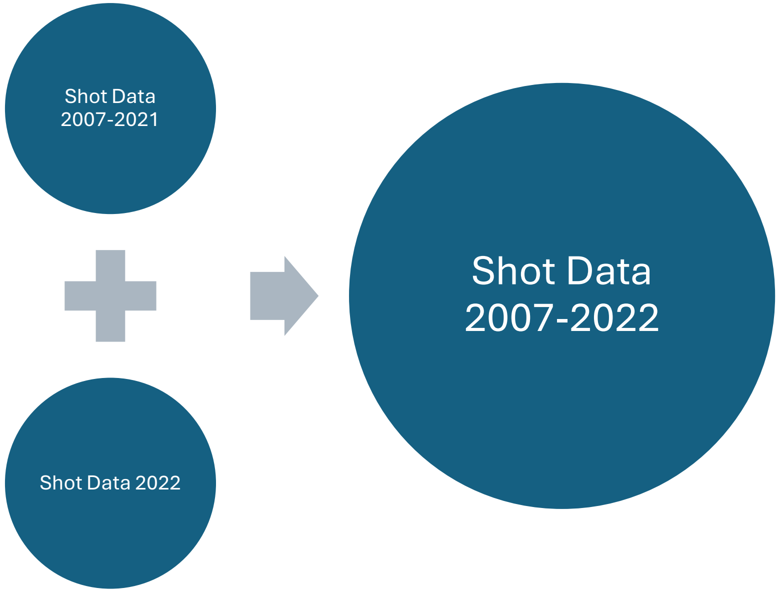
Data Overview

Source: NHL shot event data for every game including both regular season and playoffs from 2007 to 2022

- Kaggle- NHL Data - Player, team, and shots data from 2008-2023
- Money Puck - https://moneypuck.com/about.htm

	Raw Data	Cleaned Data
Number of Records	~1.7 million	~1.2 million
Number of Features	124	47
Missing Values	Present in several key features	Dropped
Duplicates	Present in some key features	Removed
Categorical Variables	13 unencoded features	All encoded (One-Hot, Target, and Frequency)
Outliers	Present	Removed
Time Frame	2007-2022	2007-2022

Data Wrangling



Missing Values

- Present in several key features of the merged data set
- Missing values dropped since they accounted for less than 1% of total data

Feature Reduction

- Data set had redundant identifiers and features that could lead to data leakage.
- Domain-informed feature selection was used to retain only features with a clear relevance to goal-scoring prediction

Outliers

- Data contained some impossible values such as having only 2 or 7 skaters on the ice and shots from behind the goal.
- All outliers were dropped.

Data Analysis

Key Features

- Shot Type
- Distance from Goal
- Shot Angle
- Situation
- Player Position
- Shooter Handiness
- Shot Rebounds

Data Analysis – Shot Type

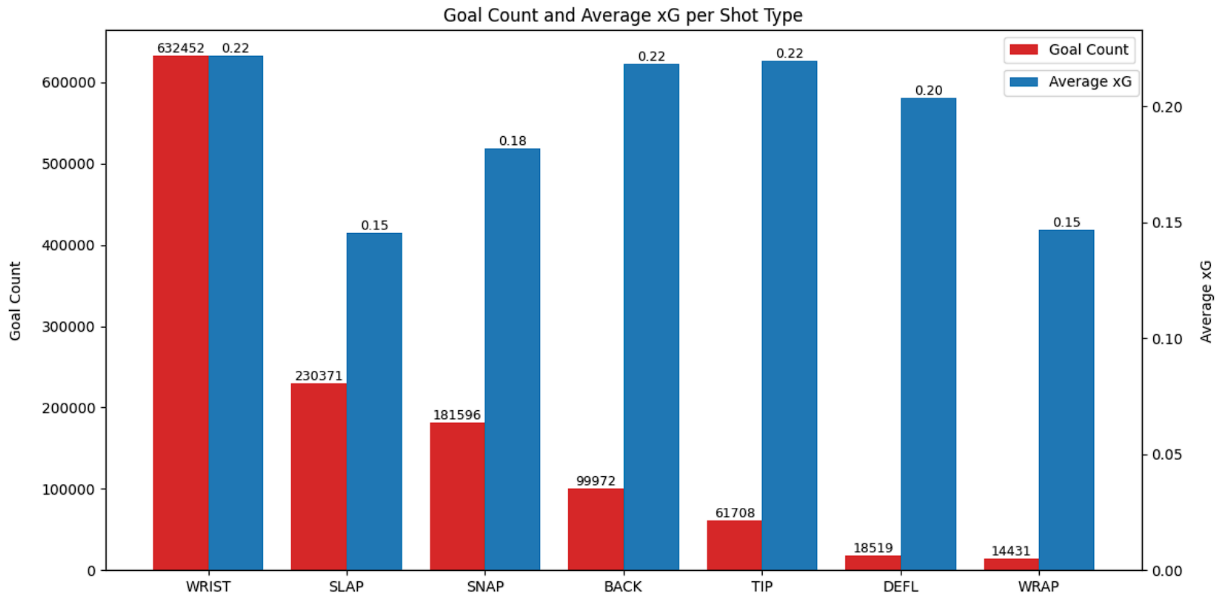


Figure1: Goal count and average xGoal by shot type

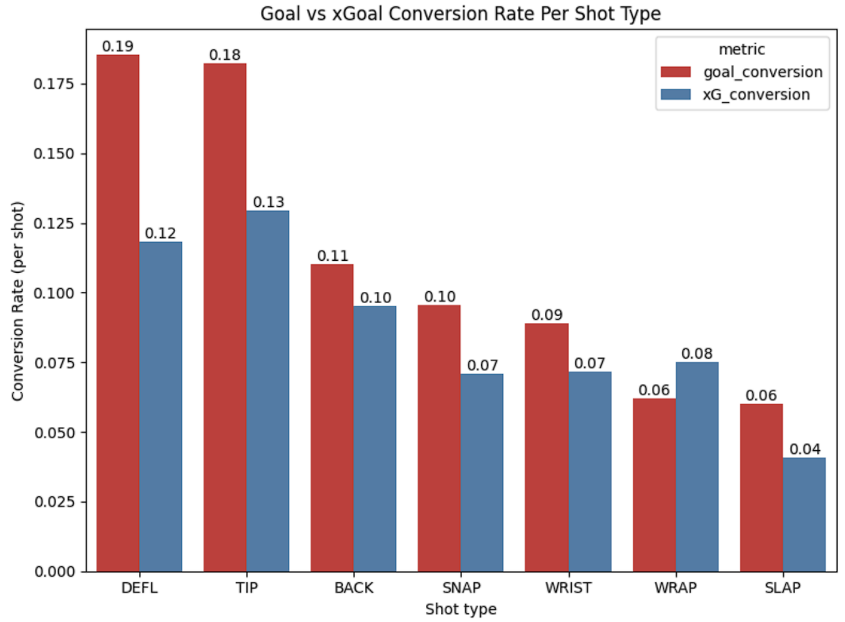


Figure 2: Goal conversion and xGoal conversion percentage rates by shot type

Data Analysis – Shot Distance

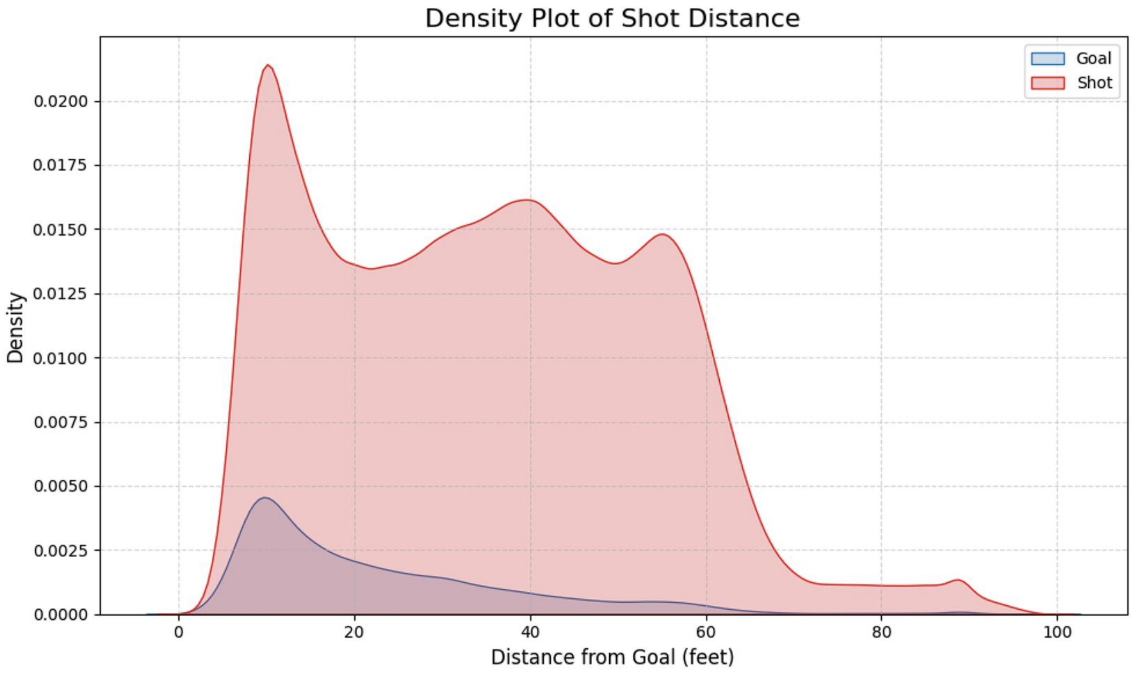


Figure 3: Density Plot of shot distances for all shots and all goals

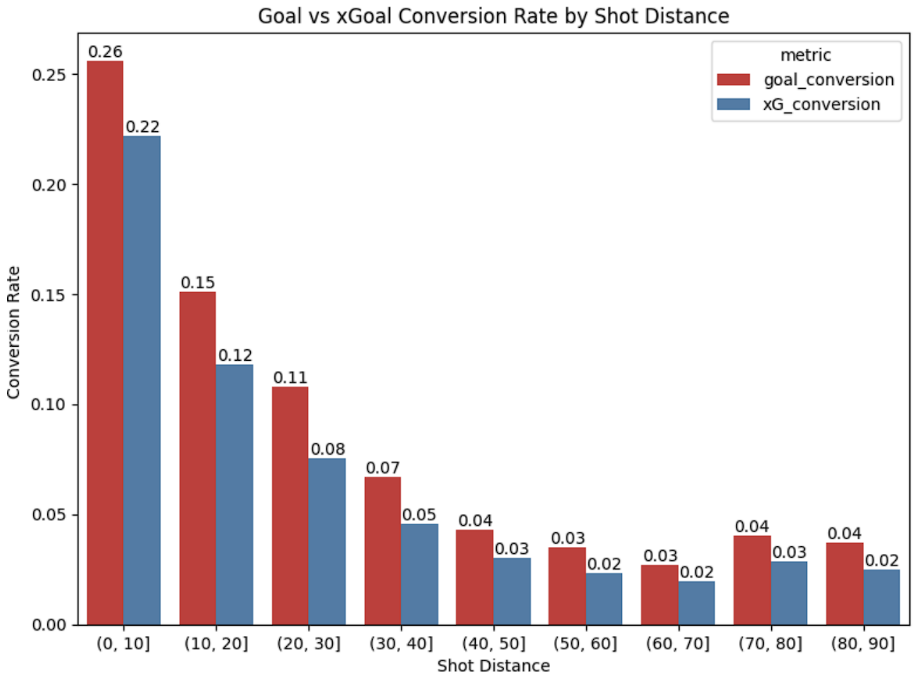


Figure 4: Goal conversion and xGoal conversion by shot distance

Data Analysis – Shot Angle

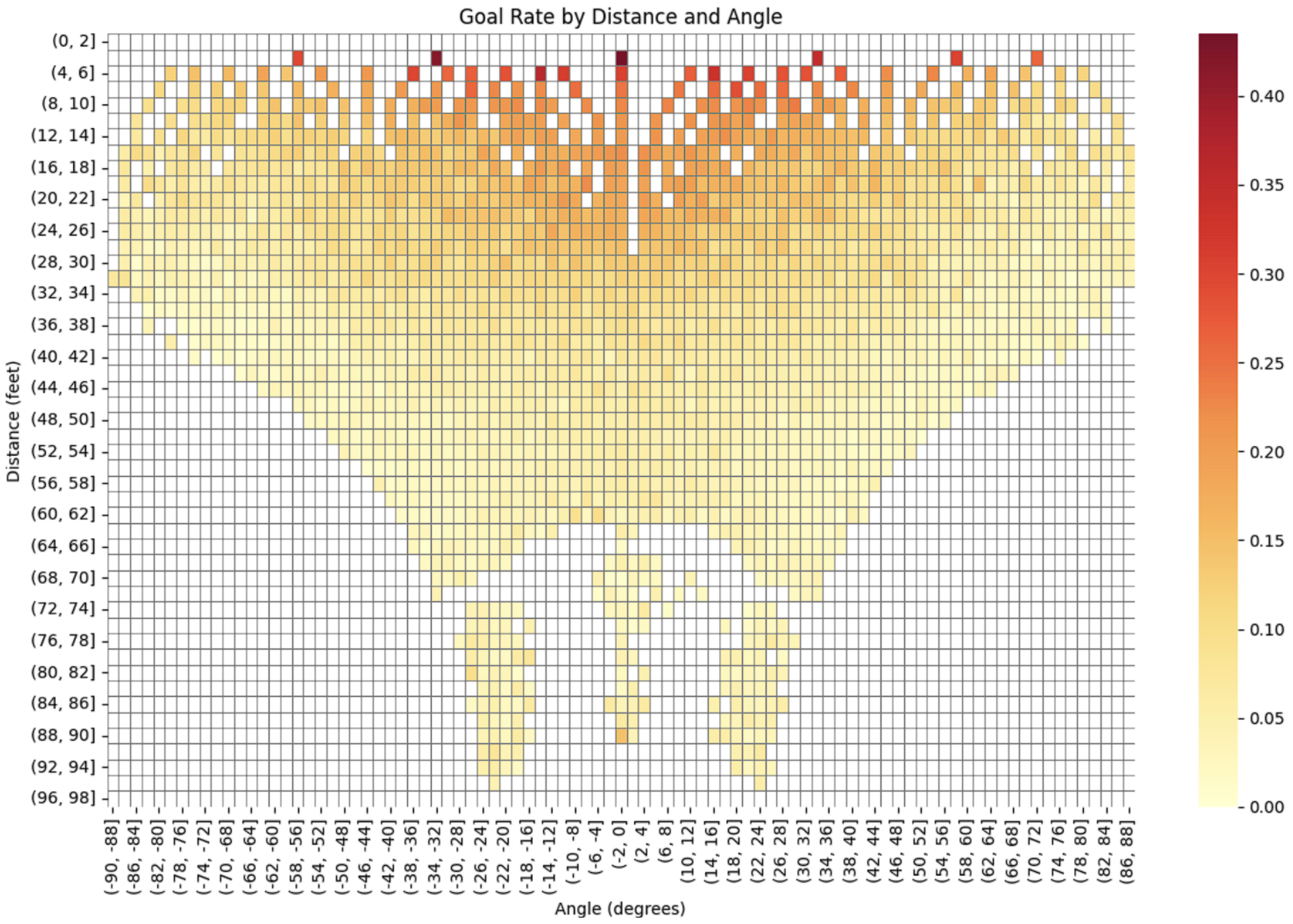


Figure 5: Heatmap of goal rate by shot distance and shot angle. Minimum of 100 shots

Data Analysis – Situation

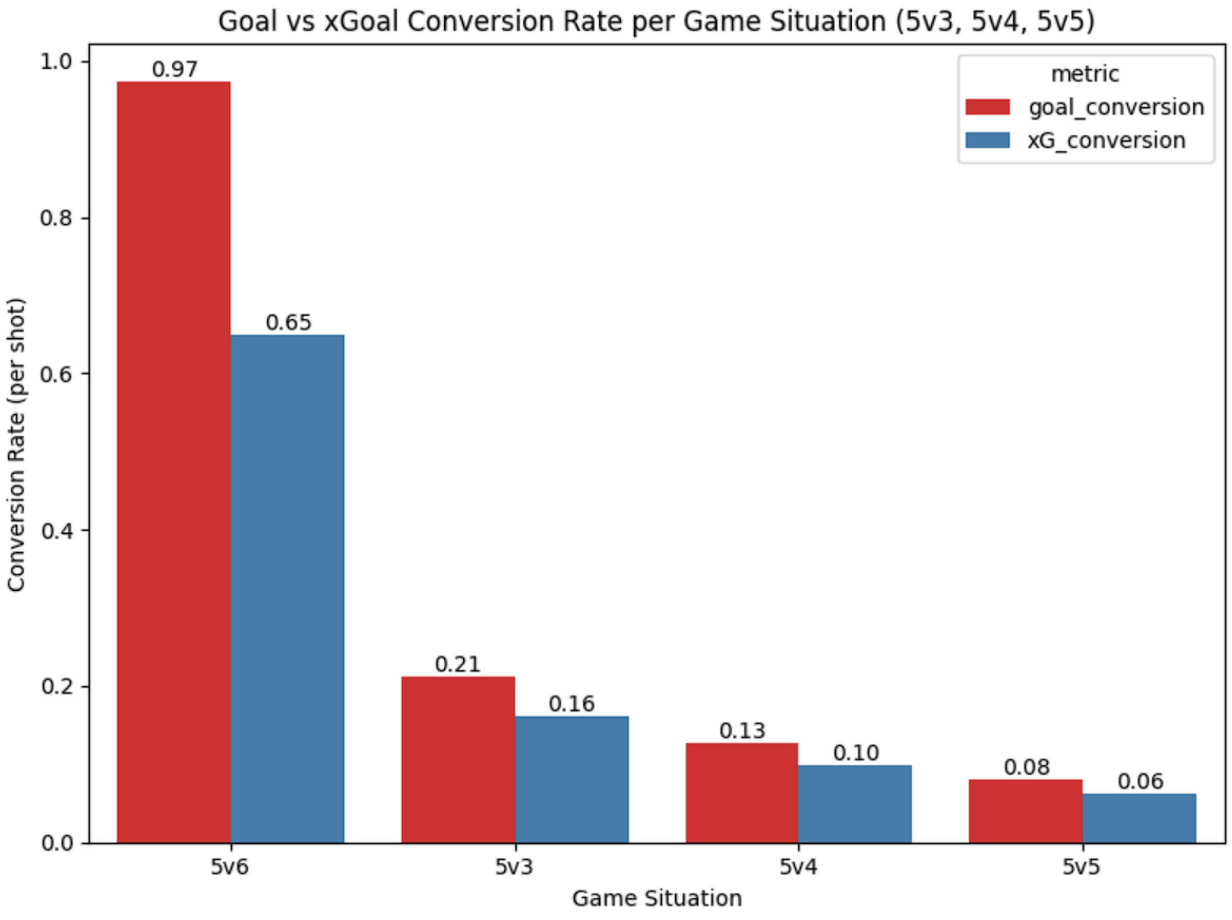


Figure 6: : Goal conversion and xGoal conversion per game situation

Data Analysis – Player Position

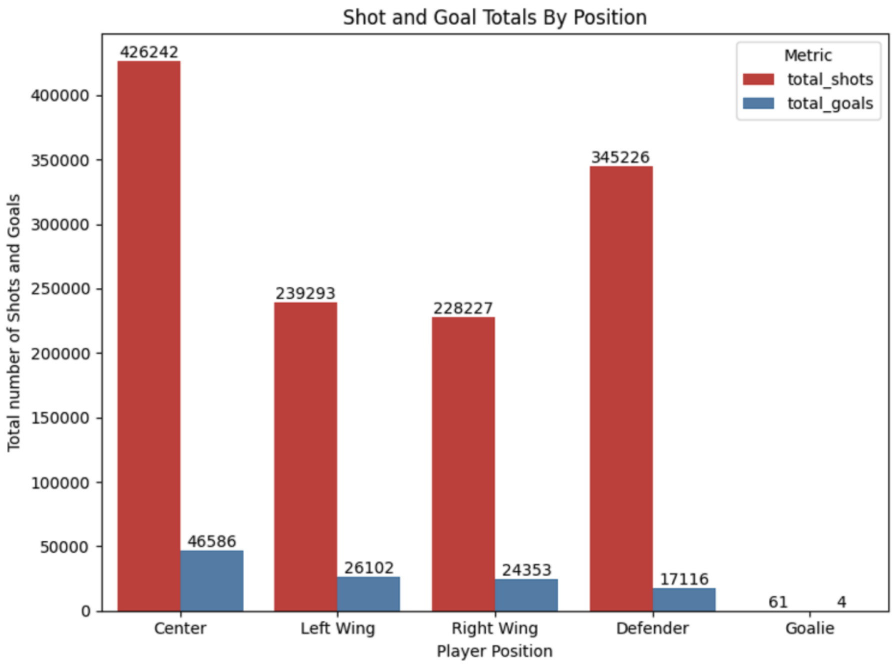


Figure 7: Total shots and total goals by player position

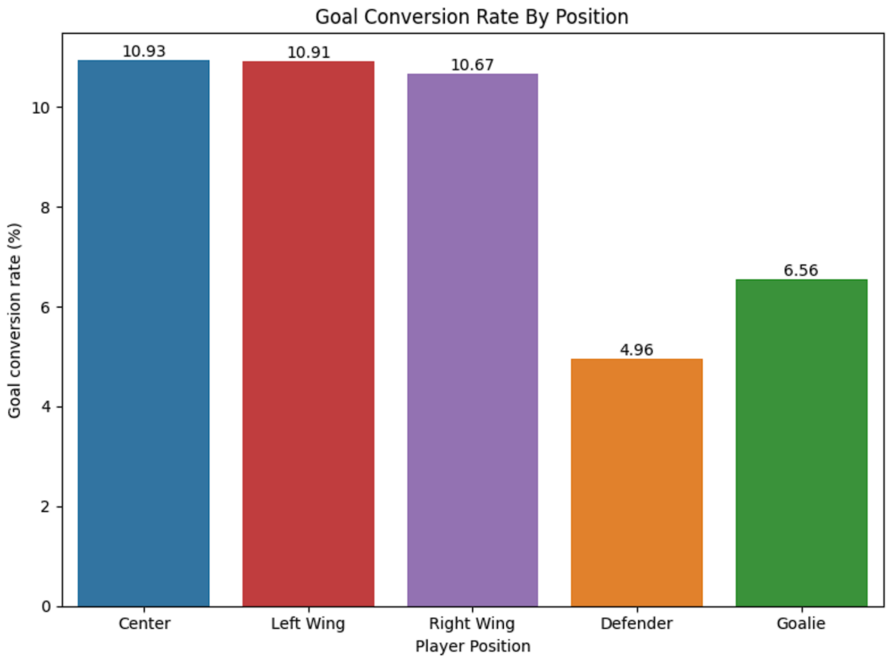


Figure 8: Goal conversion by player position

Data Analysis – Shooter Handedness

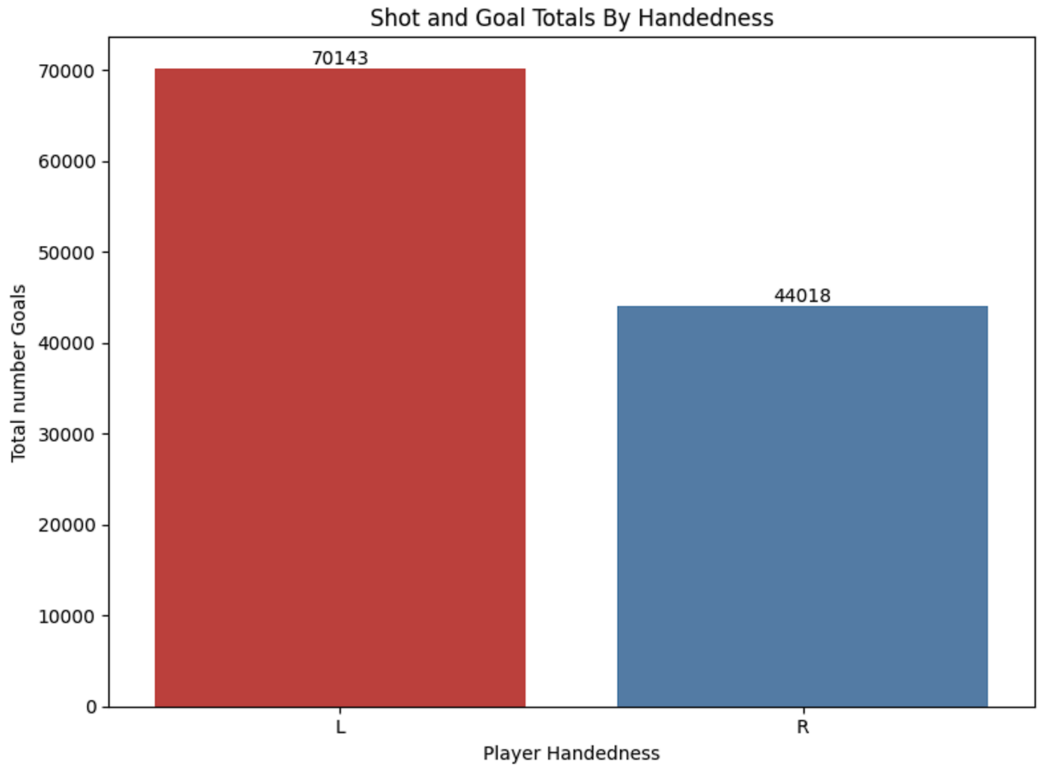


Figure 9: Goal totals by player handedness

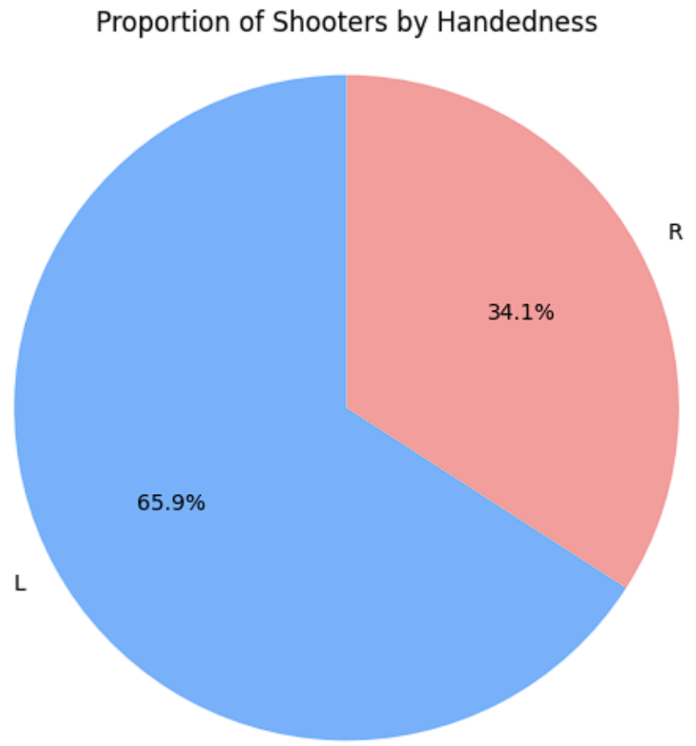


Figure 10: Proportion of Left-handed and Right-handed shooters

Data Analysis – Shooter Handedness

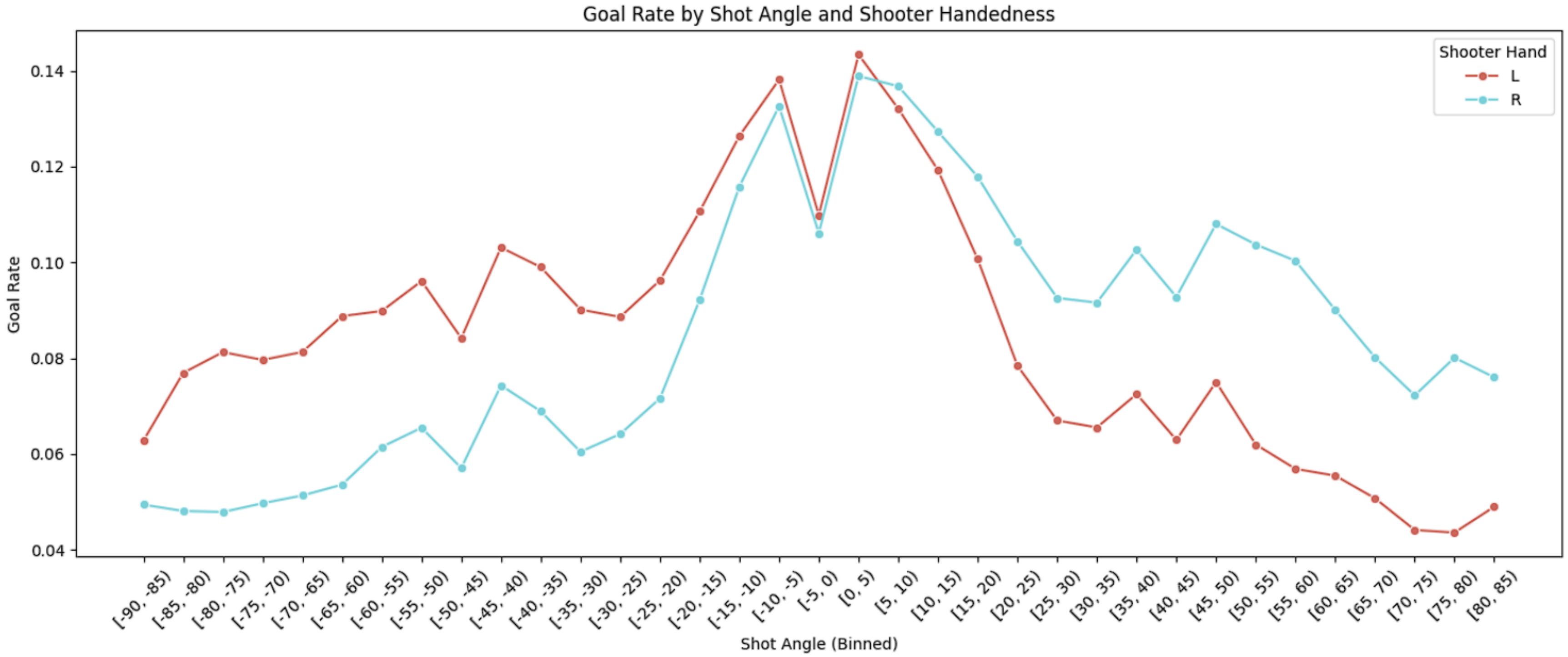


Figure 11: Goal rate across different shooting angles for right and left handed shooters

Data Analysis – Shot Rebound

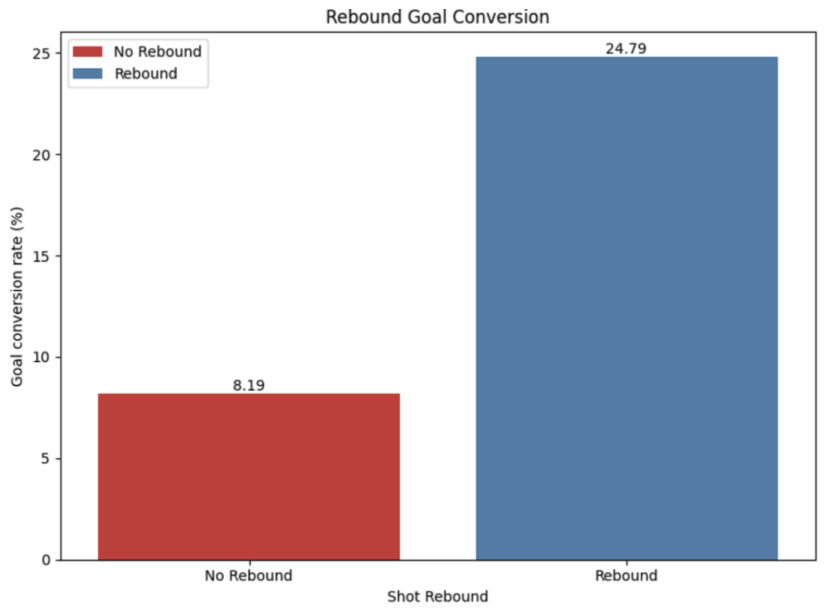


Figure 12: Goal conversion rate for rebounded and non-rebounded shots

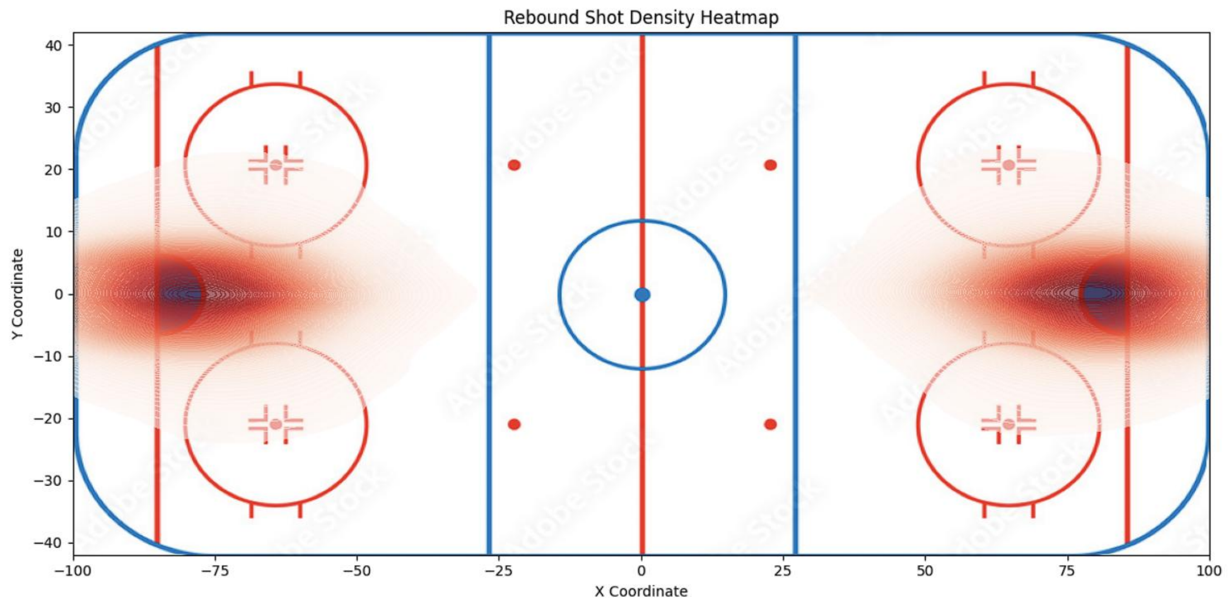


Figure 13: Rebound shot density locations on the ice

Modeling Approach

Model	Strengths	Why It's a Good Fit
Logistic Regression	- Simple & interpretable - Fast to train - Good baseline	- Clear understanding of how features (distance, angle) influence goal probability
Random Forest	- Handles non-linear patterns - Feature importance built-in - Robust	- Captures complex feature interactions (e.g., rebound × shot type) - Low risk of overfitting
XGBoost	- High accuracy - Built-in support for imbalance - Highly tunable	- Excels with tabular data - Best recall for identifying likely goals

Model Performance

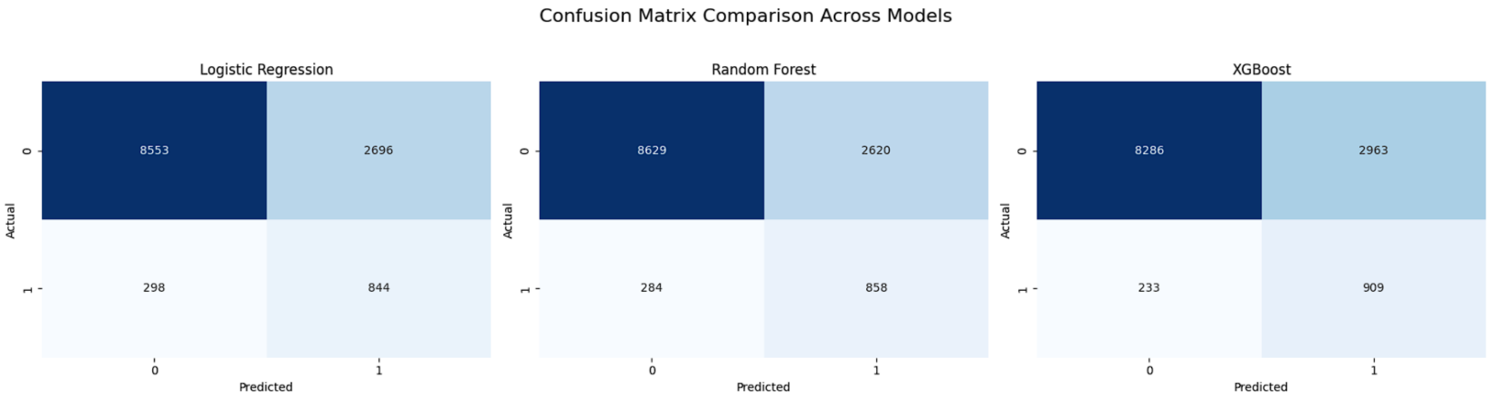


Figure 14: Confusion Matrix comparison of all 3 models.

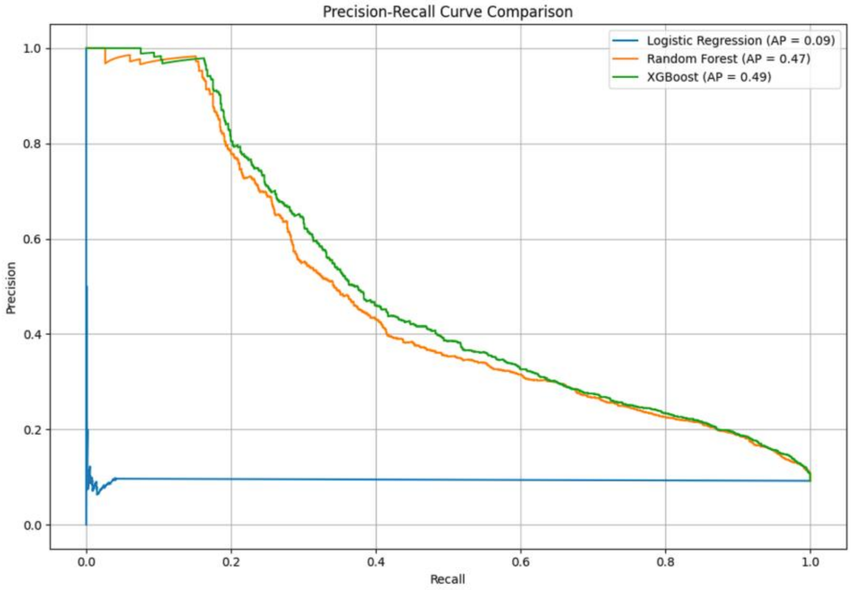


Figure 15: Precision-Recall Curve comparison of all 3 models.

Business Value

Player Value Forecasting.

Use the model to predict the expected number of goals for each client in the upcoming season.

Goals (especially for forwards) heavily influence contract negotiations, media value, and endorsement potential.

TB Sports can the expected number of goals for each client to informed decisions about player contracts, trades, and investments to optimize their clients careers.

Negotiation Leverage to Increase Client Contracts.

Compare a player’s predicted performance to similar players with higher salaries.

This information can help demonstrate underutilized or undervalued talent.

For example, based on model predictions player A is projected to outperform player B (who has a $5M contract), so increasing player A’s new contract from $3.8M to $5.5M is fair compensation for player A.

Client Acquisition

The model can help identify breakout candidates, players likely to outperform expectations.

This can help TB Sport sign emerging talent before market value explodes.

For example, the model can identify low-salary players with high xGoals and expected breakout performance and allow TB Sports to approach them before other agencies.

Performance-Based Endorsement Strategies

Use model projections to help structure endorsement deals based on expected goals.

Sponsors want measurable outcomes and using data from the model can help close endorsement deals and increase endorsement deal contracts.

For example, if the model projects player A to score 25+ goals. TB Sports can pitch endorsement deals with performance bonuses if player A meets or exceeds that goal total.

Injury Recovery & Performance Justification

The model can show that a player's xGoals or shot quality remained high even if their actual goals dipped (e.g., due to bad luck or post-injury recovery).

This will provide data-driven context to down seasons.

For example, if player A had a low goal count, but his expected goals remained high, this indicates a strong bounce-back next season.

Limitations & Future Work

Limitations

Class Imbalance	Skewed target classes impacting recall and precision
Complex Nonlinear Relationships	Interactions between features are nonlinear and complex, limiting simple models
Timeframe	Limited time constrained model experimentation
High Dimensionality	Large feature set leading to potential issues with noise and overfitting

Limitations & Future Work

Future Work

- **Neural Networks**: Explore deep learning models to capture complex, nonlinear feature interactions.
- **Advanced Ensembles**: Implement stacking or blending of multiple models to improve predictive power
- **Selective Sampling**: Use clustering or other unsupervised methods to sample more representative data for the majority class.
- **Incorporate New Data**: Add positional or tracking data to better capture game context and player movement

Conclusion

Key Insights

- ✅ Rebound shots are 3x more likely to result in goals
- ✅ Shot distance and angle are critical predictors
- ✅ XGBoost outperformed other models with best balance of recall and precision

Business Value	Informs smarter scouting decisions by identifying high-efficiency shooters
	Goes beyond raw goal counts to evaluate player potential
	Adds depth to contract and performance assessments
Future Potential	Incorporate puck/player tracking for richer context
	Explore neural networks and ensemble blending techniques
	Deploy as an API or interactive dashboard for stakeholder use