Capstone 2 Project Report

NHL Goal Scoring Prediction

Travis Bates

Problem Statement

What opportunities exist for T.B Sports to increase the average client contract by 25% for the 2025-2026 season using a machine learning model 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 recently seen a decline in clients and did not have a rise in average client contracts from the previous year. With management worried that the recent financial struggles will continue to grow, they want to turn to predictive analytics to help negotiate future contracts and guide future recruitment strategies.

Summary of Findings

Our primary goal was building a predictive model for predicting how many goals a player is expected to score in the upcoming season. Goals (especially for forwards) heavily influence contract negotiations, media value, and endorsement potential. It can provide negotiation leverage with teams by comparing a player's predicted performance to similar players with higher salaries. It can also be used to identify breakout players likely to outperform expectations or used to help structure endorsement deals based on the expected goals.

DATA WRANGLING:

The raw data sets contained 1.6 million and 122,000 rows, respectively, and 124 columns. The first data set contained shot data from 2007 to 2021, and the second set contained shot data from 2022. With 124 features, dimensionality reduction was critical. I reduced the original dataset from 124 features to 63 by removing redundant identifiers, low-variance features, and features that might introduce data leakage. I also performed domain-informed feature selection, retaining only features with clear relevance to goal-scoring prediction. Null values accounted for less than 1% of the entire data set and were dropped. Outliers were identified in the remaining features with impossible values, such as only having 2 skaters on the ice or impossible shot distances. These outliers were dropped accordingly.

The final shape of my data after data wrangling was 1,239,049 rows and 63 columns.

EXPLORATORY DATA ANALYSIS:

The goal of this EDA was to understand key distributions, correlations, and patterns in the data that influence whether a shot results in a goal to help inform feature selection and model strategy.

I started by examining the class imbalance of my target feature: goals. The goal scoring percentage in the NHL from 2007 to 2022 was 6.67%. Only 114,591 out of 1,717,746 total shots were goals. This is a large class imbalance that can lead to biased predictions and poor performance on the minority class. This will have to be addressed before training any models.

I was most interested in the following key features due to their potential impact on whether a shot is a goal.

- Shot type
- Distance from goal
- Shot angle
- Situation
- Player Position
- Shooter handiness
- Shot rebounds

Shot Type

Wrist shots were by far the most common shot type for all goals scored in the dataset. There were nearly 39,000 more wrist shots scored than the nearest shot type, the snapshot. Wrist shots also had one of the highest average xGoal probability, which is the likelihood of a shot being a goal. However, wrist shots were also the most common shot from all shots in the entire dataset. This is why we then investigated the conversion rate on each shot type. What is interesting is we see that wrist shots have lower conversion rates when investigating the goal and XGoal conversion rates per shot type. Wrist shots only have a 9% conversion rate compared to deflections which have a conversion rate around 19%. We also see that shot types with the highest average xGoal probability tended to have higher conversion rates.

This does make sense when applying some domain knowledge. Shot types like deflections, and tip-ins generally occur close to the net at favorable angles. Slap shots on the other hand tend to be taken further away and take longer, meaning the goalie has more time to react to the shot. TB Sports should look at the individual players' shot type preference. Certain players may favor specific shot types and score more efficiently with them.

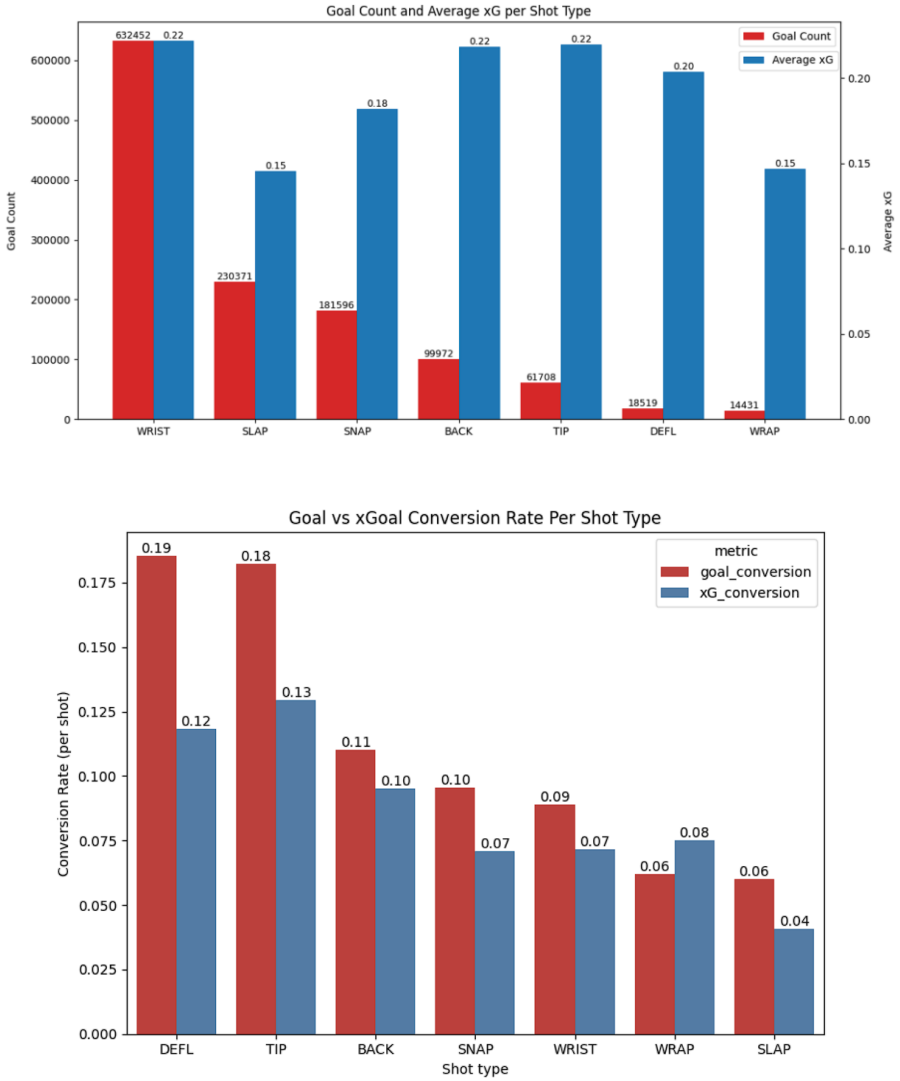

Figure1: Goal count and average xGoal by shot type (top), goal conversion and xGoal conversion percentage rates by shot type (bottom)

Distance from goal

As expected, shots closer to the goal had higher xGoal probability and resulted in higher goal rates. Most goals occurred within 32 feet or less of the net with an average distance of 23 feet. This means players can generate higher-quality scoring opportunities if more shots are taken from close range. As a sports agency looking for players who can generate high-quality scoring opportunities, TB Sports could look at players' whose tendencies lean towards more shots at close range.

Shot Angle

Shot angle is the shooting angle relative to the net. 0° is straight in front of the net, 90° is from the side boards. Similar to distance from goal, most goals are scored from tight angles close to 0°. The average shot angle of all goals in the data set was -1.5° with most goals being scored from a range of -24°and 22°.

Intuitively, one would assume that shot angle and shot distance would have a moderate to strong correlation with each other. Meaning shots that are closer to the goal and closer to an angle of 0° would produce higher goal rates. The heatmap in figure 2 explores this assumption by showing the goal rate by shot distance and shot angle. Each bin consists of 100 or more shots.

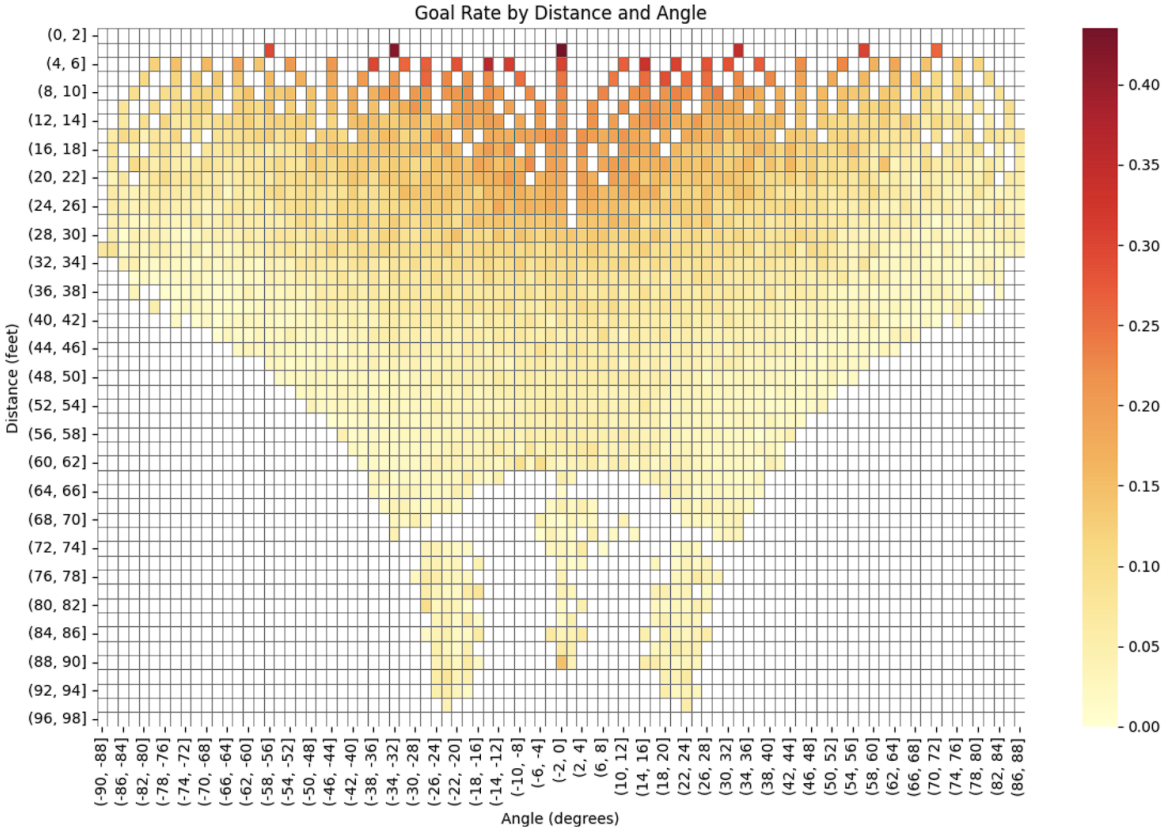

Figure 2: Heatmap of goal rate by shot distance and shot angle. Minimum of 100 shots

From the heatmap above we see our assumption seems to hold true. Shots that are closer to the goal and at angles close to 0° had higher goal rates than shots that are from further distances and wider angles.

TB sports should consider player position when looking at shot distance and shot angle data. Positions like defensemen will likely be taking shots from a further distance and poor angels due to their position on the ice.

Situation

The total percentage of all goals that were scored on the power play was 16.92 %. The standard power play, which consists of one team having 4 players on the ice and one team having 5 players on the ice (4v5), was not the only situation that was investigated. Other situations including 3v5 and 5v6 were also explored. Goal conversion rates, as expected, were the highest in 5v6 situations because there was not goalie in the net when the shot was taken. This situation had a 97.3% goal conversion rate. The next highest goal conversion rate was in the 3v5 situation when one team has a 2-player advantage on the ice. This had a goal conversion rate of 21.3% where the standard powerplay had a goal conversion rate of 12.6%. All these situations had higher goal conversion rates than even 5v5 situations which had a goal conversion rate of only 8.1%. The xGoal probability conversion rate for each situation followed the same pattern as goal conversion.

This shows that players with high shot volume on power plays may have inflated goal totals and should be considered when evaluating new players to add to TB Sports agency. It is also important to consider how team strategies differ; some generate more quality chances even at even strength.

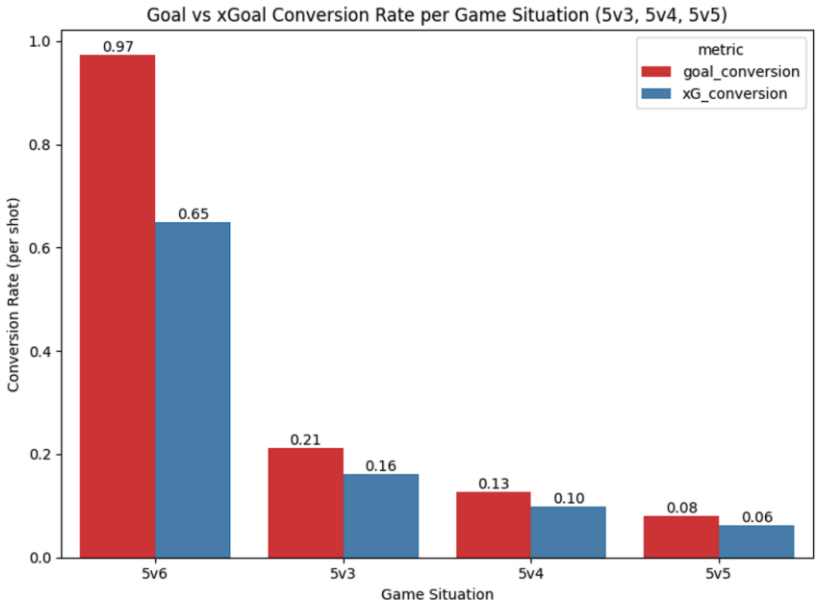

Figure 3: Goal conversion and xGoal conversion rates based on situation.

Player Position

At first glance, it would appear that a player who plays the position of Center would have a significant advantage when it comes to scoring goals. Centers outscored the nearest position, Left Wing, by 20,000 goals. However, when digging deeper and inspecting the actual goal conversion rate compared to the overall shots taken, there was no advantage between all 3 offensive positions. Centers, Left Wing, and Right-Wing positions had a goal conversion rate of 10.93%, 10.91%, and 10.67% respectively. As expected, Defenseman had the lowest conversion rate of only 5%. This is expected as their primary responsibility is not goal scoring by defending. They naturally will take less shots and have lower goal scoring abilities.

TB Sports should not compare players across different positions as there is not a significant advantage. Rather they should compare players within the same position to get a baseline for contract negotiations and scouting new clients.

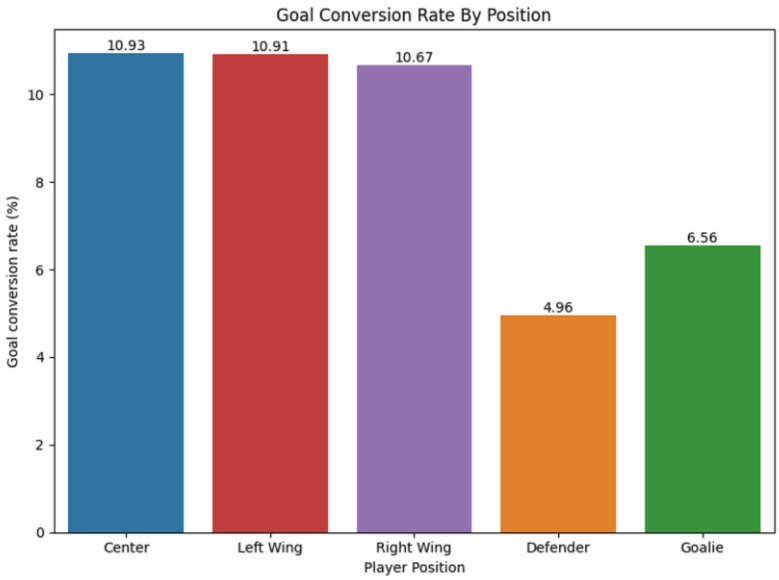

Figure 4: Goal conversion rates based on player position

Shooter Handedness

When evaluating shooter handedness among all goals in the dataset, left-handed shooters scored 1.6 times as many goals as right-handed shooters. When investigating this trend deeper, it was found that there are about 1.9 times as many left-handed shooters in the NHL than there are right-handed shooters. This imbalance alone could possibly explain why we see more goals scored from left-handed shooters.

Goalie handiness was also investigated and about 92% of goals catch with their left hand. This suggests that goalie handedness does not play a large role, if any, in the success of a goal regardless of the shooter's handedness.

Next, shooter handedness was compared with shot angle to determine if the shooter had any advantages at certain angles. As expected, left-handed shooters were able to convert more goals on the right side of the ice (negative angles) and right-handed shooters were able to convert more goals on the left side of the ice (positive angles).

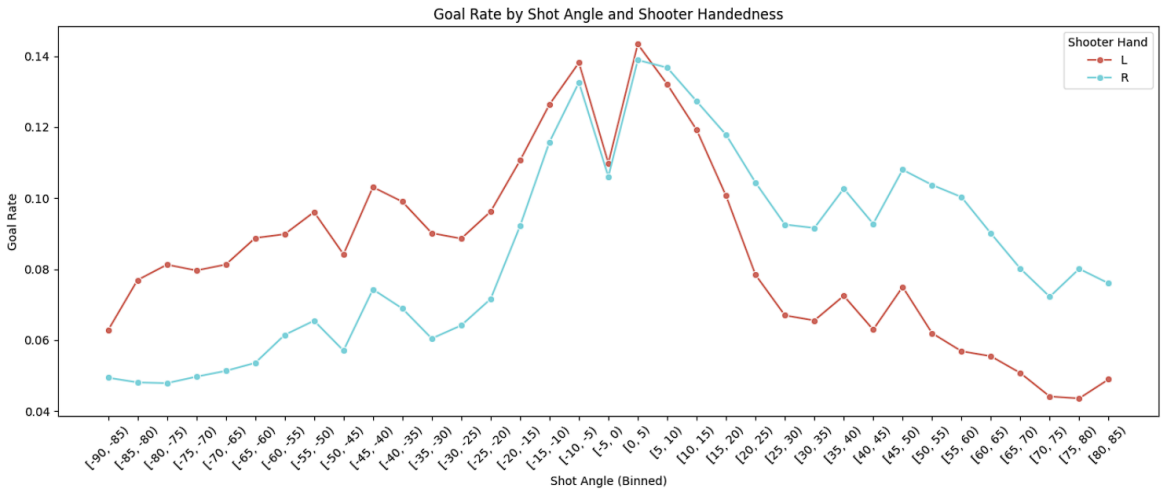

Figure 5: Goal rate by shooter handiness and shot angle.

This suggests that the players' handedness, alone, does not play a significant role in determining if a shot becomes a goal. The players position and shooting angle are more of a determining factor when combined with shooter handedness.. TB Sports would need to consider a players position and their shot location tendencies when evaluating new players rather than heavily relying on just player handedness alone.

Shot Rebond

Rebond shots are often more dangerous due to their proximity to the net and the goaltender being out of position. The data supports this claim as well. The average xGoal probability on all rebound goals was 35% with non-rebound goals only having an average xGoal probability of 18%. The goal conversion rate among rebound shots was also much higher than non-rebound shots: 25% compared to 8%. This suggests that a shot being a rebound does have a higher chance of being a goal.

TB Sports could evaluate players that may have a knack for generating or finishing rebounds. Forwards who position themselves well near the net may have more rebound goals.

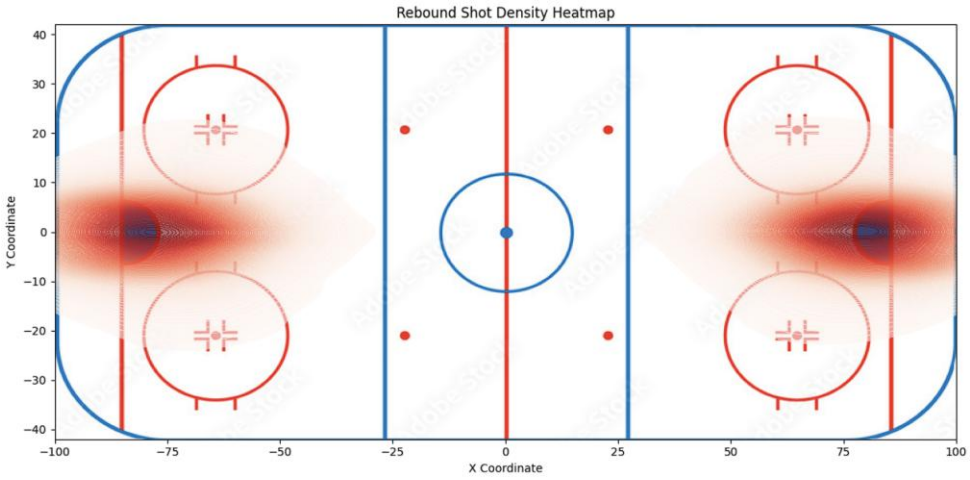

Figure 6: Heatmap of rebound shot density.

PREPROCESSING AND TRAINING:

The goal was to build a classification model that predicts whether a shot results in a goal using shot-level features from NHL players.

- Feature Selection

  Based on the exploratory data analysis on the data and applying additional domain knowledge, I reduced the dataset further from 63 features down to 47. I eliminated additional redundant features such as player names since the data set included player IDs and names would not be useful for training the model. Other features like non-adjusted coordinates were also dropped due to their potential to complicate training of the model.

- Splitting Data

  To help prevent data leakage, I first split the data after the final feature reduction and before feature encoding. Because it was a very large data set, a 98/1/1 split was used. This created a training set using 98% of the data, a test spit using 1% of the data and a validation split using 1% of the data.

- Encoding Categorical Features

  The data had several different types of categorical data that required different encoding methods. I used 3 different methods based on the categories.

  - One-hot encoding for any categorical features with only 2 levels
  - Target encoding for categories believed to have a correlation to our target. I wanted to help capture their relationship to the target.
  - Frequency encoding for the remaining columns with more than 2 levels whose relationship to the target was not as important.

- Standardization And Normalization

  Since I planned on testing a logistic regression model, I also needed to perform feature scaling. I started by applying a log transformation to all the skewed features to help with normalization of all skewed features in my data.

  After all features were normalized, I begin with the following scaling techniques:

  - Standard scaler to standardize the unbounded (non-probability based) features
  - min_max scaler on the bounded (probability based) features.

MODEL SECLECTION:

I evaluated the models using the following metrics:

- **Precision**: How many predicted goals were actually goals.
- **Recall**: How many actual goals the model correctly identified.
- **F1 Score**: Harmonic mean of precision and recall — the key metric for this use case.
- **ROC-AUC**: Overall discriminative power of the model.

Given the project's goal to identify scoring opportunities, Recall and F1 score were prioritized over accuracy. This was important because when evaluating player talent, recall reflects the model's ability to correctly identify high-value players who will score goals. TB Sports' goal is to increase number of clients and client contracts, so they must target high-value players.

We trained and tested the following algorithms:

Model	**Description**
Logistic Regression	Simple, interpretable baseline
Random Forest	Ensemble-based decision tree model
XGBoost	Gradient boosting model tuned for class imbalance

Each model also went through hyperparameter tuning. Hyperparameters were tuned using GridSearchCV and RandomizedSearchCV with cross-validation on the training set.

- XGBoost Parameters Tuned: n_estimators, learning_rate, max_depth, learning_rate, scale_pos_weight, gamma, alpha, lambda, min_child_weight
- Random Forest Parameters Tuned: n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features
- Logistic Regression Parameters: C, penalty, solver

Recursive feature reduction was also performed on all models to improve model performance. However, there was no noticeable improvement in any of the models after the feature reduction.

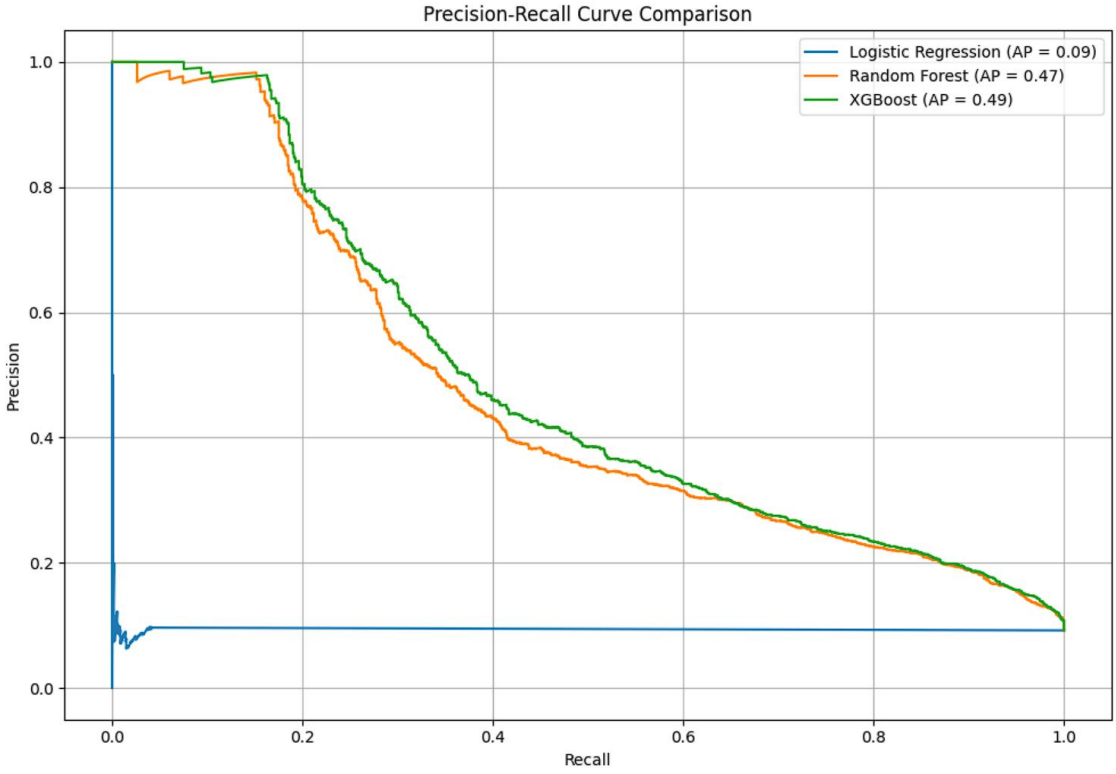

Figure 7: Precision-Recall Curve comparison between models

XGBoost outperformed other models in terms of Recall and ROC-AUC. However, all models still struggled with precision and F1 Score. This was expected in our highly imbalanced dataset and was slightly compensated for by strong recall. However, the greatest challenges with this dataset were the large class imbalance and the complexity of relationships between features when predicting goal scoring.

RESULTS AND RECOMMENDATIONS

I was not able to achieve the desired results during this project within the time constraints. The final model had a recall of 0.80 but precision of 0.23 and F1 Score of 0.36 for the target variable, goal. Since the models were trained and tested on shot data from 2007 to 2022, I ran predictions from the best performing model and compared the predictions to the actual goals scored in the 2023 shot data. The average accuracy for all goal predictions when compared to the 2023 shot data was: 19.59%.

Even though I was not able to achieve the desired model metrics there are still several valuable insights to be drawn from the data and model once improved.

Recommendations:

1. 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.

2. 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.

3. 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.

4. Performance-Based Endorsement Strategies

- TB Sports should 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.

5. 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.

FUTURE SCOPE OF WORK

There are several approaches I can take to improve model performance. Below are different approaches to help improve model performance:

- Advanced Ensembles
  - Implement stacking or blending of multiple models to improve predictive power
- Neural Networks
  - Explore deep learning models to capture complex, nonlinear feature interactions.
- 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

Once model has been improved, it can be expanded to include other statistics like assist, hits, penalty minutes, ect. However, if predicting goals was the only statistic we wanted, we could then deploy an improved model.

To deploy the goal prediction model so a sports agency could use it in real-time player evaluations, I would follow a streamlined process using accessible tools like Flask for the API and a dashboard for end users.

- First, I would serialize the final trained model using a library like joblib or pickle. This allows the model to be saved and loaded for inference without retraining.
- Next, I would build a RESTful API using Flask or FastAPI. This API would accept JSON input with shot-level features (e.g., distance, angle, shot type), process the data, and return the predicted goal probability.

- The API could feed predictions into a front-end dashboard built in Streamlit or Plotly Dash. This would let a sports agent select a player and view predicted goals, conversion rates, or compare players visually.
- To make the model accessible, I'd deploy it to a cloud platform like Heroku, AWS, or Azure. This would let users send predictions from anywhere, such as a team's scouting portal or agency CRM system.