Modeling Expected Goals (xG)

James Benasuli

Outline

- Business Problem
- Data
- Methods
- Results
- Conclusions

Business Problem

- Many NHL teams are still resistant to using advanced analytics despite the quality and quantity of game data greatly improving in recent years
- Reliance on older, less advanced metrics such as simply counting shot volume puts coaching and front office staffs at a disadvantage
- Our firm, Hockey Data LLC, has been contracted by an NHL team to build them an xG model
- Our xG model provides insights into individual and team performance allowing stakeholders to better understand the game and make better decisions

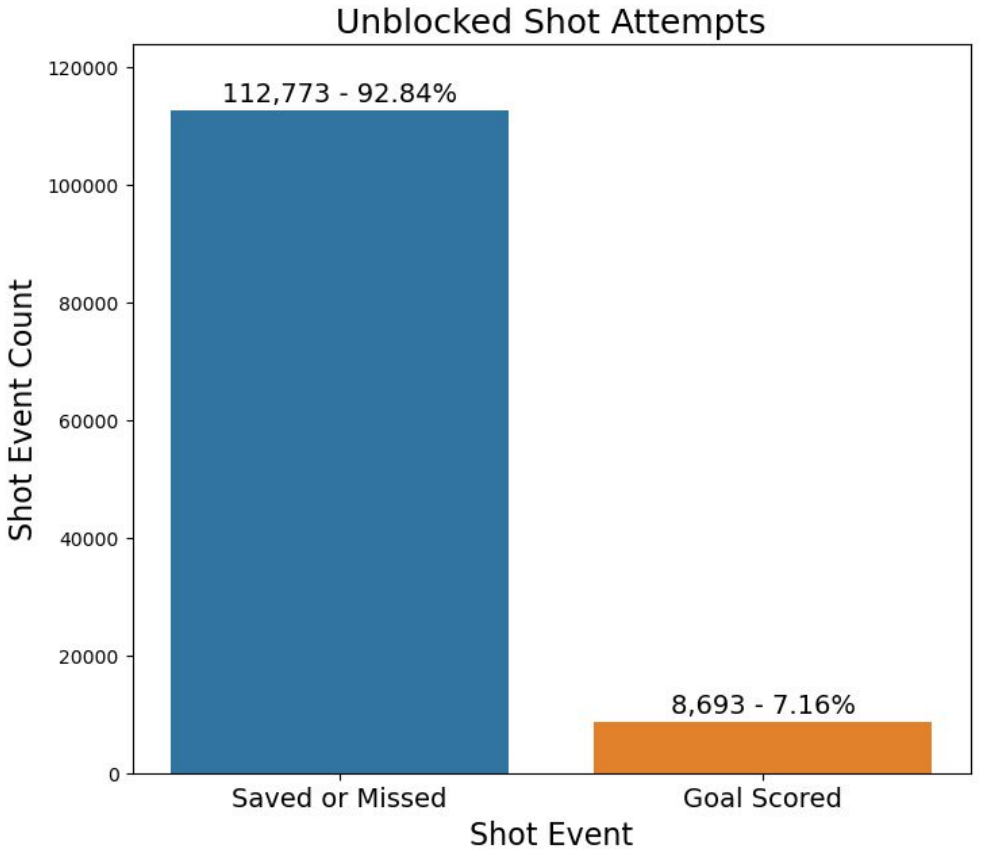
Data

- Individual shot data from the 2021-2022 NHL season
- Shot data was sourced from moneypuck.com
  - Moneypuck shot data is scraped from the ESPN and NHL websites
- Target variable is Goal (whether a goal was scored on a given shot)
- Features include shot type, shot location, game strength state (i.e. is the shooting team on a powerplay)

Why is understanding xG so important?

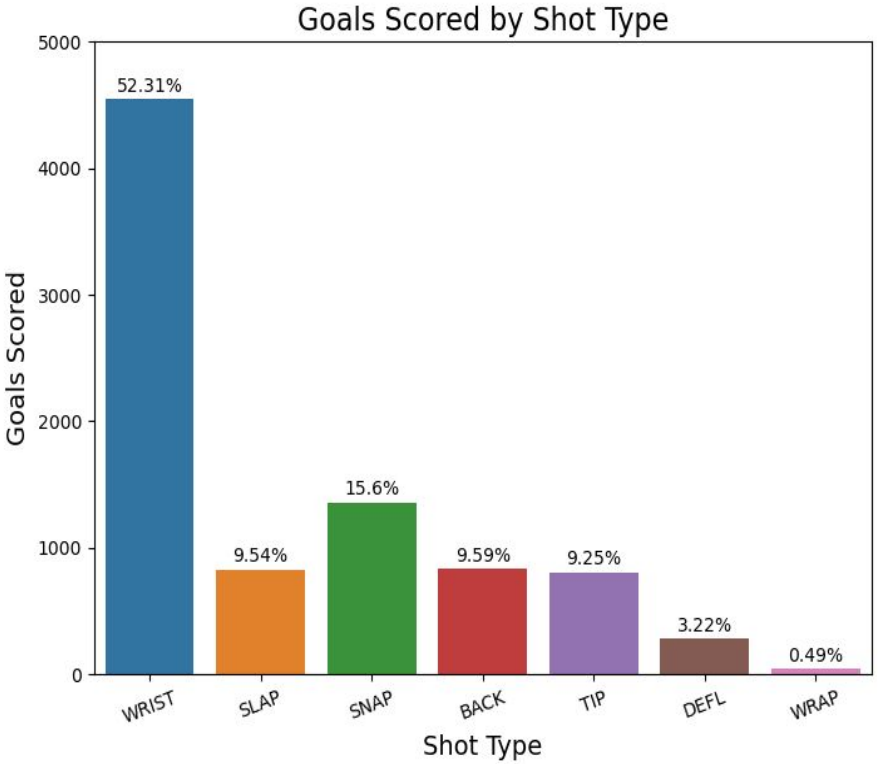
Most shots don’t go in

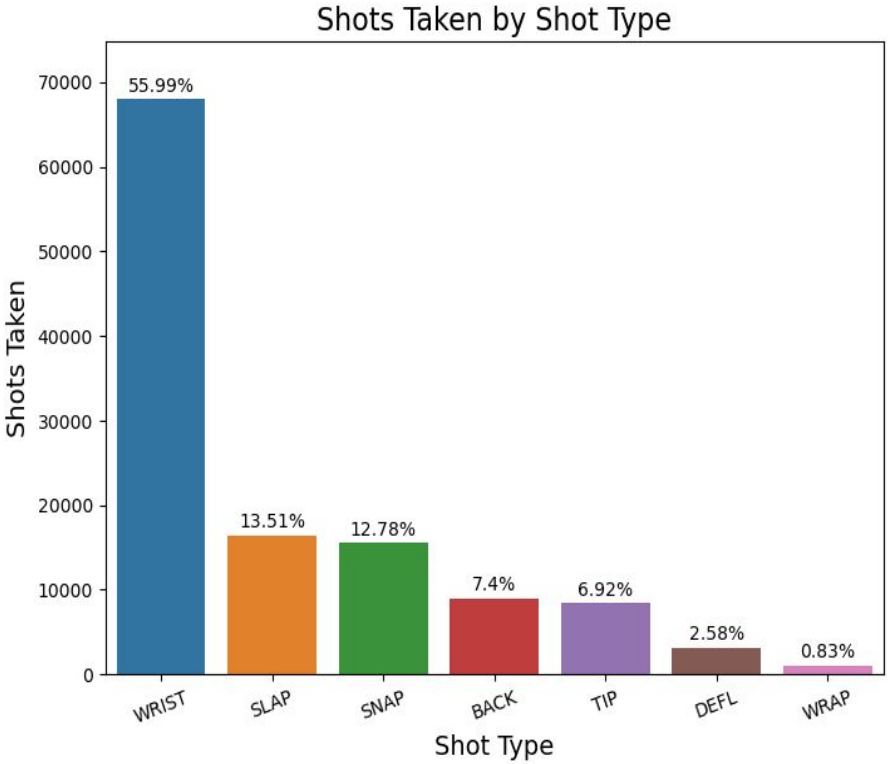
- 112,773 of 121,466 shots taken last season did not result in a goal
- Or only 7.16% of shots taken last season scored



Why is understanding xG so important?

And features such as shot type play an important role





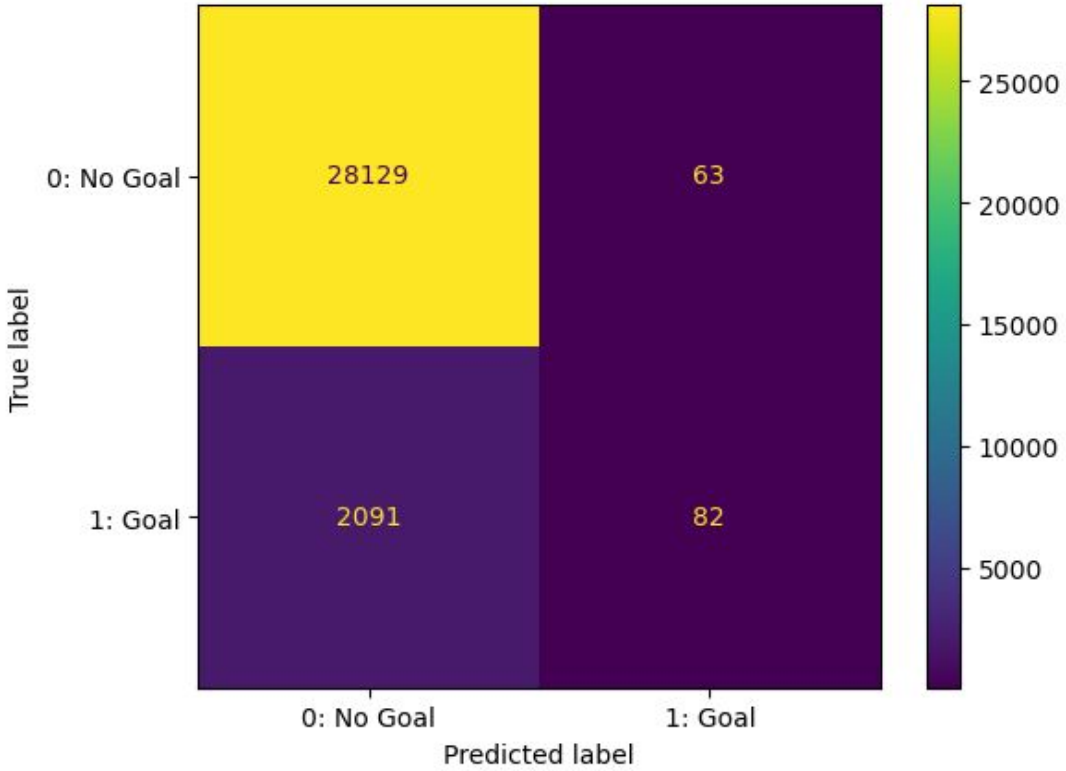
Methods

- Of the 124 features in the original dataset, 18 relating to location and game situation were selected for inclusion in our models
- This problem is treated as binary classification as we want to know the probability of any given shot to go in (1) or not (0)
- The # of players for each team on the ice at the time of a shot was used to create a game state strength metric
- Pipelines were built to handle encoding of categorical variables and scale numerical features
- Pipelines were processed using column transformers before being passed to logistic regression and decision tree models

Baseline Logistic Regression Model Results

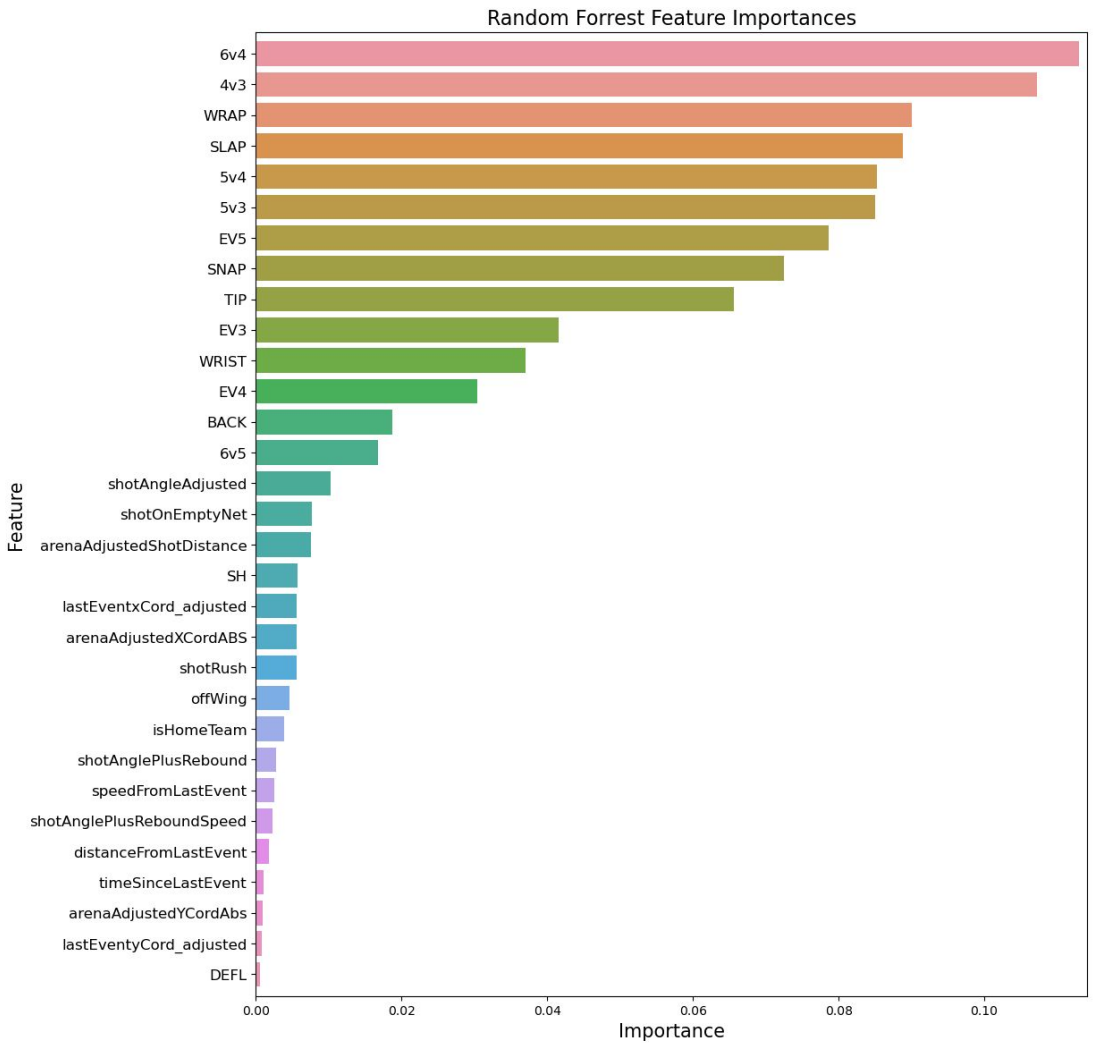
Out of the box logistic regression

- Test accuracy: 0.93
- Test AUC-ROC score: 0.756
- Test log loss score: 0.228



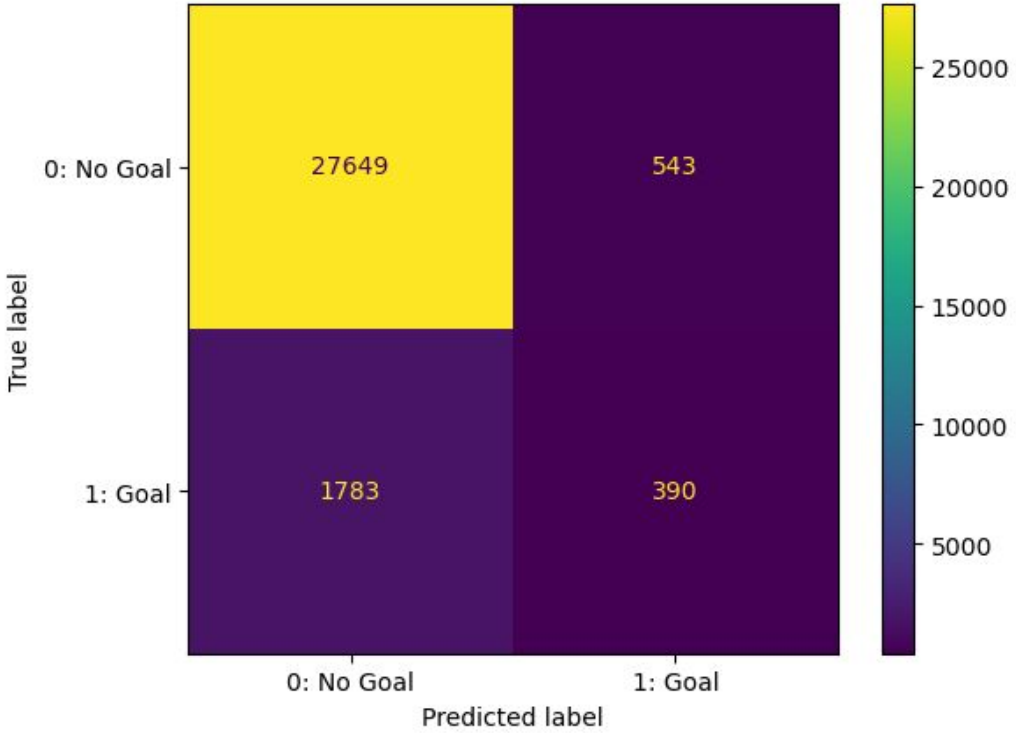
Baseline ExtraTreesClassifier

- Similar scores as baseline logistic regression
- Game state is dominating feature importance list



Random Forest with SMOTE

- Using SMOTE to oversample the minority class improved our ability to correctly predict goals
- However we traded a lot of false negatives for false positives
- Log loss jumped to .3

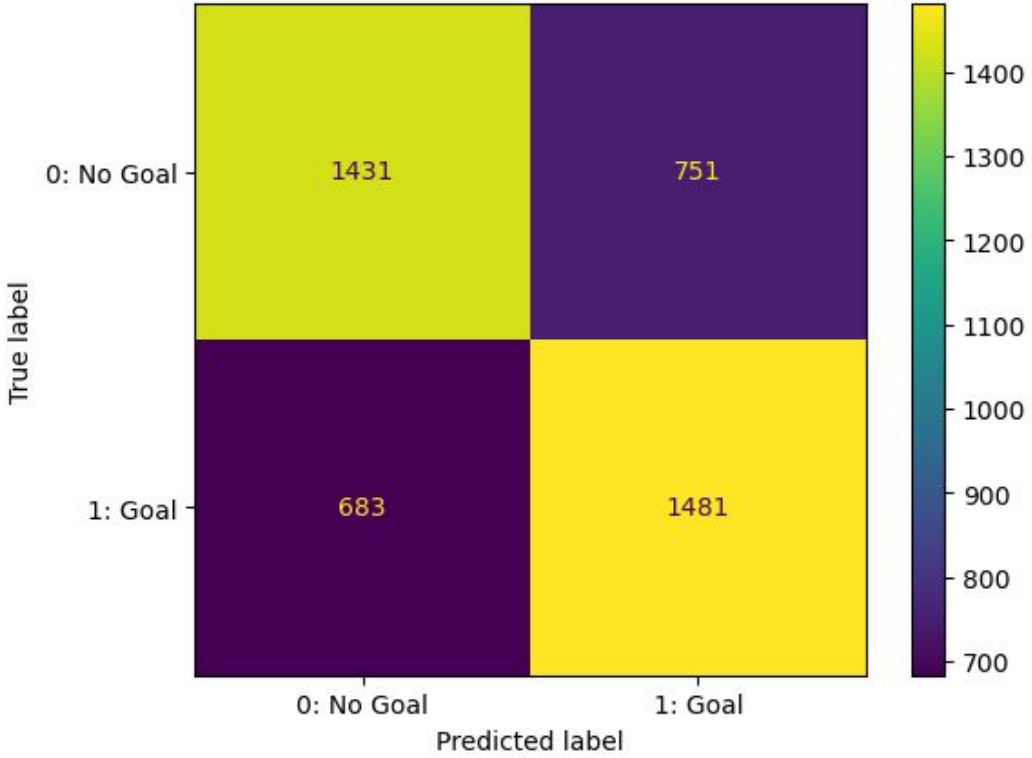


Random Forest using a subset of the majority class

Took a subset of ‘No Goal’ shots = to the number of shots which scored a goal

- Test accuracy: 0.67
- Test AUC-ROC score: 0.737
- Test log loss score: 0.601

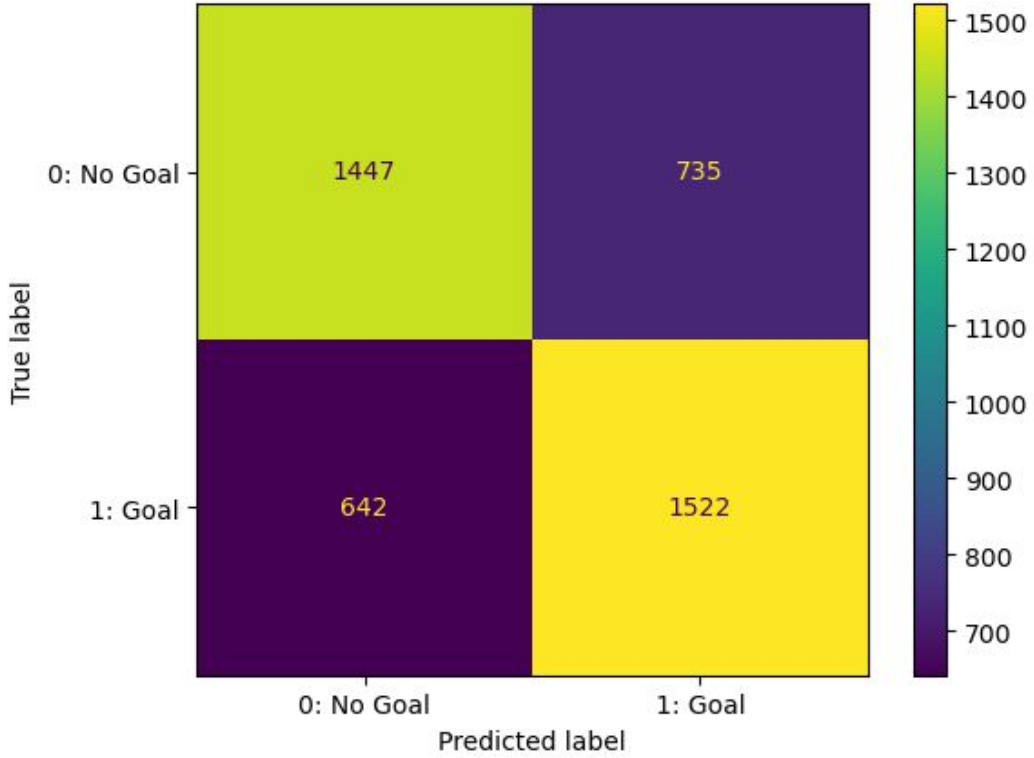
While our ability to correctly classify true positive outcomes significantly improved, log loss doubled



Logistic Regression Cross Validation on the subset

Marginal improvements over previous attempt with Random Forest

- Test accuracy: 0.68
- Test AUC-ROC score: 0.756
- Test log loss score: 0.586



Conclusions

- Modeling publicly available provides actionable insight for coaches and players, but there is room for improvement
- More work needs to be done to reduce the number of false negatives
- While the final model produced some good scores, our log loss jumped significantly
- Next steps:
  - Revisit feature selection
  - Construct separate models for different game strength classes
    - Different strengths dominates feature importance, but majority of game is played even strength
    - Feature importance might differ between states
  - Use team proprietary data
    - Introduce other features public data does not account for
    - Proprietary data could well be better than league data and provide an edge