NHL Game Prediction Modeling

Mark Shumka
November 10, 2023

Overview

Can we use data science to analyze past results of National Hockey League games and make accurate predictions of future game outcomes?

- Stakeholders would pay $millions for such a solution:
  - NHL teams
  - Sportsbooks

Dataset and Preprocessing

- Merged team-based Basic Stats with Advanced Stats
  - Supplemented with data scraped from the NHL API
- Performed feature engineering to convert raw data into comparative variables
  - E.g., Shots -> Share of Shots, Game Date -> Days Since Last Game

Initial Feature Buckets

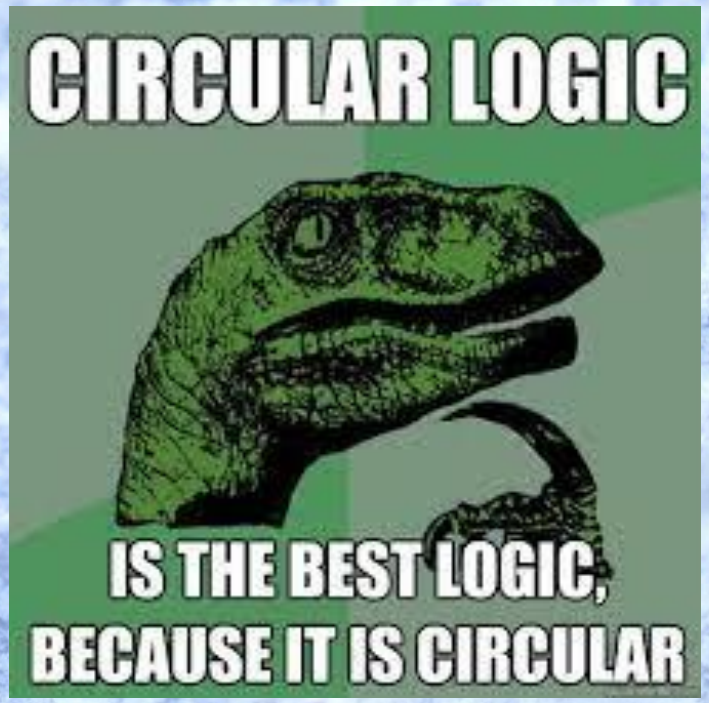
Goals	Shots	Expected Goals	Other
Goals For *# Goals Against*	Corsi Share of Shots Shooting Percentage Save Percentage High Danger Shots Share High Danger Shots Ratio	xGoals Percentage Shot Efficiency Defensive Efficiency	Home or Away Days Since Last Game Distance Since Last Game Share of Hits Share of Blocks Faceoff Percentage Power Play Percentage Penalty Kill Percentage Share of Giveaways Share of Takeaways Takeaway: Giveaway Ratio

Initial Modeling Accuracy

BUT WAIT!!

LOGICAL FALLACY ALERT

- I am trying to predict the outcome of an event using events that are occurring during the event that I am trying to predict



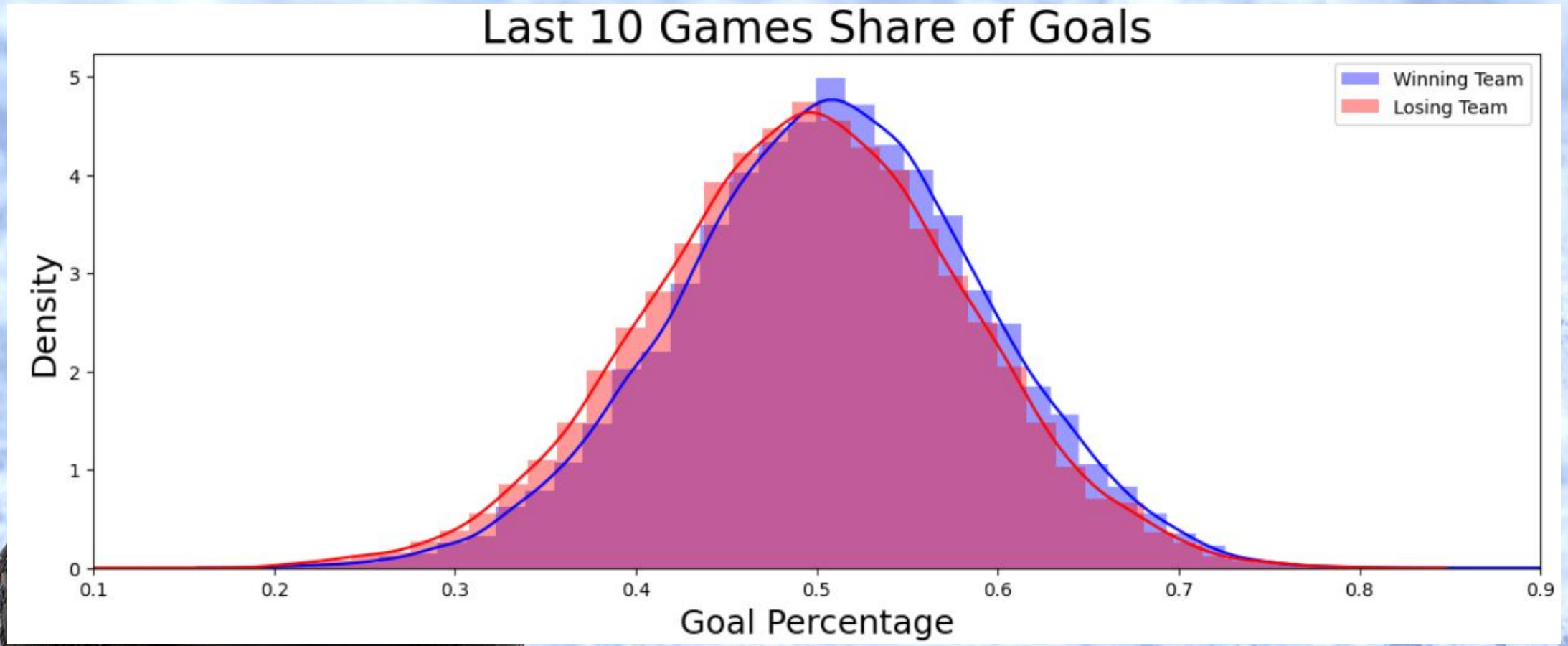
Where do we go from here?

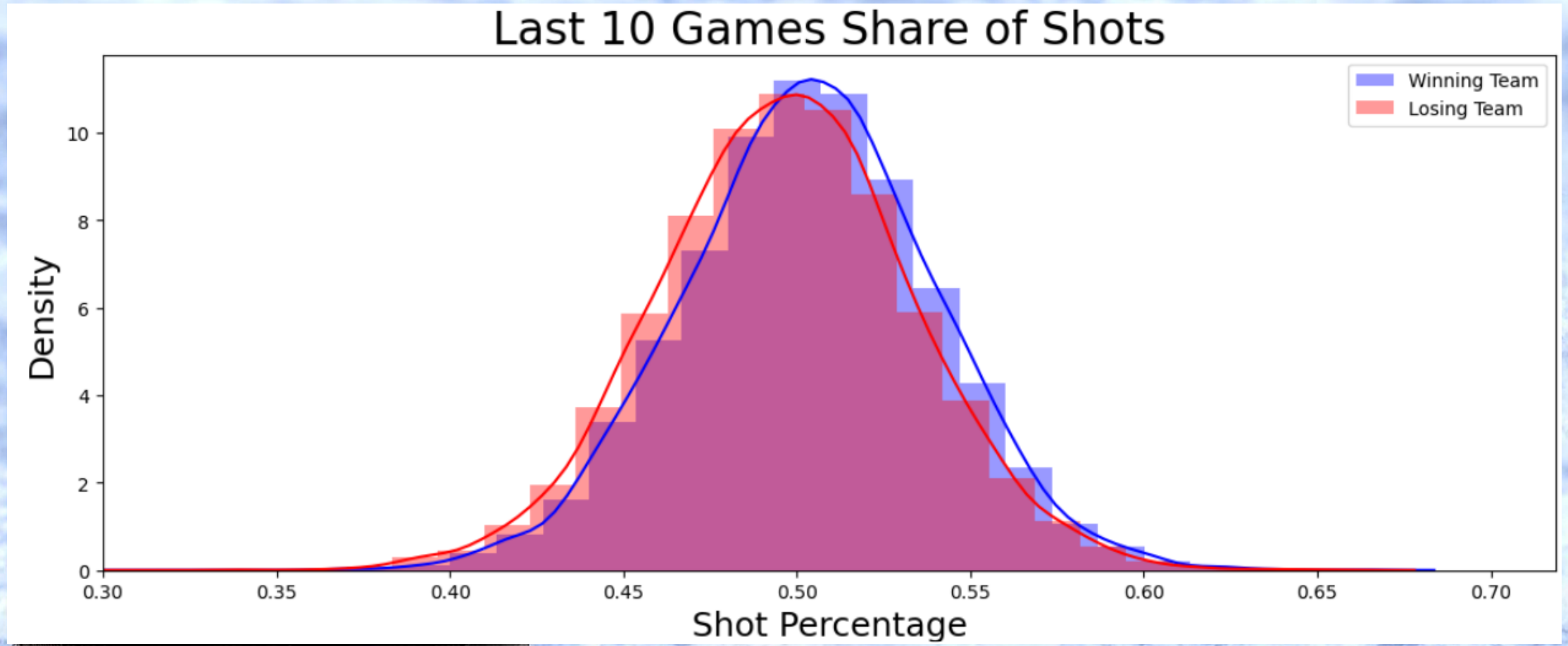
MORE Feature Engineering

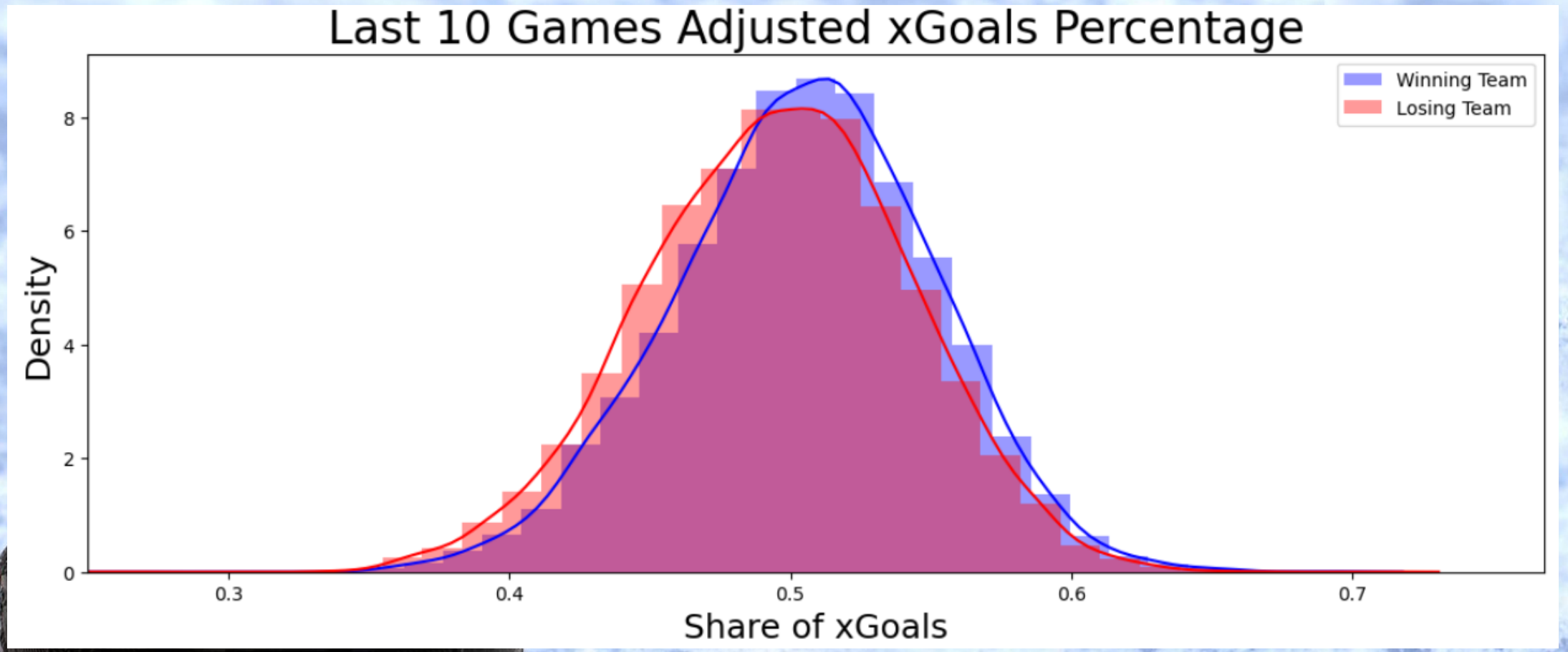
- Created additional features looking at recent historical performance of the teams playing each game
  - Trailing 10 games
- Open questions
  - Number of games?
  - Weighting

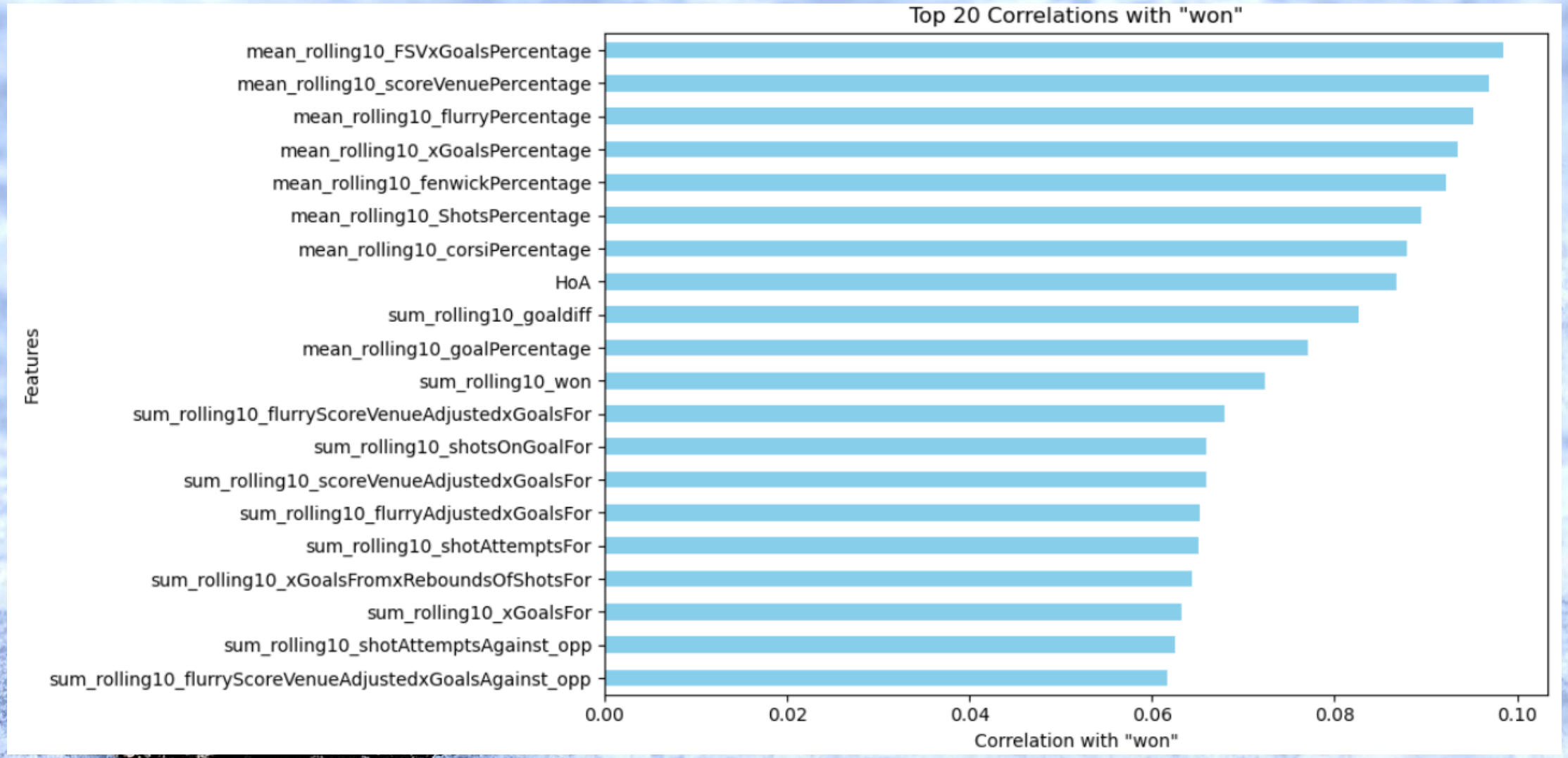
But first:

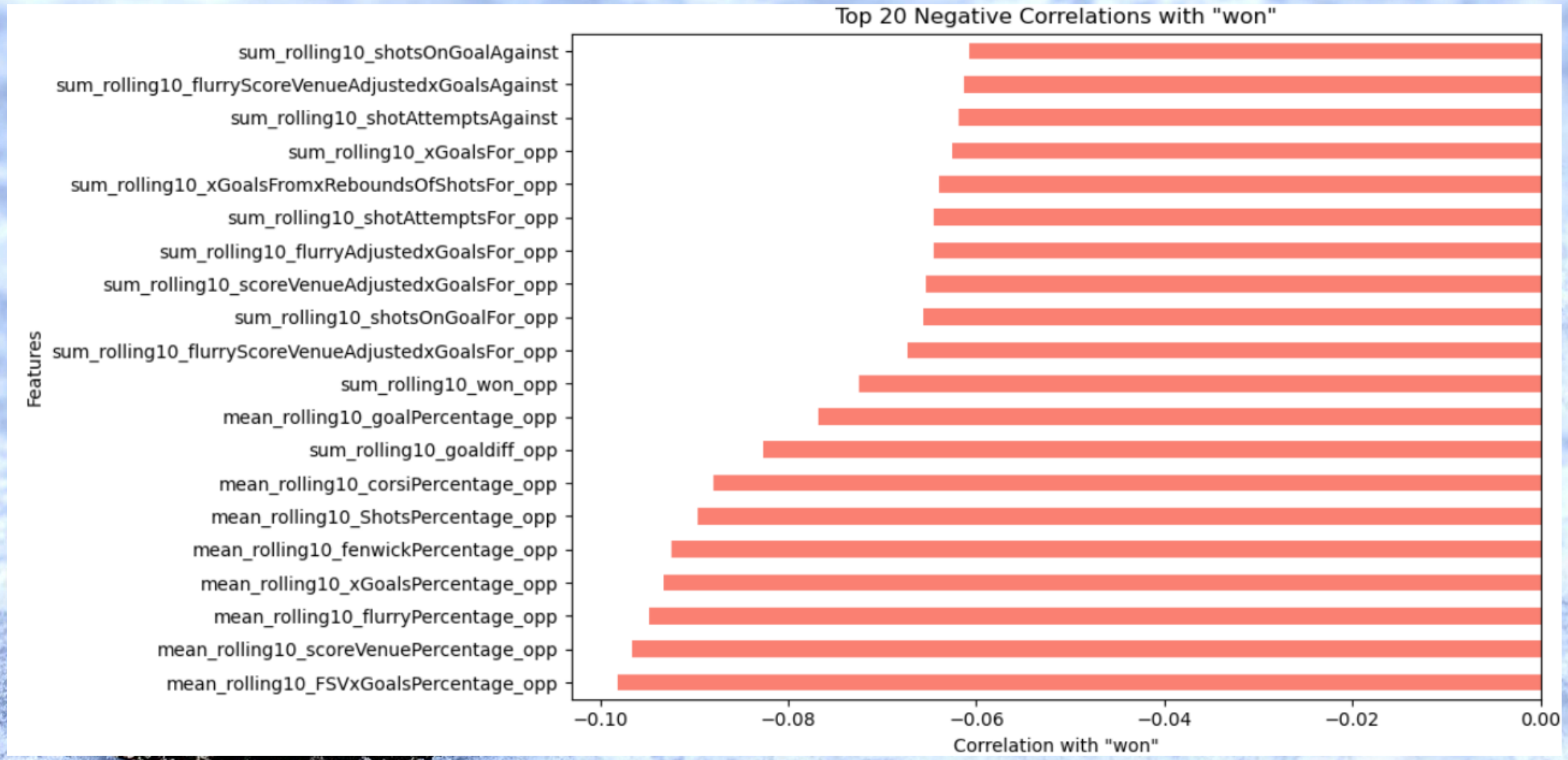
- Is there value in this approach?


Last 10 Games Share of Goals
Winning Team
Losing Team
Density
Goal Percentage
0
1
2
3
4
5
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9

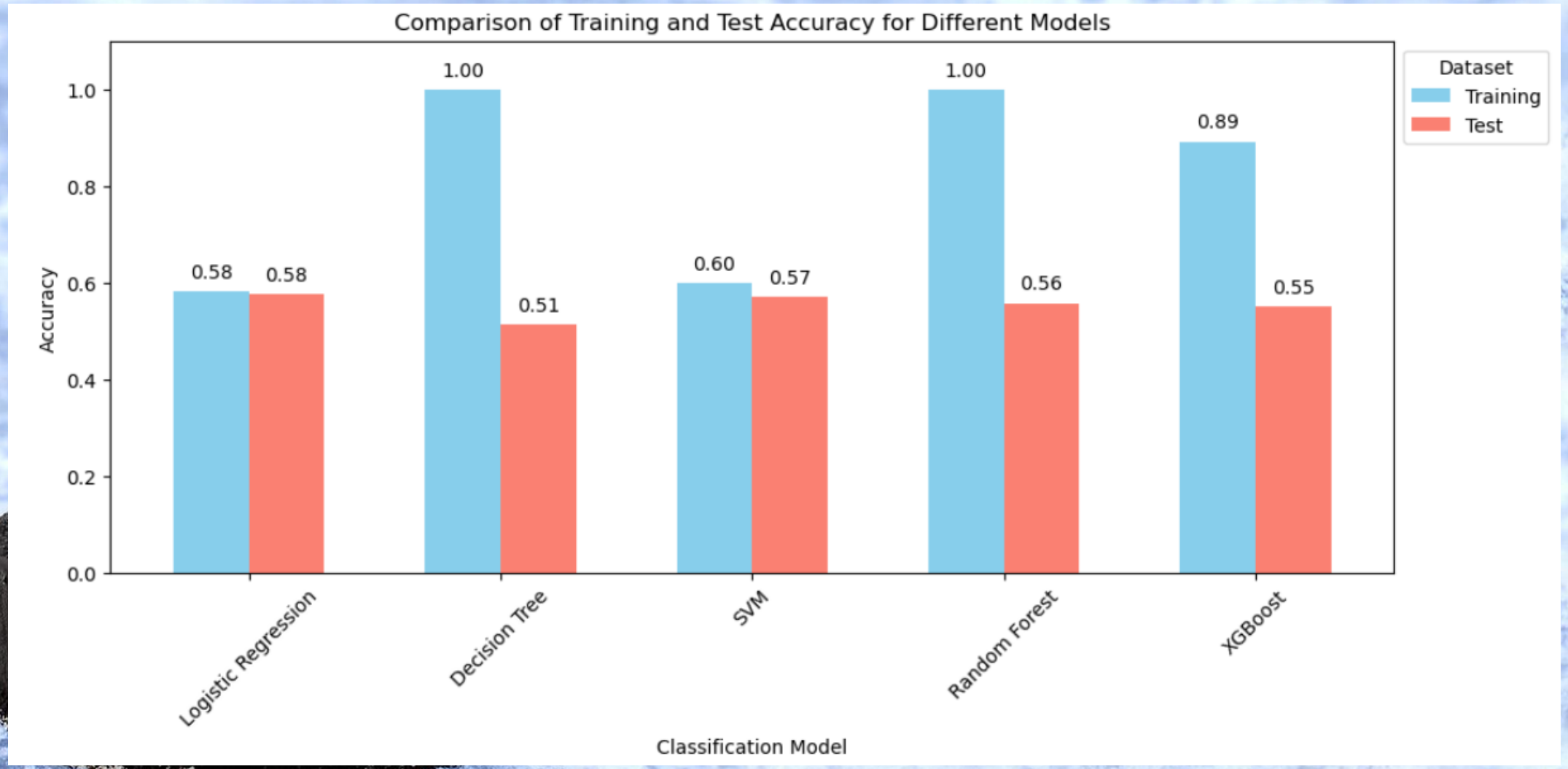

Last 10 Games Share of Shots
Winning Team
Losing Team
Density
0
2
4
6
8
10
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
Shot Percentage


Last 10 Games Adjusted xGoals Percentage
Winning Team
Losing Team
Density
0
2
4
6
8
0.3
0.4
0.5
0.6
0.7
Share of xGoals


Top 20 Correlations with "won"
mean_rolling10_FSVxGoalsPercentage
mean_rolling10_scoreVenuePercentage
mean_rolling10_flurryPercentage
mean_rolling10_xGoalsPercentage
mean_rolling10_fenwickPercentage
mean_rolling10_ShotsPercentage
mean_rolling10_corsiPercentage
HoA
sum_rolling10_goaldiff
mean_rolling10_goalPercentage
sum_rolling10_won
sum_rolling10_flurryScoreVenueAdjustedxGoalsFor
sum_rolling10_shotsOnGoalFor
sum_rolling10_scoreVenueAdjustedxGoalsFor
sum_rolling10_flurryAdjustedxGoalsFor
sum_rolling10_shotAttemptsFor
sum_rolling10_xGoalsFromxReboundsOfShotsFor
sum_rolling10_xGoalsFor
sum_rolling10_shotAttemptsAgainst_opp
sum_rolling10_flurryScoreVenueAdjustedxGoalsAgainst_opp
Features
0.00
0.02
0.04
0.06
0.08
0.10
Correlation with "won"


Top 20 Negative Correlations with "won"
sum_rolling10_shotsOnGoalAgainst
sum_rolling10_flurryScoreVenueAdjustedxGoalsAgainst
sum_rolling10_shotAttemptsAgainst
sum_rolling10_xGoalsFor_opp
sum_rolling10_xGoalsFromxReboundsOfShotsFor_opp
sum_rolling10_shotAttemptsFor_opp
sum_rolling10_flurryAdjustedxGoalsFor_opp
sum_rolling10_scoreVenueAdjustedxGoalsFor_opp
sum_rolling10_shotsOnGoalFor_opp
sum_rolling10_flurryScoreVenueAdjustedxGoalsFor_opp
sum_rolling10_won_opp
mean_rolling10_goalPercentage_opp
sum_rolling10_goaldiff_opp
mean_rolling10_corsiPercentage_opp
mean_rolling10_ShotsPercentage_opp
mean_rolling10_fenwickPercentage_opp
mean_rolling10_xGoalsPercentage_opp
mean_rolling10_flurryPercentage_opp
mean_rolling10_scoreVenuePercentage_opp
mean_rolling10_FSVxGoalsPercentage_opp
Features
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
Correlation with "won"

Revised Modeling Results

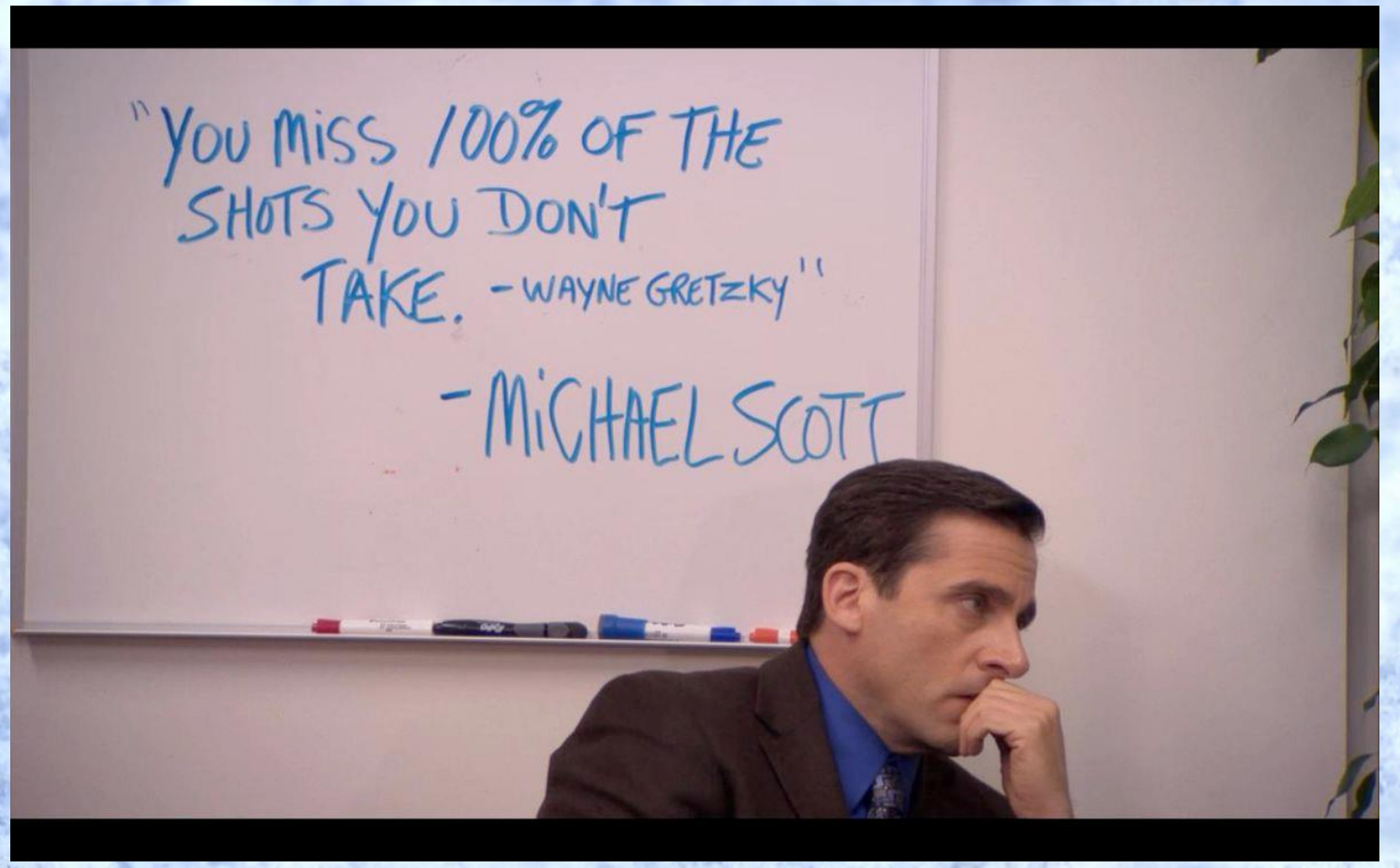


Next Steps

- Optimize pregame variables
  - Number of games to include, weighted for recency?
- Feature selection
- Model optimization and fine tuning
- How to measure success?

Questions?


VANCOUVER


"You miss 100% of the shots you don't take. -Wayne Gretzky"
-Michael Scott