Predicting Michigan Hockey goal differentials: A game-by-game look at the past 3 season

Yaya Harman

STATS 401: Applied Statistical Methods II

April 30, 2025

Word Count: 1487 words

Background

Over the past three seasons, Michigan Hockey has collected game-by-game data on itself and its opponents. By comparing their statistics to the opponent's statistics, an "Objective Sheet" is created. This sheet has evolved over the past three seasons and now contains 24 objectives (statistics) that are compared every game. In this analysis, these objectives were narrowed down to goal differential (GoalDiff), even strength net front shots on goal differential (NFDiff), even strength offensive zone possession time (OZPosDiff), overall chance wins (OverallChanceDiff), and home or away (HA). There were five neutral site games that were removed from the original data to avoid overfitting. The units for OZPosDiff are seconds. The differential is the difference between Michigan Hockey and its opponent for each game in each of the relevant statistics. The response variable is goal differential. The rest of the variables are predictor variables: three quantitative variables and one categorical variable (home or away). The overall goal of this analysis is to best predict the goal differential of a game based on a combination of predictor statistics.

Analysis – Creating the Model

Initial Model Evaluation

The initial model predicts goal differential using all four initial predictor variables: NFDiff, OZPosDiff, OverallChanceDiff, and HA. The initial model's assumption of linearity is reasonable because the observations in the residuals plot are randomly distributed (Figure 1). There's no visual pattern indicating a non-linear distribution. The residuals plot also shows that the initial model's assumption of constant variance is also reasonable because the observations in the plot are randomly distributed and don't show a splaying of the data (Figure 1). The observations at the tail ends of the Q-Q plot are slightly deviating from the reference line (Figure

1). However, because this deviation is fairly minor and the data has more than 30 observations, the assumption of normality is reasonable due to the Q-Q plot and the Central Limit Theorem. The initial model has a R^2 value of 0.2948, suggesting the initial model is a weak fit (Figure 1). Less than 30% of the variability in goal differential can be explained by the initial model. Additionally, the RMSE indicates that the model's average error is roughly 2.535 goals which is a large variation in error for a hockey game (Figure 1).

Log Transformations

This model does not require any log transformations. Goal Differential and the three other quantitative predictive variables all distribute a relatively normal distribution (Figure 2). There is no mild to heavy right-skew exhibited by any of the quantitative variables that needs to be resolved by a log transformation.

Interactions

The interaction between OverallChanceDiff and Home & Away shows no evidence that there is a difference in the effect of overall chance differential on goal differential between home and away games. This is shown by the 0.863 p-value in the summary output and visually by the slopes of both lines of best fit being similar (Figure 3). Due to the lack of statistical evidence, this interaction will not be included in the final model.

Add Predictors

Two predictors, faceoff proportion differential (FODiff) and opponent conference (Conf) were added to the initial model. FODiff requires no transformation as it is normally distributed (Figure 4). Opponent conference is either the Big Ten or non-conference. Adding FODiff to the model increased the adjusted R^2 value from 0.2677 to 0.2893, indicating that FODiff improved the model (Figure 4). This model also continues to follow assumptions of linearity, constant

variance, and normality as seen in the residual and Q-Q plot (Figure 4). Adding Conf to the new model with FODiff increased the adjusted R^2 value from 0.2893 to 0.3413, indicating that Conf also improves the model (Figure 5). Again, this new model continues to follow assumptions of linearity, constant variance, and normality as seen in the residual and Q-Q plot (Figure 5). This is the final model.

Analysis – Interpreting the Model

Diagnostics Plots/Assumptions

As discussed above, the final model continues to follow assumptions of linearity, constant variance, and normality as the initial model did. The residual plot shows randomly distributed observations with no obvious linear or non-linear pattern (Figure 5). The residual plot also shows constant variance because there is no obvious splay in observations (Figure 5). Similar to the initial model, the Q-Q plot shows possible signs of the normality assumption being violated because of the tail end of the observations straying slightly from the reference line (Figure 5). However, similar to the initial model, it is very minor, and there are enough observations to assume normality from the Central Limit Theorem.

Independence

This data likely has multiple independence assumption violations due to numerous factors. Firstly, this data has been collected across multiple seasons with different players, coaches, and staff. The team playing every game is not the same, nor is the coaching strategy, preparation, etc. Unquantifiable effects on a team, such as the concept of "momentum," may impact a team's performance from one game to the next. Additionally, Michigan Hockey plays the same Big Ten Conference teams every season that are also experiencing the same changes Michigan Hockey is. The non-conference teams we play every year are different programs, and

are often weaker programs. This will also impact normality. Overall, this is an imperfect data set that is likely violating independence assumptions.

Model Fit

Unfortunately, this is a very weak model. The R^2 value is 0.3779 which means that only 37.79% of the variability in goal differential can be explained by the model (Figure 5). The RMSE is 2.404, meaning that the model's average error is roughly 2.4 goals which is a lot of goals in the context of a hockey game (Figure 5). A strong model in the context of this data would have an RMSE of less than 1 goal.

Multicollinearity and Overfitting

The Variance Inflation Factor for each quantitative variable does not exceed 5 (Figure 6). This indicates that multicollinearity does not exist in this model. This model is not overfit based on two rough rules of thumb. The sample size of the data is 109 observations which is greater than 60 which is 10x the number of predictors (6). Secondly, each group in both categorical predictors (HA and Conf) have more than 10 observations (Figure 7). This model could be overfit if neutral site games had been included in the final data set because only five neutral site games were recorded. As discussed in the Background, neutral site games were removed from the data to avoid overfitting.

Predictor Results

There are three predictors that are statistically significant. There is some evidence to suggest that faceoff proportion differential impacts goal differential (Figure 5). Controlling for all other predictors, on average, the model estimates that for every 1% increase in faceoff proportion differential, the game's goal differential is expected to decrease by 0.039 (Figure 5). There is strong evidence that there is a difference in conference impact on goal differential.

Controlling for all other predictors, on average, the model estimates that the goal differential in a non-conference game will be 1.642 goals higher than Big Ten games (Figure 5). There is very strong evidence that overall chance differential impacts goal differential. Controlling for all other predictors, on average, the model estimates for every 1 more chance won, the goal differential is expected to increase by 0.189 goals (Figure 5). These three predictors are all statistically significant, however not all are practically significant. A 1% increase in faceoff differential is hard to quantify, it would be more significant if it was the differential in number of faceoffs won. The difference between non-conference and in-conference games being 1.642 goals, while statistically significant and seemingly practically significant (because 1.642 goals is significant in a hockey game) is less practically significant in context. Non-conference games don't play as big a role in making the postseason as winning in-conference games. The most practically significant predictor is overall chances. The final model predicts that on average, for every 6 more chances won, the goal differential will increase by approximately 1. With teams winning an average of approximately 25 chances a game, a difference of 6 chances is fairly reasonable. This is intuitive because if a team is getting more scoring chances, they are more likely to have one of those chances find the back of the net.

Conclusion

Overall, based on R^2 and RMSE, this is a poorly fitted model that does not accurately predict goal differential. This model found that the most statistically significant predictor is overall chances which is an unsurprising finding because the more scoring chances a team has, the more likely a team is to score.

A future model would explore independence violations in the data collection process and would likely narrow the data set down. My next steps would be to use this data to create a logistic regression model to examine how these predictors impact win probability.