Tale Of Two Great Goal Scorers: An Analysis Of How Alex Ovechkin And Wayne Gretzky's Goals Impact Their Teams' Chances Of Winning

Xiaotang (Jeffrey) Zhou

04/12/2022

Introduction

If you follow the National Hockey League (NHL) in any capacity, you would likely know the names of Alexander Ovechkin and Wayne Gretzky. If you don't follow the NHL or don't know one or both of these names, an introduction to them can best be summed up as describing the two greatest goal scorers in NHL history.

To go further in depth, Wayne Gretzky played for four NHL teams between 1979 and 1999, a 20-year career which saw him set over 60 NHL records that still stand to this day, including the record for most career regular season goals with 894, earning him the nickname "The Great One" and the honour of being the only player to have his number 99 retired all around the NHL as the greatest complete player in NHL history. On the other hand, Alexander Ovechkin currently plays for the Washington Capitals, the team that drafted him in 2004, and is widely considered to be the best goal scorer of this generation of NHL players, having amassed 776 career NHL regular season goals as of April 12 2022 and is considered the player most likely to break Gretzky's record of 894.

As a result of Ovechkin having a legitimate shot of breaking Gretzky's goals record as he recently signed a 5-year contract extension with the Capitals, a lot of questions have been raised about whether he can do it and also about who is the greatest goal scorer of all time, with the pro-Ovechkin side claiming that as Gretzky played in an era where a final score of 8-7 was not uncommon for an NHL game, Ovechkin already holds the title of greatest goal scorer as he scored his goals in a time when defensive strategies and goalie equipment have improved dramatically, while the pro-Gretzky side claims that Gretzky was able to score his goals in a time when wooden sticks were used and actions that would be considered clear penalties today would be brushed off as "part of the game", and therefore Gretzky is still the greatest goal scorer.

From this standpoint, it is easy to fall down the rabbit hole of trying to determine whether Ovechkin or Gretzky is the greatest goal scorer of all time, as this comparison will attempt the impossible task of comparing two players with completely different playing styles from completely different eras. Instead, this project will seek to quantify, from a probabilistic point of view, the impact of Ovechkin and Gretzky's goal-scoring prowesses on their teams' odds of winning regular season games. This way, the impact of having different eras present in our analyses will be minimized, as there will not be any direct interaction between the models, with a comparison of Ovechkin and Gretzky only taking place after the models have been built and interpreted.

To get a preliminary snapshot of how each player's teams have performed in games where they scored goals, consider the following bar plots depicting the frequencies of the results of these games:

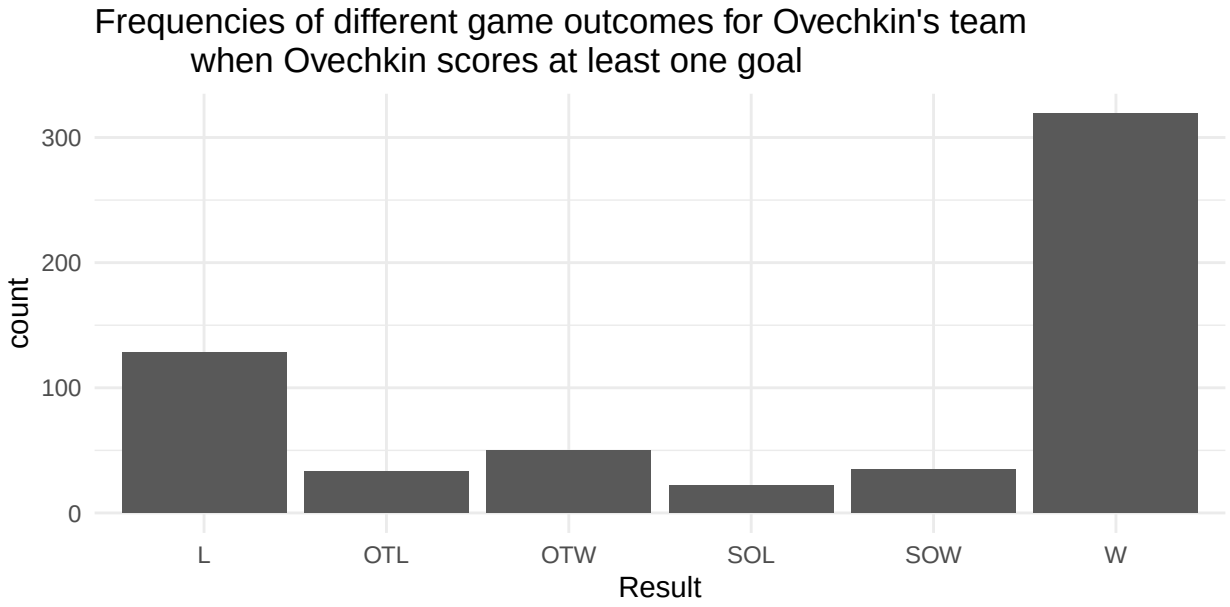

Figure 1: Frequencies of different game outcomes for Ovechkin's team when Ovechkin scores at least one goal

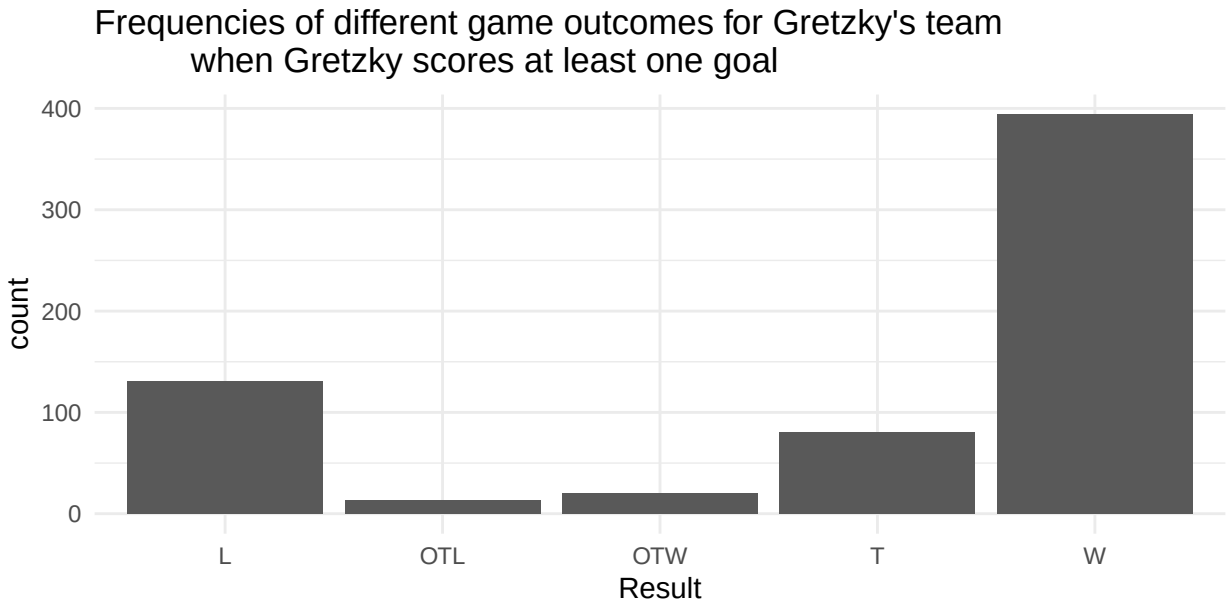

Figure 2: Frequencies of different game outcomes for Gretzky's team when Gretzky scores at least one goal

It is easy to see from the tall bars in the "W" columns in both charts that when either player scores, their team tends to win, which is expected as both players are regarded as key contributors to their teams' success. We can also explore the same results frequencies for games in which Ovechkin or Gretzky did *not* score a goal for their teams:

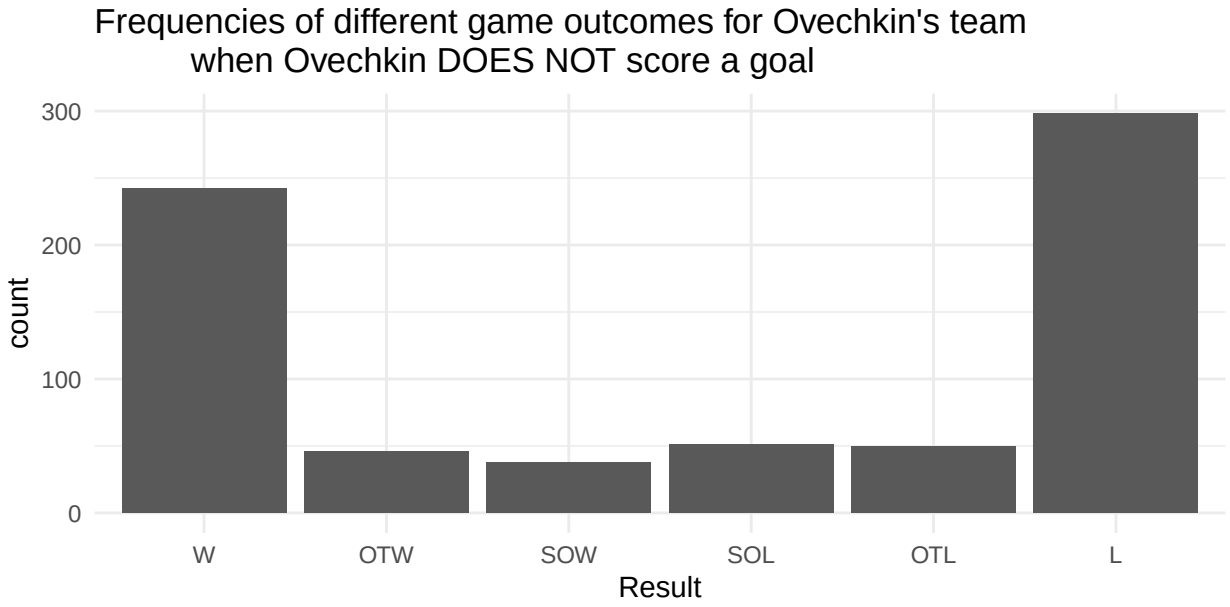

Figure 3: Frequencies of different game outcomes for Ovechkin's team when Ovechkin DOES NOT score a goal

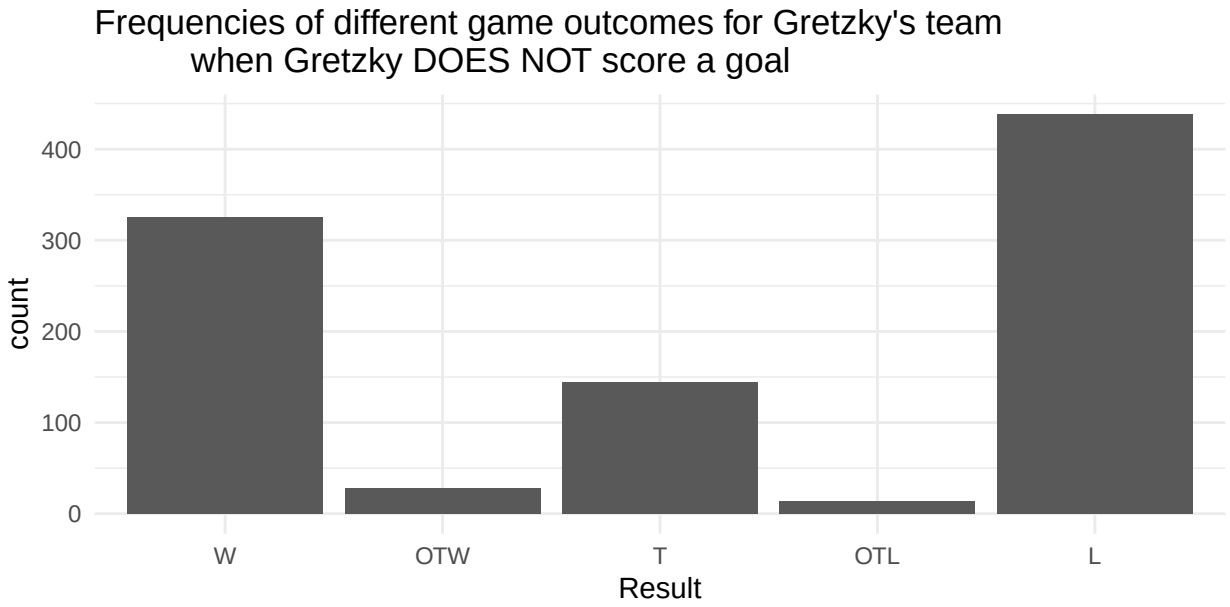

Figure 4: Frequencies of different game outcomes for Gretzky's team when Gretzky DOES NOT score a goal

In contrast to Figure 1 and Figure 2, Figure 3 and Figure 4 from above show that when Ovechkin and Gretzky *don't* score, their teams' loss frequencies actually surpass their win frequencies. This is further proof of how instrumental both players are to their teams' success.

Methods

Data Scraping and Cleaning

There were a total of five datasets that were used in the analysis. The first two datasets contained information on all 894 of Gretzky's career regular season goals while the other contained information on all 776 of Ovechkin's career regular season goals as of April 12 2022. These were derived from Stathead's "Goal Finder" feature (accessible **here**: this is the page for Ovechkin, to find the page for Gretzky click the link and type in Gretzky in the search bar at the top of the page), where data on all of the goals scored by anyone who has ever played in the NHL can be found.

The next two datasets contained information on all of the regular season games that Ovechkin and Gretzky *could have* played during their careers as of April 12 2022, defined to be *all* of the regular season games that were played by the team that each player was a part of, regardless of whether the player himself played in that game or not. For example, Gretzky started the 1995-96 season with the LA Kings but was traded to the St. Louis Blues on February 27 1996, so his corresponding dataset includes the first 64 regular season games played by the LA Kings up to and including February 26 1996, and then changes teams to include the last 21 regular season games played by the St. Louis Blues in the 1995-96 season. These datasets were derived from Hockey Reference's "Schedule & Results" section (accessible **here**: this is the page corresponding to the 2021-22 Washington Capitals for whom Ovechkin currently plays), where data on all of the games ever played by every NHL team to have ever existed can be found.

The last of the five datasets is a table of team names and their abbreviations. This table was manually created in R and it was used primarily (for cleaning purposes) to match an abbreviation of a team's name to that team's full name and vice versa. This dataset is available in the "Data" folder of the project GitHub under the name `name-table.csv`, as are the previously mentioned datasets containing information about all of the goals scored by Ovechkin and Gretzky (available under the names `ovi-data.csv` and `gretzky-data.csv`, respectively) and all of the games that Ovechkin and Gretzky could have played during their careers (available under the names `ovi-potential-games.csv` and `gretzky-potential-games.csv`, respectively).

Model Building

The first step of the model building process was to combine the datasets so that the information about Ovechkin and Gretzky's goals can be matched to the result of the games in which the goals were scored. However, since both players are not only known for scoring just one goal in a game but several, with Gretzky holding the record of 189 for most career multi-goal (2 or more) goal games and Ovechkin currently in third place with 157 (as of April 12 2022), a straight merge of the datasets for each player would not only create an untidy mess with lots of cells having missing (NA) values but it would also introduce bias to the model's predictions, as games where they scored more than 1 goal will be given multiple rows, which will cause the model to count the results of these games more than once. To mitigate this issue, it was decided that for each game in which Ovechkin or Gretzky scored, the *count* of goals would instead be used to indicate both whether they scored and how many goals they scored, while different categories of goals (i.e. first, second, third, or overtime period goals, even strength goals, power play goals, etc.), would also be aggregated by game. This also made games in which they didn't score easy to create rows for, as all that would need to be done was to put 0's in all of the columns that count goals.

The next step of the model building process was to choose the type of model that would be fit to the newly created datasets. As this project seeks to assess from a probabilistic standpoint how Ovechkin and Gretzky's scoring prowesses affect the odds that their teams win games, the response of the model was the results of all of the games Ovechkin and Gretzky could have played while the explanatory variables were the ages of Ovechkin and Gretzky, whether the game was a home game or an away game, and all of the aggregated goal totals created previously. Now, although the response is categorical, its possible values can be given the ordering, from highest to lowest, of "W" (a win where no extra tie-breaking hockey was played), "OTW" (a win where sudden death overtime was needed to break a tie), "SOW" (a win where a penalty shootout was needed to break a tie because no team scored in overtime (only relevant for Ovechkin because games

tied after overtime prior to 2004 ended as ties)), "T" (a tie (only relevant for Gretzky)), "SOL" (a win where a penalty shootout was needed to break a tie because no team scored in overtime (only relevant for Ovechkin)), "OTL" (a loss where sudden death overtime was needed to break a tie), and "L" (a loss where no extra tie-breaking hockey was played). With this ordering of the categories, the type of model fit to the data was the ordinal logistic regression model, which performs logistic regression to predict a response that is categorical but ordered.

To satisfy the model assumption that no multicollinearity would be present between predictors, it was decided that the age of the player, the count of empty net goals scored in a game, the count of penalty shot goals scored in a game, and the total count of goals in a game would not be used as predictors, and that a binary indicator of whether a game was a multi-goal game for Ovechkin or Gretzky would be added and used instead. Also, to satisfy the proportional odds model assumption that the relationship between all pairs of outcome groups are the same, it was decided that three models would be built for each player, where all three models would include the same predictors with the exception that the count of even strength goals (when the scoring team and the opposing team have the same number of players on the ice) scored in a game, the count of power play goals (when the scoring team has one or more players on the ice than the opposing team due to a penalty previously taken by the opposing team) scored in a game, and count of shorthanded goals (when the opposing team has one or more players on the ice than the scoring team due to a penalty previously taken by the scoring team) scored in a game would be evenly split among the three models.

Results

After fitting the three models for each player, the following odds ratios and corresponding 95% confidence intervals were obtained:

Table 1: Summary table of odds ratios and corresponding 95% confidence intervals for the ordinal logistic regression model fit on even strength goals for Ovechkin

	Odds Ratio	Corresponding 95% Confidence Intervals
Away Game	1.4052	(1.1461, 1.7237)
Multi-Goal Game	1.1412	(0.6840, 1.8978)
First Period Goal Count	0.5371	(0.3776, 0.7604)
Second Period Goal Count	0.6198	(0.4441, 0.8631)
Third Period Goal Count	0.5108	(0.3595, 0.7231)
Overtime Period Goal Count	0.9677	(0.5013, 1.8782)
Even Strength Goal Count	0.8171	(0.6142, 1.0869)

Table 2: Summary table of odds ratios and corresponding 95% confidence intervals for the ordinal logistic regression model fit on power play goals for Ovechkin

	Odds Ratio	Corresponding 95% Confidence Intervals
Away Game	1.4041	(1.1454, 1.7225)
Multi-Goal Game	1.1424	(0.6848, 1.8996)
First Period Goal Count	0.4451	(0.3181, 0.6192)
Second Period Goal Count	0.5136	(0.3740, 0.7028)
Third Period Goal Count	0.4219	(0.3130, 0.5655)
Overtime Period Goal Count	0.7948	(0.4294, 1.4749)
Power Play Goal Count	1.1852	(0.8896, 1.5788)

Table 3: Summary table of odds ratios and corresponding 95% confidence intervals for the ordinal logistic regression model fit on shorthanded goals for Ovechkin

	Odds Ratio	Corresponding 95% Confidence Intervals
Away Game	1.4058	(1.1467, 1.7245)
Multi-Goal Game	1.1426	(0.6849, 1.8999)
First Period Goal Count	0.4753	(0.3477, 0.6460)
Second Period Goal Count	0.5522	(0.4126, 0.7371)
Third Period Goal Count	0.4355	(0.3267, 0.5777)
Overtime Period Goal Count	0.8205	(0.4446, 1.5197)
Shorthanded Goal Count	4.2644	(0.6633, 33.8469)

Table 4: Summary table of odds ratios and corresponding 95% confidence intervals for the ordinal logistic regression model fit on even strength goals for Gretzky

	Odds Ratio	Corresponding 95% Confidence Intervals
Away Game	1.9135	(1.5765, 2.3245)
Multi-Goal Game	0.8012	(0.4584, 1.3836)
First Period Goal Count	0.5712	(0.3846, 0.8443)
Second Period Goal Count	0.632	(0.4525, 0.8811)
Third Period Goal Count	0.3082	(0.2099, 0.4486)
Overtime Period Goal Count	1.4206	(0.1574, 12.9742)
Even Strength Goal Count	0.8378	(0.6166, 1.1405)

Table 5: Summary table of odds ratios and corresponding 95% confidence intervals for the ordinal logistic regression model fit on power play goals for Gretzky

	Odds Ratio	Corresponding 95% Confidence Intervals
Away Game	1.9097	(1.5730, 2.3205)
Multi-Goal Game	0.7944	(0.4534, 1.3752)
First Period Goal Count	0.4509	(0.3229, 0.6252)
Second Period Goal Count	0.4920	(0.3647, 0.6612)
Third Period Goal Count	0.2459	(0.1796, 0.3336)
Overtime Period Goal Count	1.2247	(0.1384, 10.9465)
Power Play Goal Count	1.6058	(1.1643, 2.2129)

Table 6: Summary table of odds ratios and corresponding 95% confidence intervals for the ordinal logistic regression model fit on shorthanded goals for Gretzky

	Odds Ratio	Corresponding 95% Confidence Intervals
Away Game	1.9106	(1.5736, 2.3217)
Multi-Goal Game	0.8459	(0.4834, 1.4631)
First Period Goal Count	0.5163	(0.3729, 0.7108)
Second Period Goal Count	0.6192	(0.4658, 0.8216)
Third Period Goal Count	0.2802	(0.2061 0.3777)
Overtime Period Goal Count	1.1100	(0.1275, 9.7393)
Shorthanded Goal Count	0.3947	(0.1949, 0.7379)

Results and Conclusion

Results Discussion

There are three main observations that can be made from the odds ratios in the tables. First, it can be seen from the six tables that across all three models for each player, when an away game is played as opposed to a home game, the odds of either players' team moving from a regulation loss to an overtime loss, shootout loss, tie, shootout win, overtime win, or regulation win increases at the same rate regardless of what type of goal in terms of strength is the predictor in their models. That is, across all three models for Ovechkin and all three models for Gretzky, an away game as opposed to a home game for either players' team causes these odds to increase by about 40 percent for Ovechkin and 91 percent for Gretzky. This shows that both Ovechkin and Gretzky's teams were better road teams than home teams, which is expected for Ovechkin as the Washington Capitals, who he has played for his whole career, have always seemed to perform better on the road than in front of their home fans, but surprising for Gretzky as teams that he has played on during his career never had a reputation for playing better on the road than at home, which makes it all the more interesting that the percent increase in odds for Gretzky is over double that for Ovechkin.

Another observation that can be made is that for all three of Ovechkin's models, the odds ratio corresponding to the binary predictor (0 or 1) indicating whether Ovechkin scored at least two goals in a game hovers around 1.14, which means that for a unit increase in this variable, or in other words for a multi-goal game registered by Ovechkin, the odds that his team moves from a regulation loss to an overtime loss and so on increases by 14 percent. In contrast, for all three of Gretzky's models, this odds ratio ranges approximately from 0.79 to 0.85, which means that for a multi-goal game registered by Gretzky, the odds that his team moves from a regulation loss to an overtime loss and so on actually decreases around 15 to 21 percent. This means that a multi-goal game registered by Ovechkin is more important to his team's success than an analogous multi-goal game registered by Gretzky, which can possibly be explained by the fact that over Ovechkin's whole career he has been relied on as the main source of goals for his team, and so a multi-goal game by him is more important to his team's ability to achieve better results than a multi-goal game registered by Gretzky, who was not relied on nearly as heavily by his teams to score goals as Ovechkin is.

Lastly, it can be seen that the odds ratios corresponding to the predictors aggregating first, second, and third period goals that are below 1 don't make very much sense in the context of the data or the question that is being answered. That is, these odds ratios that are below 1 imply that for a one unit increase in the number of first, second, or third period goal, the odds that Ovechkin or Gretzky's team moves from a regulation loss to an overtime loss and so on somehow *decreases*. This can be attributed to the fact that there are more games in which Ovechkin and Gretzky *didn't* score than there are games in which they scored, and the predictors that aggregate first, second, and third period goals take 0 as their value a lot of times, which will cause the odds ratios to get pulled down.

Limitations

There were two main limitations in the results of this analysis. First, the most obvious limitation is that due to there (naturally) being more games where Ovechkin and Gretzky *didn't* score for their teams than there were games where they did, misleading coefficient estimates and odds ratios were assigned to the predictors aggregating first, second, and third period goals that implied that a one goal increase in first, second, or third period goal being scored by either player in a game somehow meant a *decrease* in the odds that their respective teams could move from a regulation loss to a overtime loss and so on. This could possibly have been avoided had a different choice in predictors been made, or the model building process had been different, with one possible alternative approach being to build a model just for games where Ovechkin and Gretzky scored and using the remaining games where they didn't score as a sort of control group. The second limitation is that some of the variables in the original dataset, such as player age and total count of goals in a game, could not be included as predictors in the final model due to problems with satisfying model assumptions. This could be avoided next time with a different type of model that is more robust to multicollinearity and problems with the relationship between all pairs of outcome groups being the same, which are the assumptions that forced the decision to drop these variables as predictors.

Conclusion

All in all, this analysis of data on all of Ovechkin and Gretzky's goals and all of the games they possibly could have played during their careers concluded that both Ovechkin and Gretzky's teams had more success in games away from their home arena than they did in games at their home arena, and also concluded that a game in which a multi-goal game registered Ovechkin was more beneficial to Ovechkin's team than a multi-goal game registered by Gretzky would be to Gretzky's team. This can be attributed to the fact that for all of Ovechkin's career, he has been tasked as the main source of goals while Gretzky had less of that responsibility on teams he played for, and it also shows that Ovechkin is a more relied upon player as a source for his team's success goal-scoring wise than Gretzky was.

However, this project did not come without its flaws and limitations, as it can be seen that the dataset and the predictors used to fit the final model were flawed in a way that yielded some conclusions that were simply nonsensical in the context of the question. Thus, future iterations of similar projects could employ methods that are more robust to multicollinearity and other common problems in building prediction models, which would allow for stronger comparisons to be made and better conclusions to be drawn.

References:

Alex Ovechkin: Goal finder. Stathead.com. (n.d.). Retrieved April 21, 2022, from https://stathead.com/hockey/goal_finder.cgi?request=1&order_by=scorer&match=goallist&season_start=1&is_playoffs=N&grouping=p&season_end=-1&age_max=99&scorer_id=ovechal01&offset=0

Every NHL Team. Hockey Reference. (n.d.). Retrieved April 21, 2022, from https://www.hockey-reference.com/teams/

Lee, E. (2019, May 29). Ordinal logistic regression on World happiness report. Medium. Retrieved April 21, 2022, from https://medium.com/evangelinelee/ordinal-logistic-regression-on-world-ha ppiness-report-221372709095

Rawat, A. (2019, January 6). Ordinal logistic regression. Towards Data Science. Retrieved April 21, 2022, from https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-re gression-1ee699274cf5

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wayne Gretzky: Goal finder. Stathead.com. (n.d.). Retrieved April 21, 2022, from https://stathead.com/hockey/goal_finder.cgi?request=1&scorer_id=gretzwa01&scorer_id_select=Wayne%2520Gretzky&utm_medium=hr&utm_source=player-page-links&utm_campaign=hockey&___hstc=88549636.f725800c032be8dc0bf471dd62470201.1649632332067.1650588069497.1650599016599.5&___hssc=88549636.3.1650599016599&___hsfp=4176485162